



UNIVERSITAT DE
BARCELONA

From the discovery of epistatic events in Type 2 Diabetes Mellitus to the study of related gene expression regulatory variation

Lorena Alonso Parrilla



Aquesta tesi doctoral està subjecta a la llicència **Reconeixement- NoComercial – SenseObraDerivada 4.0. Espanya de Creative Commons.**

Esta tesis doctoral está sujeta a la licencia **Reconocimiento - NoComercial – SinObraDerivada 4.0. España de Creative Commons.**

This doctoral thesis is licensed under the **Creative Commons Attribution-NonCommercial-NoDerivs 4.0. Spain License.**

TESI DOCTORAL

**FROM THE DISCOVERY OF EPISTATIC EVENTS IN
TYPE 2 DIABETES MELLITUS TO THE STUDY OF
RELATED GENE EXPRESSION REGULATORY
VARIATION**

Lorena Alonso Parrilla
2018-2022



UNIVERSITAT DE
BARCELONA



FROM THE DISCOVERY OF EPISTATIC EVENTS IN TYPE 2 DIABETES MELLITUS TO THE STUDY OF RELATED GENE
EXPRESSION REGULATORY VARIATION

From the discovery of epistatic events in Type 2 Diabetes Mellitus to the study of related gene expression regulatory variation

Programa de doctorat: Biomedicina (HDK05)
Tesi realitzada a: Barcelona Supercomputing Center
Memòria presentada per: Lorena Alonso Parrilla



Doctorand: Lorena Alonso Parrilla



Supervisor: David Torrents Arenales

Tutor: Jose Luis Gelpi Buchaca

DEDICATÒRIA I AGRAÏMENTS

Dedicatòria i agraïments

Hola, soc la Lorena Alonso Parrilla, autora d'aquesta tesi. Abans de començar a llegir, m'agradaria poder presentar-me, ja que penso que potser això pot ajudar a algú en un futur; a vegades és important sentir-se identificat per a perdre la por i fer el primer pas endavant. Soc una matemàtica a la qui agrada la matemàtica aplicada centrada en la Biomedicina. Així, el 2011 vaig acabar la llicenciatura de Matemàtiques a la UB, el 2013 vaig aconseguir un màster d'Estadística Aplicada a la UNED i el 2016 vaig completar un màster en Biomatemàtica, bioinformàtica i genòmica computacional a la UOC. Des de l'últim any de carrera vaig estar treballant com a programadora web i analista de dades i, acabat els màsters, vaig a entrar al Barcelona Supercomputing Center (BSC) com a Research Engineer al grup del David Torrents. Dos anys després, el 2018, vaig començar el doctorat en Biomedicina en el mateix grup. I ara, el 2022, ja tinc més de 35.7 anys, visc amb la meva parella que m'ha acompanyat en tota aquesta aventura des de gairebé principis de la carrera, i tinc un fill amb poc més de 0.5 anys. Aquesta tesi, són els apunts i els resultats d'aquests quatre últims anys d'estudi, i han estat escrits intentant que pugui ser un document entenedor per sí mateix, proporcionant diversos materials i fonts de consulta. Per aquest motiu, es fa un gran èmfasi en introduir i aprofundir en els conceptes de Biologia i metodologies informàtiques, ja que resulta fonamental per a entendre les anàlisi i els resultats obtinguts.

Feta la presentació, només tinc paraules d'agraïment cap a totes aquelles persones que han facilitat que aquesta tesi surti endavant. Primer de tot, agrair els que han posat diners per a que jo i aquest projecte de I+D+i (R&D i Innovació) sortim endavant, amb la beca BES-2017-081635 finançada pel MCIN i per "FSE Invertint en el teu futur". Vull agrair al BSC les bones condicions que ens donen com a centre, no només a nivell laboral, sino també a nivell de visibilitat, instal·lacions i, perquè si no fos perquè tenim el Mare Nostrum, aquesta tesi no tindria sentit. A tota la gent del centre, des de recursos humans, gestió documental i de beques, support, helpdesk, finances,... ja que si no fos perquè estan allà, per la seva amabilitat, ajuda, paciència i bon tarannà, moltes de les gestions, instal·lacions de software, hardware que s'han hagut de fer durant la tesi haurien estat més complexes, haurien portat més temps, i per tant, haurien dificultat l'avanç de la recerca i l'obtenció de resultats. Gràcies per facilitar la vida d'aquesta estudiant. En especial vull expressar el meu agraïment a la gent del departament de Life Sciences, que fan un gran esforç dia a dia per a crear el millor ambient d'aprenentatge per a tots, per fomentar el diàleg científic, els vincles i col·laboracions amb els diferents departaments, i per donar-nos un espai de treball on podem practicar la divulgació científica; tots aquests esforços ens ajuden a créixer des del primer dia. A més, en particular al Dr. Jose Luis Gelpi, vull agrair-li que hagi estat el meu tutor de tesi. De veritat que considero que he tingut molta sort perquè, per a tot el que he necessitat quant a seguiment i gestions amb la universitat, m'ho ha facilitat moltíssim i això... és molt d'agrair.

Bueno, Dr. David Torrents, que sé que no t'agraden aquest tipus de formalismes però acceptes que de tant en tant jo sigui molt clàssica, gràcies per acollir-me i donar-me la possibilitat de poder fer el doctorat al teu grup. Sé que ha de ser un repte obrir la porta a un personatge com jo, per la meva particular forma de ser, fer i expressar-me; i per la gran falta de coneixements que tenia. No pots arribar a imaginar lo feliç que he estat en aquests darrers 6 anys en el grup, m'has obert les portes de casa i a sobre m'has donat totes les eines i facilitats per a que aprengué. Mai m'ha faltat ajuda al voltant perquè tu t'has encarregat de posar-me a treballar amb els millors mestres que tenies a l'agenda i gràcies a això, he pogut aprendre moltes coses. Tant ha estat així que ara, fins i tot, ja puc parlar de ciència amb tu amb més tranquil·litat. Tot i així, sabem que encara tinc molt a aprendre i que de tant en tant encara dic alguna barbaritat. Gràcies per fer de guia, conseller, a vegades casi un pare, per donar-me reptes, trencar-me la ment i els esquemes quan estava més segura i per donar-me confiança quan estava menys convençuda, però sobretot, gràcies per donar-me aquesta oportunitat. M'he sentit tan bé des del primer dia, que per primera vegada a la vida, he volgut continuar en una feina durant tants anys. Ara el problema serà que quan una està tan bé en algun lloc, doncs no vol marxar.

I tot això no hauria estat possible sense el Txema, la Montse i la Romina, que van ser els encarregats de fer-me una entrevista de feina per a fer de Research Engineer en un projecte europeu; el TIGER. És molt curiós que des del primer moment em vaig sentir ben còmoda amb vosaltres. Recordo aquella entrevista com si fos ahir i semblava que us coneixia de fa molt de temps. Us estic súper agraïda d'haver-me triat com a candidata per a poder entrar al grup, per tota la guia i tota l'ajuda que m'heu donat durant aquest temps. Feu que treballar sigui un plaer al vostre costat. Però a banda de tot això, gràcies per la vostra paciència, comprensió i per aguantar el meu nervi o rebeldia de tant en tant. Si d'alguna cosa estic segura és que si tingués que repetir la mateixa història, us triaria com a companys i guies de camí. En particular, Txema, he après tantes coses de tu, a nivell laboral i com a persona, de veritat que em fascina trobar-me a algú que tingui tanta empena i esperit crític i de superació; ha estat dur seguir-te el ritme a vegades però m'ha encantat fer projecte, reunions i discussions amb tu. Gràcies per tot el que m'has ensenyat. Romina, em va encantar començar el projecte amb tu, aprenent i barallant-nos amb la plataforma, les dades i les màquines virtuals. És un gust trobar-se pel camí a una persona tan predisposada sempre a donar un cop de mà i amb aquest somriure a la cara. Gràcies per ajudar-me amb les primeres batalles, per fer-me companyia i per escoltar-me quan estava de queixa. La teva pau, comprensió i el teu somriure, sempre han estat de gran ajuda. I Montse, què més a dir que un se sent molt afortunat de que estiguis sempre allà. Lo teu és impressionant; per a una conversa de feina, per a arreglar un codi, per a veure què fer amb unes dades, per a processar algo al mare, per donar suport moral... la pregunta és què fariem sense tu? Gràcies per tot això i per estar sempre per nosaltres.

But talking about TIGER without mentioning all the people involved in the project doesn't have any sense. For this reason, I want to thank all the T2Dsystems Consortium for trusting in me to participate in this huge project, which has represented a major challenge for me. Particularly I want to thank them also for their patience, since they have been listening my updates on the platform in each follow-up session without showing any signal of boredom; this has been a very good practice for me. Thank you for helping me to grow up. Here I must specially thank Dr. Miriam Cnop for giving support and pushing the project until the publication, Dr. Jorge Ferrer for his support in some calls when David was not able to attend and for facilitating us the publication of the cASE method, and many thanks to Anthony, it has been a pleasure to share this project with you and Ignasi, to be with you in our regular meetings, and a pleasure to work with you during your visit to the BSC.

Perquè sí, quan un comença un projecte s'hi va trobant a més gent pel camí. I així vaig tenir la sort de conèixer a l'Ignasi; que en aquella època era encara un estudiant de PhD. Curiós va ser el fet que em parlava en anglés i jo pensava... amb aquest nom i aquests cognoms... i en anglés... Anda que no em vaig posar contenta a Sitges quan vaig veure que parlaves català! Em vas caure súper bé (compte, no només per parlar català ehh); la cosa que des d'aquell moment vaig sentir que ja hi havia algú més a l'equip. I déu ni do quin "fitxatge"! La veritat que estic encantada d'haver-te conegut i treballar junts amb el TIGER, però més contenta estic des que vas entrar al BSC de postdoc. Gràcies per donar-me la oportunitat de poder treballar amb tu en el projecte d'epistàsia; ha estat dur, perquè ets súper crític i no et talles ni un pèl a l'hora de dir que algo està malament, però amb tu un sempre està segur perquè sap que aquí s'està fent ciència de la bona. A banda que és molt divertit, de tant en tant, fer un gràfic curiós que et faci explotar el cervell. Pues què et puc dir havent estat tan gran mestre... si és que fins i tot m'has ensenyat a llegir papers, amb la teva idea de fer un journal club, Ignasi! Saps que segueixo amb gust les teves passes... fins i tot hem tingut un nen gairebé alhora xD. De tot cor, moltes gràcies per posar-me en bon camí.

Claro, no puedo seguir adelante sin agradecer al Dr. Juan Ramón González el haberme ofrecido la posibilidad de colaborar en su proyecto de inversiones y, por habernos dado el empujón inicial y el soporte necesario con epistasia. Debo confesar que fue muy estimulante, a la vez que

esperanzador, encontrar a un matemático trabajando en genómica. Otro ejemplo a seguir! Mil gracias Juan Ramón.

And project by project I also get to the hands of Cecilia, my desktop mate in the office; and what a mate! Since you get to the BSC I've enjoyed a lot learning by your side. Your love for science is so big that you make it so easy to participate and to discuss about everything. I can imagine that it was the reason for you to decide joining the Journal Club. It was very nice to broaden the discussion with both Dani and you to other germline topics. Thank you guys for sharing all those Monday mornings with me, talking about science in English, learning how to read, explain and criticize (in a positive manner) a paper. All this time shared with you has been of great profit for me; very enriching but also funny. Last Ceci, thank you for helping me in collaboration with Ignasi, to improve my writing skills while writing the Mathematics review; your organized way of behaving encompassed with the discipline in your work and speech have been a great guidance for me.

Y así ha ido avanzando esta tesis; como algunos comentaban en el grupo, he sido muy afortunada de compartir proyecto y contar con el apoyo y guía de 3 postDocs; y no están nada equivocados. Sin embargo, esa afirmación está incompleta porque mi fortuna no sólo es haber contado con la ayuda de Txema, Ignasi y Cecilia, sino también de contar con todo un grupo de investigadores que siempre están dispuestos a echar una mano, a discutir de ciencia y no ciencia, a compartir su trabajo y a dar su opinión crítica sobre el trabajo que uno presenta. Además sin importar el formato ya sea oficial o extraoficial, en un meeting de grupo, en una pausa para el café... siempre disponibles para ayudar a mejorar en el trabajo y como persona. En este grupo se aprende hasta a comer más sano! Gracias a todos los ya mencionados y a Mercè, Elias, Sílvia, Marta, Juan, Alex, Jordi, Luisa, Ana, Michelle, Lydia, Álvaro, Migue, Ramón e Iván por todos esos momentos compartidos, por ser un ejemplo y referente día tras día, por esos cafeticos y esas charlas apasionantes de ciencia, política, del día a día..., por escucharme cuando lo necesité, por darme vuestro apoyo cuando estaba de bajón, por aclararme las dudillas que me iban saliendo sobre la marcha, por aceptarme en el grupo y por la paciencia que habéis tenido conmigo (cuántas veces se me tuvo que repetir lo que hacía la RNA-polimerasa! xD)... gracias en definitiva por estar ahí, ser cada uno como sois y ser grandes maestros para mi. De corazón os digo que cada uno de vosotros ha contribuido en mayor o menor grado a que todo esto salga adelante; esta tesis y yo.

Y ya casi al final de mi trayectoria de estudiante de doctorado, he tenido la suerte de contar con el apoyo del tribunal de tesis, a los que quiero agradecer su amabilidad al recibir mi propuesta, su disponibilidad para hacer posible la defensa, con toda la faena adicional que esto conlleva, y sobre todo, desde la admiración a su trabajo, quiero darles las gracias por concederme el honor de poder discutir con ellos sobre genética, en particular, centrándonos en las líneas de investigación que compartimos. Així, inclouré aquí al Dr. Rafael de Cid, a qui ja li estic súper agraïda per ser el primer en dir-me que sí i haver fet que els nervis d'aconseguir un tribunal de tesi, no siguin tants. És una sort que el tema de la fibrosi quística sigui del teu gust! xD. A la Dra. Ana Viñuela, el Dr. Ferran Reverter, que en un tiempo récord aceptaron formar parte de mi tribunal, facilitándome todos los trámites requeridos por la universidad. Al Dr. Àlex Sànchez-Pla, que aún solapándose la defensa de tesis con la presentación de sus oposiciones de cátedra, se ofreció para formar parte de mi tribunal como suplente. Y la Dra. Alicia Huerta-Chagoya, que pese a no poder ofrecerle una plaza en el tribunal principal, muy amablemente me ayudó a completar el tribunal suplente en un tiempo récord también.

Finalmente, quiero dar las gracias a los de siempre; mis dioses creadores: mi papá y mi mamá, que no tuvieron suficiente con crearme y criarme sino que además me escuchan, me acompañan y me apoyan a cada paso que doy. Y así les da igual si les doy una chapa de mates, de expresión de tejidos o de pañales... son los grandes pilares de mi vida; sin ellos no hubiera podido llegar aquí. Al nen, que ha aguantado estoicamente a mi lado una vez más la inmersión en otra etapa de estudio; qué paciencia conmigo! Gracias una vez más por tu apoyo, comprensión y cariño. A mi

pequeño Andreu, que aguantó 2 días más dentro de mi panza para que yo acabara de escribir hasta la última línea de mi primera versión de tesis, acompañándome y sufriendo desde dentro la última etapa de este camino, y haciendo siestecitas una vez fuera para que yo pudiera continuar. A mis amigos, a los que están o han estado acompañándome durante el camino de tesis y no tesis; en especial a mis dos soles, porque siempre es de gran ayuda y fortalece el tener a alguien que te acepta y con quien compartir hasta el silencio que sale del corazón. Y a Auron, porque siempre está para ayudarme a desconectar de lo que haga falta.

En general, a todos los que mencioné y a todas las personas que han compartido espacio-tiempo conmigo durante la tesis y también antes, porque... todos los que nos rodean pueden ser nuestros maestros si estamos dispuestos a observarlos y a aprender de ellos. A todos aquellos que marcaron mi camino y contribuyeron a que hoy sea tan tan tan feliz escribiendo esto... muchas gracias. Y... espero no haberme olvidado de nadie con esta cabeza loca que tengo, GRACIAS por compartir estos años conmigo. Un gusto aprender a vuestro lado; llevo conmigo al futuro TODO lo que fui capaz de aprender de vuestras enseñanzas, feliz de haber estado rodeada de tan grandes personas. Y como diría un gran señor pero a mi manera... quiero acabar esta sección con un: Seguim! ^^

*“Hasta el viaje más largo empieza con un solo paso”
Proverbio japonés*

ABSTRACT

Abstract

One of the major and most challenging goals of Biomedicine during the last centuries has been the study of the human biological mechanisms, and its relation with traits and diseases. Particularly, in the case of complex diseases, such as Type 2 Diabetes (T2D), asthma or Alzheimer, special interest has been devoted to understanding the underlying molecular mechanisms that affect the development of complex diseases, and the biological processes involved in the preservation of these diseases across generations (genetic basis). In this direction, during the last decades, the advance of computing as well as the development of new DNA-related technologies has largely contributed to the faster development of methods, tools, and resources, which have enhanced the genetic study of traits and diseases. As a result of this revolution, new specialised fields such as Biomedicine, Bioinformatics, and Computational genomics have emerged to find the genomic basis of disease using computational tools. Hence, the identification of the genetic factors behind complex diseases has evolved into a multidisciplinary effort, which combines disciplines as diverse as Biology, Mathematics, Physics, Chemistry, and Information technology.

The Computational genomics field, in the context of Biomedicine, focuses on the study of the relationship between genomic changes (variants) and the predisposition or the offset of disease with the final aim of understanding, predict and prevent diseases and, ultimately, to design better treatments. In this direction, numerous contributions have been made in this field to discover variants associated with the risk of developing a disease, and to interpret these associations in terms of function. Notably, some of these contributions, such as the assembly and annotation of the human reference genome, improvements on disease characterization, the better understanding of the effects of genomic variation in different populations, or the introduction of Genome Wide Association Studies (GWAS), have represented very relevant landmarks for the advance on the understanding of the genetic basis of diseases. Particularly, the broad use of GWAS, which mostly relies on the statistical comparison between the variants present in groups of diseased and non-diseased individuals, have led to the discovery of thousands of genomic variants associated with a great diversity of complex traits and diseases.

Despite the great success of GWAS, the multiple limitations surrounding this type of approaches, has converted the study of complex diseases into a still challenging problem. Particularly, there are many elements, such as the need of analysing large cohorts of individuals, or the difficulties to generate a complete model to capture the whole complexity of common traits, which limit the discovery power of GWAS. Therefore, reducing the explanation of disease heritability, based on GWAS findings, to a small fraction. Moreover, the lack of biological and functional interpretation of the results obtained from GWAS has complicated its translation into something meaningful to be applied in the clinics. Consequently, many statistical and computational efforts have been devoted to improve GWAS discovery power, and to develop new analytical frameworks to find new disease-susceptibility variants. Additionally, other biological approaches, such as transcriptomics and epigenetics have emerged as a key to facilitate the interpretation of GWAS outcomes. Finally, the need for accessibility to this valuable genomic, transcriptomic and epigenetic information has led to the generation of a wide diversity of publicly available databases.

This is the case of Type 2 diabetes (T2D), which is a complex metabolic disorder mainly known to be caused by islet beta-cell dysfunction usually surrounded by a background of insulin resistance. T2D is an example of a common disease that has been broadly studied from the perspective of different omic layers. Particularly, the genetic study of T2D has led to the discovery of more than 700 genomic variants significantly associated with the disease, thousands of genes with a putative effect on the disorder, and thousands of target genomic regions with potential regulatory effects. However, although the genomic explanation of its heritability is estimated around 70%, approximately only 20% has been already explained and, most importantly, the use of these markers

to detect the predisposition of an individual to develop the disease is still far for the clinics. Additionally, most of these genomic signals lack functional explanation, thus representing a challenge for the understanding of disease pathophysiology.

Consequently, the general objective of this thesis is to broaden the genetic understanding of complex diseases, focusing on the analysis of T2D, by finding new disease-susceptibility loci and improving the functional interpretation of genetic markers. In this direction, the objectives of this thesis can be summarised in:

- 1) Discover epistatic groups of variants associated with T2D, applying combined machine learning and statistical approaches, and analyse their underlying molecular mechanisms to enhance the early detection of the disease and a better comprehension of its pathophysiology.
- 2) Generate a comprehensive database of human pancreatic islets gene expression regulatory variation, which integrates genomic, transcriptomic and epigenetic data related to diseases, genes and variants to improve the functional study of T2D and other islets related traits (Alonso, Piron, Morán, & et al., 2021).

Additionally, this thesis recapitulates the participation in two studies with the objectives:

- 3) Support the relevance of inversions and their effect in islets expression to improve the genetic knowledge about the shared-susceptibility of complex diseases (González et al., 2020).
- 4) Review current GWAS statistical frames to promote the development of new methods and tools that can enhance the study of complex diseases (Alonso, Morán, Salvoro, & Torrents, 2021).

Therefore, I start this document with a detailed introduction that aims to facilitate the comprehension and motivation of this study, followed by the hypotheses related to milestones 1-2), and the corresponding list of objectives. This section is followed by a report made by Dr. David Torrents, the director of this thesis, summarising my trajectory during the PhD, and detailing my contributions to the studies related to milestones 1-4) during this period. This report is followed by a brief summary of the studies presented in this thesis.

Then, for the study of milestone 1), an unpublished manuscript is provided summarising the preliminary results obtained from the analysis of variant-variant interactions and its association with T2D using machine learning and statistical approaches. Therefore, describing and discussing the last advances done, specifying the methods used, and discussing the outcomes and limitations of the preliminary analyses. Next, a publication is provided to support the results obtained from the study of milestone 2). Thus, detailing and discussing the human pancreatic islets gene expression variation results that constitute the core of the database. Additionally, two appendix sections have been provided in this document to include the publication and review related to milestones 3-4).

Finally, the global results obtained from the study of milestones 1) and 2) are summarised and discussed, and a list of conclusions is provided to briefly recapitulate the main outcomes of this thesis.

INDEX

Index/Content

Dedicatòria i agraïments	4
Abstract	10
Index/Content	14
Tables and Figures list	20
Figures	22
Tables	22
Supplemental Materials	22
Abbreviations list	24
1. Introduction	28
1.1. Biomedicine and the study of human diseases	30
1.1.1. Motivation: The study of the genetic basis of human diseases	30
1.1.2. Fundamentals of genetics and genetic inheritance	30
1.1.3. DNA alterations and inheritance patterns	32
1.1.4. Genetic variants classification	33
1.1.5. The effect of genetic variants and disease characterization	34
1.1.6. From genomic variation to its functional interpretation	35
1.1.7. Preparation of different omic data for genetic studies	36
1.1.7.1. DNA sequencing	37
1.1.7.2. DNA microarrays	37
1.1.7.3. RNA sequencing	37
1.1.7.4. Single-cell sequencing	38
1.1.7.5. Genotyping arrays	38
1.1.7.6. The evolution of sequencing and microarrays	38
1.2. Genetic studies and complex diseases	38
1.2.1. The Human Genome Project and the human genome sequence	39
1.2.2. Genetic variability maps	40
1.2.3. Discovery of variants associated with complex diseases	40
1.2.3.1. Genome Wide Association Studies (GWAS)	41
1.2.3.2. GWAS limitations	42
1.2.3.2.1. Statistical power	42
1.2.3.2.2. Complex interaction models	43
1.2.3.2.3. Lack of functional interpretation	44
1.2.3.3. Machine learning (ML) approaches	45
1.2.3.4. ML limitations	45
1.2.3.4.1. Data pre-processing	46
1.2.3.4.2. ML algorithms	47
1.2.3.4.3. Lack of functional interpretation	48

1.2.4. Molecular basis of complex diseases and functional interpretation	49
1.2.4.1. Gene expression	50
1.2.4.2. Gene expression regulation	50
1.2.4.4. <i>cis</i> -regulatory expression	51
1.2.4.4.1. Expression quantitative trait loci (eQTL) studies	52
1.2.4.4.2. Allele-specific expression (ASE) studies	53
1.2.4.5. Public genomic functional interpretation databases	53
1.3. The study of type 2 diabetes and the relevance of pancreatic islets	54
1.3.1. Metabolic pathophysiology	54
1.3.2. Genetics	55
1.3.3. Environmental factors	56
1.3.4. Epidemiology and Treatments	56
1.3.5. Disease heterogeneity	57
2. Hypothesis and Objectives	58
3. Report from the director	62
3.1. Epistasis (Unpublished)	64
3.1.1. Title	64
3.1.2. Authors	64
3.1.3. Contribution	64
3.2. TIGER publication	64
3.2.1. Title	64
3.2.2. Authors	64
3.2.3. Journal	65
3.2.4. Contribution	65
3.3. Polymorphic Inversions publication	66
3.3.1. Title	66
3.3.2. Authors	66
3.3.3. Journal	66
3.3.4. Contribution	66
3.4. Genome Wide Association Studies review	66
3.4.1. Title	66
3.4.2. Authors	67
3.4.3. Journal	67
3.4.4. Contribution	67
4. Summary of the studies	68
4.1. Epistasis (Unpublished)	70
4.2. TIGER publication	71
4.3. Polymorphic Inversions publication	72

4.4. GWAS review	73
5. Epistasis (Unpublished)	76
Tables and Figures list	78
Figures	78
Tables	78
Abstract	79
Introduction	80
Results	82
Overall strategy	82
Measuring the effect of epistasis in T2D	85
The epistatic variants functional impact and its association with T2D	86
Discussion	93
Methods	95
Discovery dataset	95
T2D case-control dataset	95
Dataset preparation	95
Machine learning approaches	95
Method selection	95
Algorithm preparation	96
Data imbalance	96
Randomness assessment	96
Maximisation of variables explanation	97
Variables redundancy	97
Missingness	97
Data availability	98
Hyperparameters adjustment	98
Genomic inspection of the results	99
Model outcomes interpretation	99
Candidate epistatic groups base genomics	99
Genomic, transcriptomic, and epigenetic functional assessment	100
Resources	100
Annotations overlap	100
Functional annotations enrichment	100
Statistical assessment	101
Logistic regression epistasis	101
Supplemental Materials	102
6. TIGER publication	122
7. Global Results and discussion	198

7.1. Epistasis	200
7.2. TIGER	203
8. Conclusions	208
8.1. Epistasis	210
8.2. TIGER	210
9. References	212
10. Supplemental Material	228
11. Appendix: Polymorphic Inversions Publication	240
12. Appendix: Genome Wide Association Studies review publication	256

TABLES AND FIGURES LIST

Tables and Figures list

Figures

Figure 1. The structure of the DNA molecule.

Figure 2. The genetic variability of the zygote.

Figure 3. Classification of variants by length and presence in the population.

Figure 4. Disease and variants classification based on the effect and behaviour of disease-related variation.

Figure 5. Central dogma of biology.

Figure 6. GWAS schema.

Figure 7. ML overfitting problem.

Figure 8. ML algorithm pipeline.

Figure 9. Evaluation of the results of the predictions made by a ML binary classifier.

Figure 10. Gene expression regulatory process schema.

Figure 11. TIGER data portal.

Tables

Table 1. Genes involved in the glucose uptake and IS process.

Supplemental Materials

Supplemental Figure 1. Machine Learning algorithms based on the type of problem to be solved.

Supplemental Table 1. Supervised Machine Learning classifiers and dimensionality reduction techniques applied in this thesis.

Supplemental Table 2. Population haplotype reference panels.

Supplemental Table 3. Data types.

Supplemental Table 4. Machine Learning models based on the learning type.

Supplemental Table 5. ML binary classifiers effectiveness and reliability measures.

Supplemental Table 6. Some popular publicly available databases with functional information.

Supplemental Table 7. Main organs dysfunction that can derive T2D.

ABBREVIATIONS LIST

Abbreviations list

1000G - 1,000 Genomes Project
A - Adenine
AdaBoost - Adaptive Boosting
Amplified antisense technologies - aRNA
ASE - Allele-specific expression
ATAC-seq - Assays for transposase-accessible chromatin sequencing
AUROC - Area Under the ROC curve
bmi - body-mass index
bp - Base pair
C - Cytosine
ChIP-seq - Chromatin Immunoprecipitation followed by sequencing
CV - Cross-Validation
DT - Decision Tree
eQTL - Expression Quantitative Trait Loci
EWAS - Environment-Wide Association Studies
FFA - Free fatty acid
FN - False Negative
FP - False Positive
G - Guanine
GoNI - Genome of the Netherlands
GTEx - Genotype-Tissue Expression database
GWAS - Genome-Wide Association Studies
GxE - Gene-Environment interactions
GxG - Gene-Gene interactions
Het - Heterozygous
hg - Human Genome
HGP - Human Genome Project
HGSC - Human Genome Sequencing Consortium
Hom - Homozygous
HRC - Haplotype Reference Consortium
IIS - Impaired insulin secretion
Indels - Short insertions and short deletions
IR - Insulin resistance
IS - Insulin secretion
KNN - K-Nearest Neighbour
LADA - Latent Autoimmune Diabetes in Adults
LD - Linkage Disequilibrium
LDA - Linear Discriminant Analysis
lncRNAs - Long non-coding RNAs
LR - Logistic Regression
MAF - Minor Allele Frequency
MCC - Matthews Correlation Coefficient
ML - Machine Learning
MODY - Maturity Onset Diabetes of the Young
mRNA - Messenger RNA
NDM - Neonatal Diabetes Mellitus
NN - Neural Network
OR - Odds ratio
PCA - Principal Component Analysis
PCR - Polymerase chain reaction

PRS - Polygenic Risk Scores
QDA - Quadratic Discriminant Analysis
RF - Random Forest
RFE - Recursive Feature Elimination
RNA-seq - RNA sequencing
ROC curve - Receiver Operating Characteristic curve
SGD - Stochastic Gradient Descent
sncRNAs - Small non-coding RNAs
SNP - Single Nucleotide Polymorphism
SNV - Single Nucleotide Variant
SV - Structural Variant
SVM - Support Vector Machine
T - Thymine
TF - Transcription Factor
TN - True Negative
TP - True Positive
TPR - True Positive Rate
TSS - Transcription Start Site
UKB - UK Biobank
WGS - Whole Genome Sequencing
XGBoost - Extreme Gradient Boosting

INTRODUCTION

1. Introduction

1.1. Biomedicine and the study of human diseases

1.1.1. Motivation: The study of the genetic basis of human diseases

The understanding of the biological mechanisms that affect the risk of developing a disease, and its preservation through different generations, has been a subject of study broadly approached during the last century by the **Biomedicine** field, and in particular during the last two decades (Quirke & Gaudillière, 2008). In this direction, different analytical frames and strategies have been designed and applied to improve the comprehension of human diseases, combining different disciplines. Particularly, during the last decades, the computational and technological revolutions have enhanced these studies by providing more sophisticated tools, and computational and analytical methods, to facilitate and support these complex analyses. The broad use of these technological advances have boosted the generation of a large volume of diverse types of data to study human genetics, making it necessary to improve information data storage techniques, management, integration, distribution, and analytical tools. Thus, leading to the creation of new specialised fields, such as **Bioinformatics**, a multidisciplinary field, which focuses on the creation and use of computational and statistical frameworks to analyse and interpret multi-omics biological data.

One of the main goals of the Bioinformatics field, specifically from the **Computational Genomics** point of view, is to study disease heritability by deciphering the contribution of genomic variation on the risk of developing a disease, and understanding how much of this genomic variation can be inherited by the offspring generation. **Heritability** is the common term used in the Computational genomics field to refer to the study of the estimated variance of a trait or disease that can be exclusively explained from the genetic point of view. In this direction, during the last decades, several efforts have been made to reach a better understanding of the **genetic basis of diseases**. Notably, the better comprehension of the human genetic architecture, and the identification and characterization of genetic markers, has been essential to improve the prediction of disease risk and diagnosis. As a result, the elucidated conclusions from the different types of analyses conducted in the Computational Genomics field can be, in such a way, translated to the clinics to early **detect, prevent** and, ultimately, to **treat diseases**. More specifically, all this knowledge has revealed some of the molecular and functional basis of human disorders, thus leading to the generation of better detection protocols, and becoming crucial in the development of new treatments (Timpson, Greenwood, Soranzo, Lawson, & Richards, 2018).

1.1.2. Fundamentals of genetics and genetic inheritance

Although the use of statistical and computational tools to perform the analyses conducted in the Bioinformatics field do not require any biological previous knowledge, a good insight in the biological basis of genetics is crucial for the preparation of the analyses, as well as for the interpretation of the results obtained. Therefore, the comprehensive biological knowledge of how conditions are transferred from one generation to the next, named **genetic inheritance patterns**, facilitates the understanding of genetic studies.

In this direction, in 1865, the fundamentals for explaining the basis of this genetic inheritance were firstly described by Mendel (S. Abbott & Fairbanks, 2016). Particularly, Mendel studies were focused on the hybridisation of 34 varieties of peas presenting clear observable differences in various traits of the plants. As a result from eight years of experiments, seven hybrid characters, such as the difference in the form and colour of the seeds, were observed in the first, second, and subsequent generations. The analysis of the prevalence or recession from each of these hybrid particularities lead to the classification of these features between **dominant**, in case they prevail, or **recessive** in case of

remission or loss across the generations. Consequently, since that point, **mutations** were defined as cell permanent and temporary associations, which follow the different **Mendelian inheritance patterns**, based on the predominance and recession of the character and its appearance in each generation (Alliance, Screening, & Services, 2009).

Notably, the advances of Mendel's work to understand the genetic inheritance were related to changes and mutations in the fertilised cell, but the principal component of the cell susceptible to mutations was still missing. For this reason, many theoretical studies suggested that changes in **genes** or **proteins**, which were known to be functional elements in the cell, could be leading to the generation of the different traits. However, it was not until 1943 when Oswald Avery experimentally proved that the sodium deoxyribonucleate or **DNA**, stored in the nucleus of the cell, was the main responsible of the genetic differences or mutations (Avery, Macleod, & McCarty, 1944; Cobb, 2014). This result was revealed from observing and analysing the transformation of specific types of cultured Pneumococcus, and led to the conclusion that genes were made of DNA instead of proteins, as it was previously thought. Moreover, since that point, the DNA sequence was defined as the 'transforming principle' and, consequently, it was stated that any chemical **DNA alteration** was the cause of different cell types and biological functions, therefore making them predictable, and transmissible in series.

The complete understanding of what is a DNA alteration requires a better comprehension of the DNA molecule. Fortunately, at that point, **the structure of the DNA molecule** had already been theoretically defined by Phoebus Levene in 1919 (Levene et al., 1919). Thus, facilitating the experimental validation of the **tetranucleotide theory**, and the confirmation of the nucleic acid as a **paired sequence**, where each base pair (bp) was generated from the combination of four nitrogenous bases: adenine (A), thymine (T), cytosine (C) and guanine (G). Each of these bases, as published by Erwin **Chargaff** in 1950, based on his chromatography studies of the DNA (Chargaff, 1950), pair in the DNA sequences as follows A-T and C-G, following the **rules** which now receive his name (**Figure 1**).

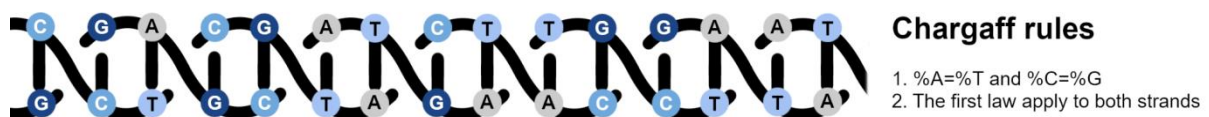


Figure 1. The structure of the DNA molecule: The DNA molecule is composed by two right-handed helical paired sequences, from which each base pair is a nitrogenous base: adenine (A), thymine (T), cytosine (C) and guanine (G). These bases pair following the Chargaff rules, which state that 1) A pair with T, and C pair with G, so that the proportion of A must be the same as the proportion of T, and the proportion of C must be the same as the proportion of G, and 2) Each DNA strand follows the first rule.

Moreover, in 1953 (Watson & Crick, 1953), Rosalind Franklin, James Watson, and Francis Crick defined the DNA structure to be composed by **two right-handed helical chains** coiled in the same axis but in the opposite direction, where each of the helix pairs connect through hydrogen bonds following the Chargaff rules. Therefore, as Avery explained, any alteration of the DNA molecule can result in differences in the biological behaviour of the nucleic acids in each of these chains, and consequently lead to an observable trait.

As a result of all these discoveries, today we know that DNA is a molecule defined by a paired sequence of nitrogenous bases, which contains the basic information for each cell type. Thus, being responsible for the different cell specialised functions. For this reason, any chemical alteration on the nitrogenous bases of its sequences can lead to a biologically functional transformation that can be observed as a trait characterising an individual, or a disease. Particularly, the variation leading to a trait or a disease can be inherited by the offspring. Consequently, these changes or mutations are

predictable and, therefore, their genetic study can lead to a better understanding of diseases, and the improvement of disease prognosis.

1.1.3. DNA alterations and inheritance patterns

The discovery of the DNA molecule and its relevance to explain diseases and traits, based on its transformation, motivates the study of the genomic mechanisms from which mutations can be acquired during the individual life. Particularly, to study how genomic variation can be transmitted to the next generations it is fundamental to remember that the DNA is organized in **23 homologous chromosomes**, from which the first 22 are the **autosomes**, and the other one, the **sexual chromosome**, defines the sex of the individual (X and Y). Moreover, any position in the human genome (**locus**) has two copies (or **alleles**): one inherited from the father *A* and one inherited from the mother *B*. Thus, humans are **diploid** organisms, and each chromosome has two identical haploid copies. These copies are named **sister chromatids**, join in a genomic region named **centromere**, and end in non-coding and highly repetitive regions named **telomeres**, which provide their structural stability. Since humans are multicellular organisms, this DNA organisation is preserved among all the different human cells. Particularly, a human being starts its existence with only one cell (**zygote**) but can reach over 30 trillion (3×10^{13}) of specialised cells when adults (A. Abbott, 2016; Sender, Fuchs, & Milo, 2016). As a result, **cell mutations** can be potentially acquired when the DNA molecule stored in its nucleus is exposed to transformations, which corresponds, depending on the stage of the cell, to the moment when the zygote is created (**germline mutation**), or when a copy of a new cell is generated (**somatic mutation**).

The DNA alterations occur during the division processes of the cell: **mitosis** (somatic) and **meiosis** (germline). However, the meiosis process is crucial to understand how mutations are inherited by the offspring and, therefore, for the study of the genetic inheritance of human disorders. Meiosis is a three step process (**Figure 2**) where:

- 1) First, each parental DNA haploid sequence is complemented with its corresponding chromatid sister.
- 2) Then, the maternal and paternal chromatids combine, during the meiotic **homologous recombination** step.
- 3) After the recombination, the chromatid sisters separate their centromeric regions, thus generating 4 alleles named **germ cells** (or gametes).

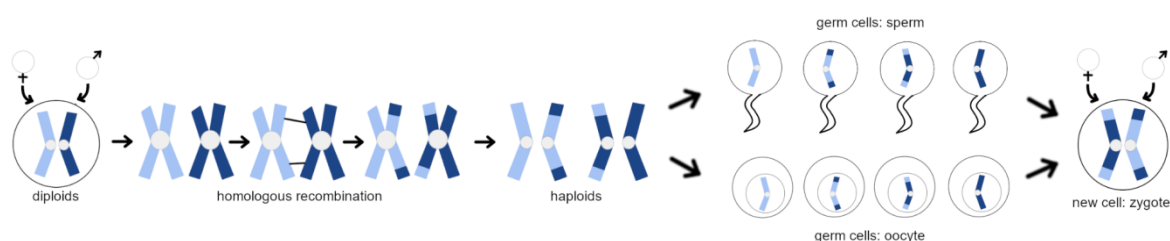


Figure 2. The genetic variability of the zygote. During the meiosis process, in diploid organisms, the DNA recombines generating four gametes that will constitute the DNA of haploid germ cells. The genetic material of two germ cells is combined during fertilisation to generate a new cell in the offspring (zygote) that inherits the variability present in the parental cells.

The combination of one maternal and one paternal gamete (**fertilisation**) leads to the generation of a new cell (zygote) (Burton, Tobin, & Hopper, 2005). As a consequence, the DNA content of the new cell inherits the variability already present in the parental germ cells and *de novo* changes occurring during the meiotic process. Each of these changes, already existing or *de novo* generated, are usually referred to as **germline genomic variation**. This type of variability, which can derive in different traits and/or diseases, represents the baseline susceptibility in complex diseases.

Particularly, during the homologous recombination, each parental DNA molecule is divided in fragments, called **linkage disequilibrium** (LD) regions or LD blocks (Slatkin, 2008). This fragmentation usually occurs in the chromosome recombination **hot spots**, which are regions of the genome more susceptible to be fragmented and recombined. As a result, LD regions contain groups of variants with a higher probability to be preserved as a block for each parent. Therefore, a group of alleles that are inherited together from a single parent is the **haplotype**, and since the combination of alleles in all loci defines the individual **genotype**, genomic variation can be studied based on the observation of the individual genotype. Indeed, based on human diploidism, each locus can have 3 possible combinations of alleles *AA*, *AB*, *BB* that can be grouped in **homozygous** (*hom.*), if the alleles are the same, and **heterozygous** (*het.*) if the alleles are different. Specifically, given a reference, for example a non-mutated cell in a population, compared to a mutated cell in the same population, these combinations can turn into **homozygous reference** if both nucleotides are the same and match the reference, **homozygous alternate** (*hom. alt.* or *BB*) if both nucleotides are the same but mismatch the reference, or heterozygous (*het.* or *AB*) when both nucleotides are different at that position, one matching and one mismatching the reference allele. Moreover, the way the alleles *A*, *B* are inherited (**inheritance patterns**), which define the different **inheritance models** (**additive**, **recessive**, **dominant**, or **heterodominant**), can lead to a different effect on the individual phenotype (Alliance et al., 2009). As a result, the evaluation of predisposition to diseases through the study of germline variant genotypes is affected by multiple factors such as LD regions, or inheritance models. Based on these factors, the study of the genotypes of germline variants, and its probability of being inherited in the offspring, can help to gain insight into the effect of genomic variation in the different traits or disorders, and to facilitate the early detection and prevention of diseases.

1.1.4. Genetic variants classification

The relevance of the study of genomic variation inside a population motivates the classification of variants to reduce the complexity of the explanation of genetic studies outcomes. Hence, variants can be categorised, in general, by their size and their presence in the population, measured by their **minor allele frequency** (MAF). As a result, in terms of their length, variants with only one nucleotide change are referred as **Single Nucleotide Variants** (SNVs), those involving a deletion or insertion between one and 50 nucleotides are **short Indels**, and the rest of genomic variants are referred as **Structural Variants** (SVs) (Escaramís, Docampo, & Rabionet, 2015). In particular, SVs are DNA regions presenting a change in copy number (**deletions**, **insertions**, **duplications**, or **copy number variation**), orientation (**inversions**), or chromosomal location (**translocations**) (**Figure 3.A**). Moreover, looking at the genomic variant frequency among the population, variants with a $MAF < 1\%$ are referred to as **rare variants**, those with a presence $1\% \leq MAF < 5\%$ are called **low-frequency variants**, and the rest ($MAF \geq 5\%$) are known as **common variants** (Bomba, Walter, & Soranzo, 2017; Eichler, 2019; Ku, Loy, Salim, Pawitan, & Chia, 2010; The International HapMap Consortium, 2005) (**Figure 3.B**). Finally, being the most common type of human variation, both low-frequency and common SNVs are known as **Single Nucleotide Polymorphisms** (SNPs). The use of this classification facilitates the interpretation of genomic population studies, improves the characterisation of disease-related variants, and promotes a better understanding of the genetic basis of human disorders.

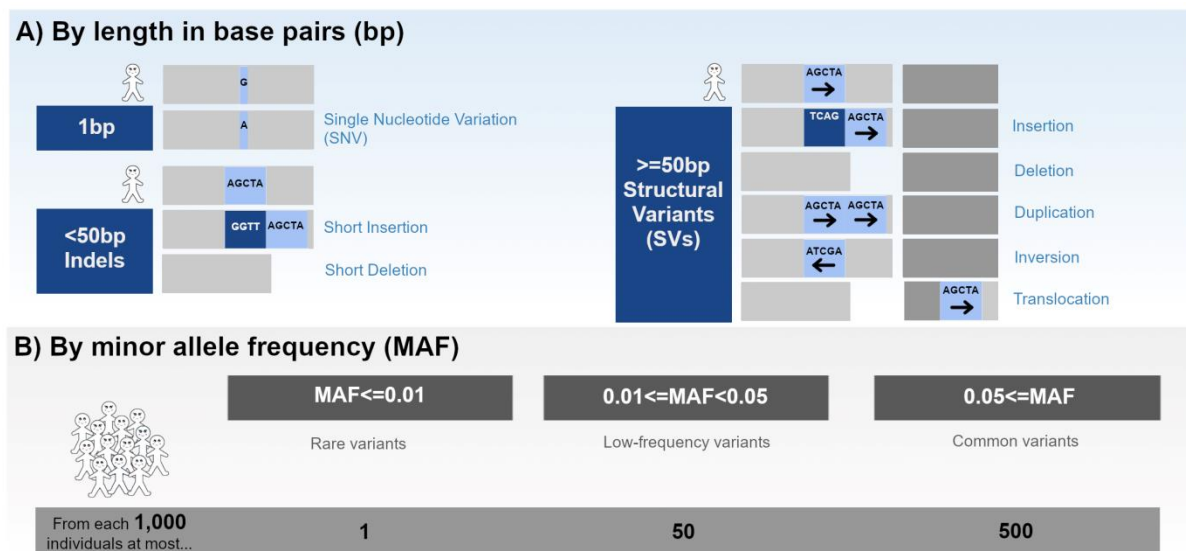


Figure 3. Classification of variants by length and presence in the population. Genetic variants can be classified by their length, and by their frequency among the population of study. The panels display: A) Variant classification by length in base pairs (bp). The original sequence is the one with a stickman on the left, and each mutated sequence is displayed below the original sequence. The light grey boxes represent the original chromosome position in the genome for the observed variant (light blue), and the dark grey boxes represent a different region of the genome. The dark blue boxes represent new inserted nucleotide sequences. B) Variant classification by the minor allele frequency (MAF).

1.1.5. The effect of genetic variants and disease characterization

The fact that different DNA alterations can result in observable traits and diseases, motivates the establishment of a categorisation of variants based on their effect on a specific phenotype. Particularly, in terms of relation with disease, germline variants can be classified as **protective**, if their contribution to the risk of developing a disease is negative, **risk** variants if they contribute positively to that risk, or **neutral** if the variant has a negligible or no effect on the phenotype of study. Consequently, in terms of this genomic association with a disease of study, protective variants reduce the individual predisposition to develop the disease, and risk variants increase this predisposition. Moreover, despite the large variety of traits that can be observed in the human population worldwide, from the genomic point of view, these attributes or diseases can only be characterised as **monogenic** or **polygenic** depending on the number of genomic variants affecting the individual phenotype, and their behaviour. Therefore, in a **Mendelian** or **monogenic** disorder, although diverse genes can be involved in its development, the effect of genomic variation in only one of these genes is enough to mediate the disease. In contrast, if the contribution of multiple genetic variants, affecting various genes simultaneously, and diverse environmental factors is needed to develop the diseased phenotype, then it is named **complex, polygenic, or common** disease (Figure 4) (Manolio, Brooks, & Collins, 2008).

A) Traits



B) Diseases

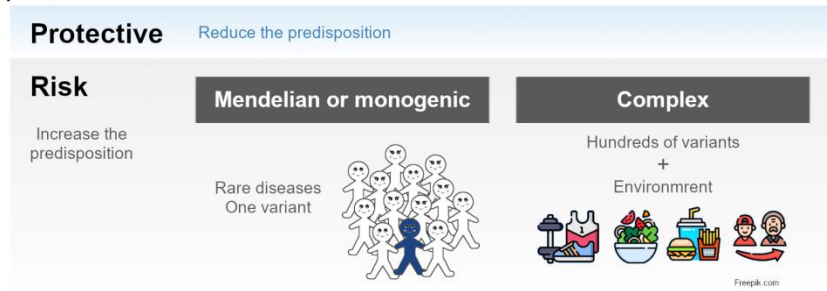


Figure 4. Disease and variants classification based on the effect and behaviour of disease-related variation. Genomic variants define the individual phenotype, thus leading to the development of A) different traits, and B) diseases. A disease susceptible variant is classified as protective (blue panel), if it helps to the prevention of the disease. In contrast, if the variant mediates the disease, it is defined as a risk variant (grey panel). In a Mendelian or monogenic disease, it is only necessary to have one variant affecting a particular gene, from the diverse genes related to the disease, to its development. In complex diseases, hundreds of variants with a low effect, affecting multiple genes simultaneously, in combination with diverse environmental factors, are needed to develop the disease.

In particular, monogenic diseases are usually rare diseases, which affect less than 5% of the population such as cystic fibrosis or polycystic kidney disease. In contrast, complex diseases, such as diabetes, asthma, or Alzheimer's disease, are broadly extended among the global population, usually affecting thousands, and even millions of individuals worldwide. Thus, converting the study of the genetic inheritance of complex diseases into one of the major goals of Biomedicine. Particularly, the better comprehension and characterisation of this genetic component, the more we will know about the molecular biology behind, and the better chances to design improved prognosis, prevention, and treatment protocols for this type of disorders.

1.1.6. From genomic variation to its functional interpretation

The study of the genetic inheritance of complex diseases, based on the analysis of the contribution of germline variation to the disease, is fundamental to find the genomic mechanisms underlying this type of disorders. However, apart from the discovery of disease-associated genetic markers, it is also necessary to understand their molecular mechanisms, since it is essential for the identification of drug targets and new therapies. Nevertheless, further knowledge is needed to convert DNA alterations into functional alterations that could explain the disease. In this direction, it is crucial to find the relation between germline variation and **cell function**. Particularly, as cell function derives from **proteins**, and **genes** are DNA segments containing the instructions for protein production, they are of special interest for the study of the effect of genomic variation on function.

The relation between genomic variation, genes, proteins, and function can be explained through the **central dogma of biology**, which constitutes the basis of molecular biology, and was published by Francis Crick in 1958 (Crick, 1958). This dogma stands on the fact that the DNA molecule is continuously transcribed into **RNA**, which then will be further translated into proteins. During the transcription, each DNA strand is copied to generate a RNA strand, transforming thymine in uracil (U). Then, the RNA is translated into **amino acids**, which are groups of three bases, to generate the different proteins that are involved in the diversity of cell functions. Each protein is composed from at least 20 amino acids (**Figure 5.A**). Therefore, DNA alterations associated with a complex phenotype can result in a change of a gene which can alter the protein function (**Figure 5.B**).

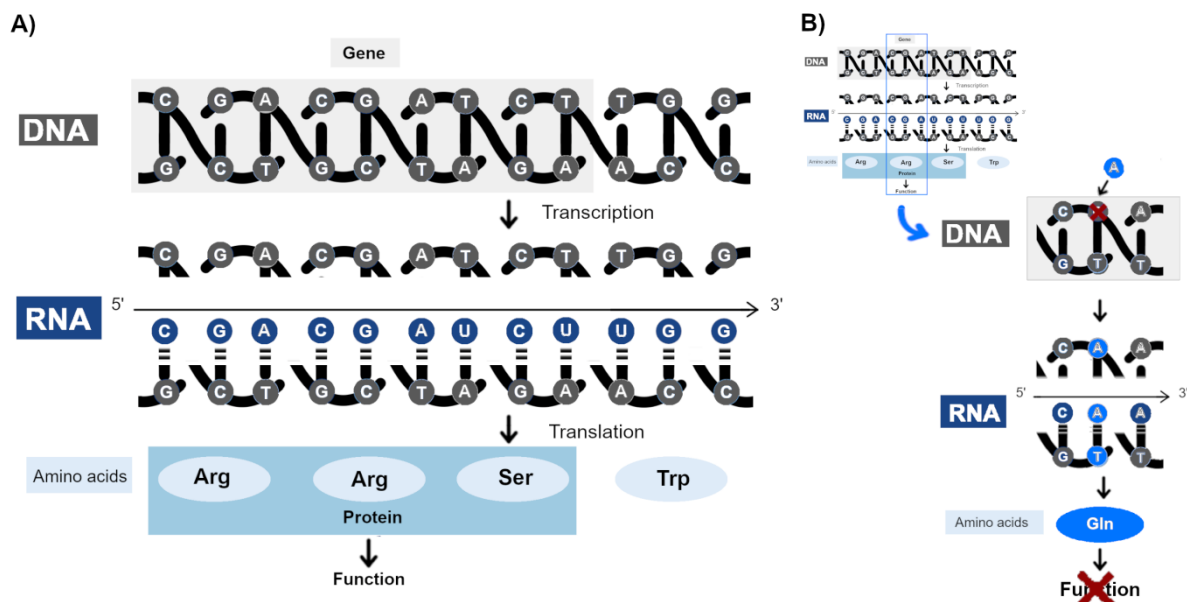


Figure 5. Central dogma of biology. The DNA molecule (top) is transcribed into RNA (middle) to then be translated into an amino acids sequence (bottom).

A) The genes, which contain the instructions for protein production in the DNA, result in groups of at least 20 amino acids that generate a protein.

B) DNA alterations can be translated into an amino acid change, which alters the protein function.

As a result, the relation between variants and genes or functional regions enhances the detection of the pathways mediating the disease or its connection with the symptoms. For this reason, after the identification of disease-susceptibility loci, the use of genomic annotations to find candidate functional regions or putatively associated genes, targeted by the alterations, has been broadly extended. Moreover, humans have hundreds of different types of **specialised cells**, which multiply during the mitotic process, to facilitate the human development, growth, and regeneration (Arendt, 2008; Vickaryous & Hall, 2006). The specialisation of each of these cells enhances the performance of the specific tasks that each human organ or tissue requires to ensure its function. Thus, suggesting that the study of the deterioration of specific types of cells or disease-related tissues, based on the presence of genomic variation, can help to improve the explanation of the functional interconnections between genetic variability and common diseases.

1.1.7. Preparation of different omic data for genetic studies

The relevance of the DNA molecule to understand how a chemical transformation of any of its chains relates to a different trait or disease, how frequently are those changes inherited through different generations, and which are the functional interconnections between this genomic variation and diseases, evidence the importance of exactly determining the DNA sequence of an individual. As a result, a wide diversity of experimental methods and computational tools has been developed during the last decades to allow the inspection of the DNA and RNA sequences. Thus, including methods which facilitate the analysis of specific positions, such as **microarrays**, or tools that determine the complete sequence of a region, such **sequencing technologies**. Particularly, the introduction of sequencing and microarray methodologies changed the paradigm enhancing the advance of genetic studies. In contrast with previous studies, which focused on the analysis of cell function to find the pathways related to disease, studies based on the use of these new methods start with the identification of variants associated with disease to, then, find a putative relation with cell function that mediate human disorders.

1.1.7.1. DNA sequencing

The analysis of the DNA sequence of an individual is crucial for the discovery of any possible alteration in its chains that can derive in the different phenotypes. Therefore, after many efforts, in 1977, the **Sanger sequencing method** emerged finding the way of determining an individual DNA sequence (Sanger, Nicklen, & Coulson, 1977). To improve the accuracy of the results obtained by applying this method, other complementary techniques, or correspondingly priming the opposite strand with the same process, were further recommended. Notably, the many advantages supposed by the simplicity of the performance of this methodology, the fewer artifact bands observed in the process, and the possibility to sequence between 15 to about 300 nucleotides from the priming site, enhanced its commercialisation process by Applied Biosystems in 1986. This commercialisation, and the continuous updates of DNA sequencing methods, promoted its broader use to gain insight in genomics, thus, including *de novo* assemblies of the genome, individuals resequencing, and other clinical and biochemical applications (Shendure et al., 2017).

1.1.7.2. DNA microarrays

The fact that changes on the DNA sequence can affect some specific cell functions, and lead to a diseased phenotype, evidenced the need of techniques to compare different cells or individuals' DNA. Despite many molecular biology based methods were developed to facilitate this analysis in a separate manner, it was not until 1983, when Tse Wen Chang published the basis to generate **DNA microarrays**, a method that allowed the simultaneous analysis of multiple cells (Tse-Wen Chang, 1983). This technology facilitates the comparison between the different tested cells, with lower reagent consumption, and minimising the test time. The many advantages of this method led to its rapid commercialisation by Affymetrix, Agilent, Applied Microarrays, Arrayjet, Illumina, and others. Moreover, the generalisation of this methodology led to diversifying the analyses performed, therefore extending them for example, to the analysis of gene expression levels, methylation, or alternative splicing (Gonzalo & Sánchez, 2018; Schena, Shalon, Davis, & Brown, 1995).

1.1.7.3. RNA sequencing

Despite the great success of DNA sequencing and microarrays, the expensive cost of Sanger sequencing and its difficulties to uniquely map to the genome, and the limitations of microarrays derived from the need of previous genome sequence knowledge, or the difficulties to precisely compare measures between independent experiments, motivated the introduction of **high throughput** and **Next Generation Sequencing methods** (Z. Wang, Gerstein, & Snyder, 2009). Particularly, in 2009, Zhong Wang presented the **RNA sequencing (RNA-seq)** methodology as an alternative to previous methods. This method facilitated the mapping and quantification of the transcriptome in reads with between 30 and 400 bases. RNA-seq technology presented many advantages in contrast with previous methods, such as the no need of a priori knowledge of the genomic sequence of a model organism, not having upper quantification limits, the reduction of background signals, more accuracy, and a lower cost. In contrast, some challenges and complications surrounded this method, such as many difficulties related to library constructions, the need for big storage, and requirement of new methodologies to process large amounts of data. Despite these disadvantages, the development of RNA-seq was revealed as fundamental, therefore enhancing the commercialisation of this methodology by several companies, such as Illumina, Qiagen and ThermoFisher Scientific. As a result, the analysis of RNA-seq was crucial to offer a global view of the transcriptome of various species, to revise gene annotation, to identify novel transcribed regions, to detect new splicing events, and to find sequence variations. Thus becoming essential for interpreting the functional elements of the genome, specific cells and tissues, and being promoted as the key for understanding development and disease.

1.1.7.4. Single-cell sequencing

Although the application of RNA-seq and expression microarrays on large numbers of cells to analyse complete expression profiles, and to understand how many and which genes are particularly expressed in a tissue or an organ resulted successful, the heterogeneity of the functions of the cells present in any tissue or organ limited the discovery to the average expression of the genes studied. These limitations evidenced the need of simultaneously analysing a diversity of cell types in more complex organisms, to calculate the expression in any single cell type. In that direction, Ernest Kawasaki proposed **single-cell sequencing** technology in 2004 (Kawasaki, 2004). This technology is based on collecting enough RNA for probe array production from a variety of cells that can be representative of a single cell population in a tissue or organ. The expected scientific advances from the use of this technology lead to its commercialisation by Fluidigm, Clontech, and 10xgenomics. Particularly, the use of single-cell sequencing has led to a better understanding of the biology of cells, and to gain insight in some diseases with highly heterogeneous tissue-related cells, such as cancer.

1.1.7.5. Genotyping arrays

The broad use of sequencing methods led to the easy characterization and identification of SNPs, and, as a consequence, to the development of **genotyping arrays** or **SNP arrays** in 1998 (D. G. Wang et al., 1998). Genotyping arrays facilitated the screening of SNP genotypes in a large-scale. Particularly, their broad commercialisation by different producers, such as Affymetrix, Agilent, Illumina and Niblegen, together with the knowledge about the existence of over 1.4 million SNPs (International Human Genome Sequencing Consortium, 2001), has facilitated the creation of genotyping arrays that can evaluate more than one million SNPs for thousands of individuals at the same time (Lamy, Grove, & Wiuf, 2011). Thus converting the use of genotyping array technology into a more economical and viable technique for the study of human genetics, and enhancing ancestry assessment, allele-specific expression studies, association with disease, and somatic changes detection (LaFramboise, 2009). Moreover, the use of SNP arrays can mediate the identification of genetic markers related to disease based on familial studies, the analysis of linkage disequilibrium in isolated populations, association analysis in case-control individuals, loss-of-heterozygosity studies, to measure genetic distances between populations, and can also be used for parental and pedigree assignment (Vignal, Milan, SanCristobal, & Eggen, 2002).

1.1.7.6. The evolution of sequencing and microarrays

In summary, sequencing and microarray processes focus on recovering the exact nucleotide sequence from a DNA molecule. This facilitates the identification of molecular elements with a potential relation with disease, and improves the knowledge of the biological functions of different cells. As a result of the success of the broad use and commercialisation of these technologies, all of them have been continuously evolving since their presentation. As a result, the quality of the material produced has improved, for example increasing the number of base pairs obtained in a sequence, and including the information of thousands of cells for single-cell analyses (Shendure et al., 2017). Moreover, the sequencing cost has been reduced from 3 billion dollars to less than 1 thousand dollars to obtain a complete individual DNA sequence ('The Cost of Sequencing a Human Genome', 2021). All these improvements have converted the use of these technologies into something fundamental for the advance on the genomic study of complex diseases.

1.2. Genetic studies and complex diseases

The great advances promoted by the development of experimental methods and computational tools in the Biomedicine field have enhanced the genomic study of diseases. Particularly, the introduction of these new technologies has represented a change in the paradigm of genomic studies, thus, facilitating the simultaneous analysis of multiple individuals to **find variants associated with complex disorders**, which can be further analysed to **understand their functional**

implications in disease predisposition. In this direction, the broad use of these methods by large Consortia has resulted in a big progress where many milestones have been achieved. These advances include the procurement of **the first assembly of the Human reference Genome**, which has been of great relevance for the discovery of disease-susceptibility loci, but also to improve their functional interpretation. Moreover, these technologies have allowed the generation of **genetic variability maps**, which have played a key role in the study of population variability, and have improved the detection of **disease-associated signals**. These achievements have represented an enormous progress in disease comprehension. However, at the same time, all these subjects are still nowadays a matter of study, discussion, and improvement, thus representing the starting point, and a solid basis for most of the current genomic studies.

1.2.1. The Human Genome Project and the human genome sequence

The development of sequencing technology and its commercialisation led the International Human Genome Sequencing Consortium (HGSC), in 1990, to announce **the Human Genome Project (HGP)**, which had the global goal of obtaining **the first assembly of the Human reference Genome sequence** (Venter et al., 1998). As a result of this large Consortia effort, by the nearly end of 2004 the project was finished with approximately covering 99% of the euchromatin genome. That corresponds to 2.85 billion (2.85×10^9) paired bases of the human genome (International Human Genome Sequencing Consortium, 2004). The inspection from the accurate sequence obtained (10x) showed that approximately 5.3% of the euchromatic genome are segmental duplications, and it contains more than 1.4 million SNPs (which occur at a rate of 1 per 1,300 bases). This valuable information was fundamental for the creation of new diagnostic tests based on the SNP association with diseases or traits. Moreover, a gene catalogue of 22,287 gene loci (34,214 transcripts, 19,438 known genes, 2,188 predicted genes, and an estimate of 20,000–25,000 protein-coding genes), and a list of transposable elements, GC content, and CpG islands were generated, thus, constituting a comprehensive human genomic database, and providing the scientific community with a great resource of functional information.

Different **assemblies** of the hg correcting previous errors have been released till now. Particularly, the last more known and broadly accepted by the scientific community are GRCh37.p13 or **hg19** from 2013, and GRCh38.p13 or **hg38** from 2019. However, these versions are still missing the remaining 8% of the genome. Thanks to the advances made by PacBio and Oxford nanopore sequencing technology (Eid et al., 2009; Jain et al., 2018), as well as the new developments in assembly, polishing, and validation, the Telomere-to-Telomere (T2T) Consortium announced the release of a new version of the hg sequence (**T2T-CHM13v1.1 assembly**) addressing the remaining gaps (The Telomere-to-Telomere Consortium, 2022). As a result, 3.055 billion (3.055×10^9) bp sequence of the hg are now known, including pericentromeric and subtelomeric regions, novel genes and segmental duplications, ampliconic gene arrays, ribosomal DNA (rDNA) arrays, the X chromosome, and 16,569 bp of mitochondrial genome.

The broad use of the different human reference genome assemblies, as well as the large genetic databases generated by the HGSC and the T2T Consortium, has benefited multiple genetic studies. Particularly, the discovery of a list with more than 1.4 million SNPs has facilitated the improvement of genotyping arrays, thus enhancing the detection of disease-associated loci. Moreover, the creation of a genes catalogue has promoted the functional interpretation of the disease-susceptibility loci in terms of gene function. Additionally, the generation of a database of functional regions has enhanced the analysis to find putative relations between genomic variation and disease regulatory mechanisms. All this knowledge has derived in a better understanding of the biology behind the human genome, and multiple benefits for human health. For this reason, as the current assembly is monoploid, meaning that it is based on only one human haplotype, in 2021 a new initiative from the Human Pangenome Reference Consortium raised to sequence 350 genomes with

the aim of properly capturing the genomic diversity in human population (Miga & Wang, 2021; Reardon, 2021).

1.2.2. Genetic variability maps

The extensive use of genotyping arrays for the study of the genomic variation across different populations has promoted the achievement of very relevant milestones, such as the development of **large population genetic variability maps**. These **haplotype reference panels** contain the haplotype of thousands of individuals evaluated in different loci, thus procuring a valuable source of information for downstream genomic analyses. Particularly, population maps are crucial to find differences between individuals from the same population, and to compare the genetic variability between different ancestries. Thus, facilitating for example the understanding of how those differences can affect disease predisposition or protection.

To generate the first haplotype reference panel, the International HapMap Consortium set out in 2002 **the International HapMap Project** (The International HapMap Consortium, 2003). This project aimed to genotype at least one common SNP every 5 Kilobases in euchromatic regions in 270 individuals from four different ancestries in Africa (Yoruba), Asia (China and Japan), and Europe (Utah). As a result, in the **Phase I** of the project, approximately 1.3 million SNPs were genotyped. In the **Phase II**, published in 2007, a further 2.1 million SNPs were successfully genotyped on the same individuals, finding one SNP every 1 kb (The International HapMap Consortium, 2007). The resounding success of the HapMap study was followed by diverse initiatives aiming to extend the discovery of genetic markers in different populations, and to provide a deeper characterization of those genetic markers in the population. These projects involved the inclusion of larger sample sizes in the analyses, the incorporation of much lower frequency variants, the analysis of single-populations, and the combination of different sequencing techniques (whole-genome sequencing (WGS) and whole-exome sequencing (WES)) with genotyping array data. The success of these initiatives required its promotion by large consortiums such as **the 1,000 Genomes Project Consortium** (1000G) (The 1000 Genomes Project Consortium, 2015), **the Genome of the Netherlands Consortium** (GoNI) (The Genome of the Netherlands Consortium, 2014), **the UK10K Consortium** (The UK10K Consortium, 2015), **the Haplotype Reference Consortium** (HRC) (The Haplotype Reference Consortium, 2016), and **the TopMed** program (Taliun et al., 2021) (**Suppl. Table 1**).

As a result of all these efforts, it is known that more than 99.9% of the bases in a human single cell are shared in all people. Therefore, the genomic differences presented by the comparison of an individual genome with the reference comprehend between 4.1-5.0 million sites (The 1000 Genomes Project Consortium, 2015). However, it has been estimated that, in the world's human population, about 10 million sites vary such that both alleles are observed at a frequency of $\geq 1\%$, thus constituting 90% of the variation in the population (The International HapMap Consortium, 2003). Interestingly, after the alignment with the reference genome, more than 400 million variants, including SNPs and short Indels, have been lastly reported (Taliun et al., 2021).

1.2.3. Discovery of variants associated with complex diseases

Common diseases are broadly extended among the worldwide population, affecting between thousands and millions of individuals, thus converting their genetic study into a major health problem. Nevertheless, the fact that complex diseases are the consequence of the combination of multiple genetic and environmental factors (Manolio et al., 2008), has complicated their study, as well as their underlying biological understanding (Craig, 2008; Mitchell, 2012). First, the genetic component of complex diseases is affected by the contribution of the small effects of multiple genomic variants, thus defining its polygenic nature (McCarthy et al., 2008). Particularly, the heritability of most complex diseases has been estimated between 20-80%. However, still only a small fraction of this estimation

has been already recapitulated, thus constituting the **missing heritability problem** (Manolio et al., 2009). In addition to this complex genetic nature of common diseases, the multiple environmental factors affecting the disease such as clinical variables, or population structure obscure their analysis. Thus, converting the discovery of variants associated with complex diseases into a still challenging computational problem, which demands robust statistical models such as those underlying **Genome Wide Association Studies (GWAS)** or **Machine Learning (ML)** approaches.

1.2.3.1. Genome Wide Association Studies (GWAS)

To address the study of the genetic inheritance of complex diseases, **Genome-Wide Association Studies (GWAS)** have been broadly applied during almost the last two decades (R. J. Klein et al., 2005). In short, this study seeks disease-associated variants, as those that are significantly more (or less) present in patients, compared with control non-diseased individuals (**Figure 6**). Therefore, it is common to start from genotyping array data to evaluate millions of variants, simultaneously, to find possible genotype-phenotype associations. Particularly, GWAS are statistical approaches which analyse the genotype of thousands of individuals from the population of study (**cohort**) in search for disease association.

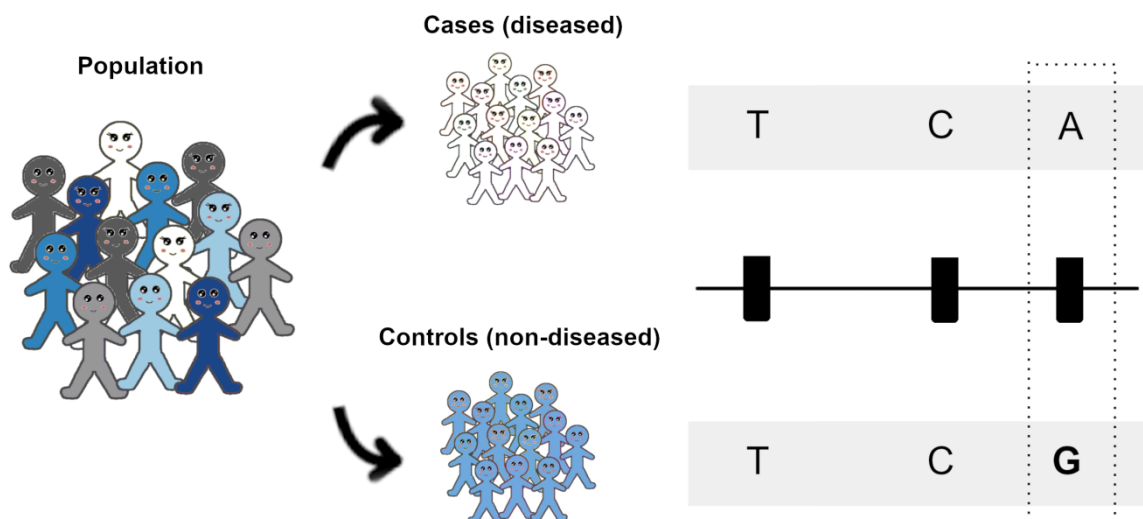


Figure 6. GWAS schema. In a binary GWAS, thousands of individuals from the population of study are first classified in diseased (white stickmen) and non-diseased (blue stickmen). Then, their genotypes are statistically compared. Each variant is tested in a single independent manner to find putative associations with the disease based on the comparison of its allelic frequency among the case-control populations (dotted square).

The diversity of GWAS models facilitates the analysis of these associations with quantitative and qualitative measures, which define the phenotypes of complex traits or diseases (**see Genome Wide Association Studies review**). GWAS involves the use of contingency tables, logistic regression, regression model extensions, and Bayesian regression approaches. All these methods test the association in a single independent manner for each variant included in the analysis. For this reason, the outcomes obtained from a GWAS, also named **summary statistics**, include, for each variant, a **multiple testing corrected p-value** standing from the association test with the disease, and the corresponding measure of its **effect (odds ratio (OR) or beta)** on the risk of developing the disease.

The combination of the success of GWAS and the general interest for its applicability in the study of complex diseases, has led to the development of several tools to enhance, improve, and facilitate the performance of this method (Uffelmann et al., 2021). Consequently, this methodology has been broadly applied to analyse the effect of genomic variation on a wide diversity of complex traits and common diseases, thus promoting the discovery of thousands of variants significantly

associated with the trait or disease inspected, where each variant contributes with a small fraction to the explanation of the risk to develop disease (McCarthy et al., 2008). As a result of GWAS success, **large catalogues of variants associated with complex diseases** have been created and made publicly available (Beck, Hastings, Gollapudi, Free, & Brookes, 2014; Buniello et al., 2019; K. Watanabe et al., 2019). Hence, providing the research community with a great resource of information that includes the association results for more than 276 thousand variants associated with more than 4 thousand traits or diseases (Buniello et al., 2019).

1.2.3.2. GWAS limitations

Despite the vast contribution of GWAS to the characterization of complex diseases, there are many limitations surrounding this methodology (Génin, 2020; Tam et al., 2019; Visscher et al., 2017; Wray et al., 2013). The diversity of factors contributing to GWAS limitations are enclosed in each of the steps involved in this type of study, thus including the input data, the statistical methods, and the results obtained from the analyses. However, the way all those factors affect the discovery encompass problems mostly related to the **statistical power**, and the challenges surrounding the inclusion of **complex association models** in the study. Moreover, the results obtained from these methods lack of **functional interpretation**, thus defining the boundaries for the understanding of the molecular mechanisms underlying disease (**see Genome Wide Association Studies review**).

1.2.3.2.1. Statistical power

There are many causes that affect the **statistical power** of a GWAS to find a significant disease-susceptibility loci association. These factors include the **allelic frequency** of the variant on the trait of study, the **sample size**, the **number of variants** that are included in the analysis, the underlying **genetic model**, the **genetic heterogeneity** of the trait in the population of study, and the **variability present inside the population**.

The power to detect a disease-associated locus is usually related to the effect of the variant on the trait of study. More precisely, variants with a higher effect on the disease are easier to capture. However, their **allelic frequency** tends to be lower in the population, usually in an inverse correlation with their effect (McCarthy et al., 2008). Therefore, favouring the detection of common variants, with usually a modest effect (OR between 1.05-1.3), than low-frequency variants or rare variants (Tam et al., 2019). In the same manner, this reasoning also applies to the higher detection of SNPs and short Indels, in contrast with SVs, which tend to be underrepresented in GWAS (**see Polymorphic Inversions and TIGER publications**).

To overcome the detection power limitation, based on the **sample size**, different approaches such as the analysis of **larger sample sizes**, **meta-analysis**, or the use of **WGS** data has been proposed (Wainschtein et al., 2022). The expensive costs of WGS have benefited the use of large-scale initiatives or meta-analysis. Particularly, large Consortia have been established to analyse bigger cohorts and to generate public and private genetic biobanks (Swede, Stone, & Norwood, 2007), thus facilitating the availability of genotype and phenotype data of thousands of individuals (**see TIGER publication**). The accession to these larger individual cohorts has reinforced the possibility of improving GWAS discovery, granted a better phenotype classification, and facilitated the opportunity of identifying more genetically homogeneous groups (**see Polymorphic Inversions and unpublished Epistasis**). Moreover, meta-analyses, which are based on the statistical combination of publicly available GWAS summary statistics results, have been commonly used in the same direction. Hence, resulting in an improvement on the discovery based on the reduction of false-positive findings, and a gain of detection power due to the increase of sample size (**see TIGER publication**).

Additionally, the probabilities of finding a GWAS signal associated with the disease increase with the **number of genetic markers** that can be tested. Particularly, under the LD background surrounding this type of study, where an associated common signal resulting from GWAS can be

masking a real rare causal variant in LD with the first, the maximisation of the number of genetic markers evaluated becomes crucial. Fortunately, the advances made to generate haplotype and genotype reference population panels, have facilitated this task. Thus, the common practice to increase the number of variants analysed in a GWAS, is to apply to the genotyping array information available for each individual in the study a **quality control**, followed by **phasing** and **imputation** techniques (**see Polymorphic Inversions and TIGER publications, and unpublished Epistasis**). The main goal of using these methodologies is to infer the genotype for multiple individual variants, from which the genotype is missing or unknown in the genotyping array data (Lo, 2014; Marchini, 2019). As a result, the use of imputation has facilitated the inclusion of millions of variants in GWAS analysis, thus improving the discovery power of these approaches and, consequently, enhancing the identification of new loci significantly associated with complex disorders.

Moreover, it has been settled that the power to detect a genomic variant associated with a complex disease through GWAS maximises when the test matches the underlying **inheritance model** of the causal allele (Lettre, Lange, & Hirschhorn, 2007). However, the common practice in GWAS is to analyse variants under the additive model (**see unpublished Epistasis**). Indeed, despite the recognized contribution of GWAS analyses under the additive model to the explanation of a large fraction of complex diseases heritability, there are many genomic variants that follow a non-additive inheritance model (recessive, dominant, or heterodominant). Therefore, the variants following non-additive models tend to be poorly detected or completely disregarded in the vast majority of current GWAS. Consequently, the simultaneous test of different genetic models has been suggested as a successful approach to gain statistical power to detect disease susceptibility loci and, therefore, to improve the knowledge based on the genetic architecture of complex diseases (Guindo-Martínez, Amela, & et al., 2021; Pozarickij, Williams, & Guggenheim, 2020).

Finally, the **genetic heterogeneity** of an observed trait in the population, as well as the **variability present in the population of study**, also affects the GWAS discovery power. In the case of disease heterogeneity, the multiple clinical variables related to complex diseased phenotypes, as well as comorbidities, can dilute specific clinical groups related signals, thus reducing the detection to the most common susceptibility loci between groups of the same disease (**see Polymorphic Inversions publication**). For this reason, although this strategy has been a valuable resource to find some of the genetic mechanisms underlying complex diseases, the discovery of variants related to more specific groups of individuals has been proposed as a crucial step towards precision medicine. As a result, different initiatives have emerged to create subclassifications of diseased individuals based on clinical variables. These patient stratifications have facilitated the possibility to perform more homogeneous GWAS based on these sub-phenotypes and to find their etiological differences (Ahlqvist et al., 2018; Ahlqvist, Prasad, & Groop, 2020; Mansour Aly et al., 2021). In a similar manner, the different allele frequencies and LD patterns emerging from the different ancestral backgrounds have also limited the possibility of extending or replicating the results in other populations. Therefore, constituting an impairment for underrepresented populations, and reducing the genetic understanding of complex diseases to the most commonly studied populations, such as European ancestry populations (**see Polymorphic Inversions and TIGER publications, and unpublished Epistasis**). However, despite the narrowed GWAS discovery behind this population genetic heterogeneity, multi-ancestry studies have shown that still a big fraction of common variants are shared across different ancestries (J. Chen et al., 2021; M.-H. Chen et al., 2020). In contrast, those studies have also supported the relevance of ancestry-specific analysis to find the genetic particularities of each population. Thus, opening a new avenue for population-specific GWAS, and a more global representation of different ancestry populations in genetic studies.

1.2.3.2.2. Complex interaction models

Despite the undeniable success of GWAS to find variants associated with disease, the statistical models usually applied to perform the phenotype-genotype association tests have limited its

discovery. Particularly, although the software specifically developed to perform these analyses has facilitated this task, the large number of variants that are expected to be analysed simultaneously in a GWAS converts the genomic study of disease association into a computational challenging problem. As a result, although complex traits are known to be affected by the combination of multiple genetic and environmental components, current GWAS evaluates the effect of single independent variants (Tam et al., 2019). Consequently, the identification of genomic loci under more complex models, such as **gene-gene interactions (GxG)** (see **unpublished Epistasis**), and **gene-environment interactions (GxE)**, are usually not considered from the analysis, thus limiting GWAS discovery, and contributing to the missing heritability problem (Manolio et al., 2009).

At the genomic level, common diseases are caused by the simultaneous combination of multiple variants each with a low contribution or effect on the disease (McCarthy et al., 2008). However, **GxG interactions** are usually reduced to consider the effects of variants additively, thus, ignoring the study of variants dependency (**epistasis**), the effect of their functional interconnections, and its association with diseased phenotypes (Mackay, 2014), or reducing it to the test of a small fraction of variants usually underlying a shared biological explanation. The main cause of this problem is the computational challenge that represents the analysis of epistasis, where for example billions (10^{12}) of tests are needed just to analyse the complete set of pairwise interactions between 500,000 SNPs (Marchini, Donnelly, & Cardon, 2005). Subsequently, diverse techniques such as **multidimensionality reduction analysis**, or **variants filtering** to restrict the analysis to sets of variants previously known to be related to biological regulatory functions, have been developed and applied to approach this problem (Manduchi, Chesi, Hall, Grant, & Moore, 2018; Josep Maria Mercader et al., 2008). Interestingly, regardless of the limitations derived from the reduction of the discovery dataset, a few genetic variants have been discovered which, despite having only modest significance on a phenotype individually, have an increased effect when considered jointly (Cordell, 2009; Kirino et al., 2013; Monir & Zhu, 2017) (see **unpublished Epistasis**).

Additionally, the role of multiple environmental and clinical variables on the development of a disease is known to have an effect on complex diseases. Therefore, the focus of **GxE interaction** is the analysis of the environmental factors, such as diet, lifestyle, psychosocial stress or airborne agents, and their relation with different genotype groups, in terms of disease associations (Bookman et al., 2011; Dempfle et al., 2008). GxE studies are usually approached by **Environment-Wide Association Studies (EWAS)**, which are an extension of GWAS where the environmental variables can be simultaneously tested with the genotype. However, the difficulties to measure some environmental variables, as well as the uncertainty to understand which features can be contributing to a disease, and the complexity of the underlying models, usually surrounded by a computationally expensive background, have limited their use and discovery (McAllister et al., 2017; Thomas, 2010; Zheng et al., 2020). Indeed, GxE studies have opened a gate for future studies given its relevance to understand the genomic differences between populations, which can be interpreted as the result from an adaptation process to a particular environment, or to the exposures to certain conditions.

1.2.3.2.3. Lack of functional interpretation

The study of the effects of genomic variation on the predisposition to develop a complex trait or disease involves a discovery phase, where multiple variants are proposed to be associated with the disease of study, followed by a **functional interpretation** step to identify the biological mechanisms and pathways that mediate disease. This last step is crucial to find the proteins that are involved in the disease and to find new drugs and therapies. However, despite technological advances have enhanced the discovery, the interpretation is still a challenge in genetic studies. Particularly, from the millions of variants simultaneously tested in a GWAS, the few hundreds or thousands of them which are significantly associated with the disease lack of functional interpretation. Thus, limiting the understanding of the biological consequences of a GWAS variant in relation with the disease.

Notably, this lack of functional interpretation, combined with the fact that most susceptibility locus lie outside the coding regions and are assumed to influence transcript regulation rather than gene function (McCarthy, 2010), hinders the analysis of GWAS outcomes (Tam et al., 2019). For this reason, many studies have advanced in the direction of developing and applying different methodologies to facilitate the translation of the genomic markers obtained from GWAS into relevant biological or clinical information (**see TIGER publication**). As a result, the functional annotation of variants, as well as the assessment of its association with transcriptional changes, and their overlap with epigenetic marks, constitutes a valuable tool for the understanding of the functional impact of variants on the disease. Therefore, **expression analyses, gene, pathway, regulatory elements and epigenetic marks enrichment**, are the most common approaches used to gain insight on this missing biological understanding (Cano-Gamez & Trynka, 2020; Lichou & Trynka, 2020; Manolio, 2013) (**see chapter 1.2.4., Polymorphic Inversions and TIGER publications, and unpublished Epistasis**). Additionally, the experimental assay of the results in cell lines and other organisms is applied to support or reject GWAS findings, and **Polygenic Risk Scores (PRS)**, although still incomplete, have been recently applied to GWAS summary statistics to mediate the translation of the statistical outcomes into something actionable in clinics (Kullo et al., 2022; Kumuthini et al., 2022; Lambert, Abraham, & Inouye, 2019).

1.2.3.3. Machine learning (ML) approaches

The undeniable relevance of the genomic study of complex diseases to find an explanation for the missing heritability, and to find the relation between the different omic layers to better comprehend this type of diseases, as well as GWAS limitations, has promoted the use of new analytical frameworks during the last decades. Notably, although different statistical and computational approaches were already available to analyse these problems, the use of **machine learning (ML)** and **neural networks** algorithms have been lately popularised in the Biomedicine field. All these methods rely on mathematical and statistical approaches, which can be applied to solve classification, clustering, regression and ranking problems. Particularly, for the scope of the genomic study of complex diseases, are both useful in terms of **making predictions**, but also to **find the underlying biological mechanisms of diseases**.

In short, ML methods are fundamentally based in the comparison of the variables (**features**) in a large number of **observations** from a subset of the input data. During this process, the method is able to learn about the necessary decisions to solve a particular problem, based on the features. Then, the same decisions can be applied in an independent dataset to solve the same problem (Greener, Kandathil, Moffat, & Jones, 2021). As a result, the use of this methodology in the Biomedicine field has shown its effectiveness to approach disease heterogeneity problems such as the classification of diabetic and obese individuals based on clinical variables (S. B. Cho, Kim, & Chung, 2019; Lin et al., 2021), GWAS loci prioritization (Nicholls et al., 2020), finding main effects and interaction associations with disease (Szymczak et al., 2009), and the study of epistasis (Behravan et al., 2018; Y. M. Cho et al., 2004; Manduchi et al., 2018; Niel, Sinoquet, Dina, & Rocheleau, 2015; Sheppard et al., 2021; Verma et al., 2018; Wei, Hemani, & Haley, 2014) (**see unpublished Epistasis**).

1.2.3.4. ML limitations

Although the numerable contribution of ML methods to the better understanding of complex diseases has made them gaining popularity in the biomedical field, there are still many computational and statistical challenges surrounding these procedures (Chicco, 2017; Sarker, 2021). The factors involved in ML limitations are related to the input data, the methods, and the outcomes of the study. Most of these limitations affect the effectiveness and the reliability of the methodology, and, therefore, the ability of the method to discover the correct genetic, clinical, or molecular markers associated with a complex disease, or to do a proper classification of patients. As a consequence, the **data-preprocessing**, the selection of a correct **learner**, and the preparation of **the ML pipeline** are crucial

for the analysis. Additionally, the outcomes obtained from ML algorithms are difficult to interpret, and usually lack of functional **interpretation**, thus representing an additional layer of complexity for the understanding of the underlying molecular mechanisms of disease predisposition.

1.2.3.4.1. Data pre-processing

There are many factors surrounding the data that can affect the ability of the ML method to discover the genetic and clinical variables associated with complex diseases, such as the amount of **available data**, **data type**, **data imbalance**, the **presence of outliers**, and **data missingness** (Chicco, 2017; Sarker, 2021). As a consequence, the **data pre-processing** step is crucial to prepare and curate the data previous to the application of a ML algorithm. This step is a complex process that requires a solid background to understand the data included in the study, the problems related to the type of the data, and a good comprehension of the ML model. Particularly, the data pre-processing step benefits the learning process ensuring the effectiveness of the methodology, and preventing from false positive results.

ML models are restricted to the analysis of large datasets of observations with at least **ten times the number of features** (Chicco, 2017). However, despite the large volumes of genomic data generated during the last decades, this is not always possible. For instance, if the features correspond to the number of susceptibility loci to be evaluated, there can be millions of features, while, in contrast, the number of observations or patients presenting those features will be measured in thousands. As a result, it is necessary to understand the effects of applying ML techniques in smaller datasets, such as the **overfitting** problem. Overfitting can occur during the training process when the model instead of learning memorises the features of the training set, so that it obtains excellent results during the training, but has a poor performance in any other independent dataset (**Figure 7**).

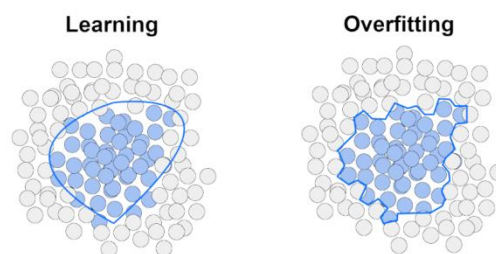


Figure 7. ML overfitting problem. Overfitting is a common ML problem, which occurs during the training step. When a ML model presents overfitting, instead of learning the relation between the variables and the output, it memorises the training features, thus resulting in poor performance in any other independent dataset. In this example, the model is expected to define a decision frontier (blue line) to classify dots in two categories (grey and blue). The left graph represents the results obtained from a good learner. In contrast, the right picture displays the overfitting case.

To avoid the overfitting problem derived from the scarcity of data, a common practice is to apply **multi-dimensionality reduction** techniques, which cover a wide variety of frameworks that range from **statistical methods**, such as K centroids or Principal Component Analysis (PCA) (Monaco et al., 2021), to more **biological based approaches** where the features are filtered based on prior biological knowledge (Manduchi et al., 2018). As a result of the use of multi-dimensionality reduction techniques, it is possible to reduce the number of features included in the data but preserve their relations, thus facilitating the application of ML algorithms.

Additionally, the presence of **imbalance** affects the performance of the ML model in such a way that the method is biased towards the selection of features related to the more representative class, thus to the detriment of the underrepresented class. To overcome this problem, there are different approaches that can be applied such as **under-sampling** by removing elements from the over-represented classes, and correcting the imbalance through **class-weighting** techniques (Chicco,

2017). However, both methodologies are prone to have an impact on the results and the applicability of the method. For instance, in terms of the applicability, an extreme reduction of observations can lead to problems related to data insufficiency. On the other hand, although class-weighting techniques are of particular interest, not all the ML methods include this characteristic, thus limiting the methodology. Moreover, in case of extreme imbalance, the weighting is not always an insurance to obtain the best results. Remarkably, a good understanding of the data and the model facilitates the choice of the best way to deal with the presence of data imbalance, and therefore, to improve the effectiveness of the method.

A similar problem occurs with any possible data-related issue that can result in **trend decisions** for the ML model, such as **missingness**, **redundancy**, or the presence of **outliers** (Chicco, 2017). Particularly, there are several statistical and computational frameworks that can be applied to for example deal with inconsistent values and outliers, such as **normalisation** in case of numeric features, or **value removal**. In contrast, in the case of missingness and redundancy, there are some ML models which are prepared to manage this type of data issues. Nonetheless, as not all the methods accept missingness or redundancy, statistical techniques such as **inference**, **transformations**, and **value approximation** are commonly used to prepare a cleaner dataset without falling into a data insufficiency problem. As a consequence, the preparation of the input dataset based on the correction of all these problems is one of the keys to improve the results that can be obtained from the ML analysis, and to ensure a good performance.

Last, the **different types of data** affect the selection of a ML learning model. Particularly, the data can be classified in **structured**, **unstructured**, **semi-structured** and **metadata** (Suppl. Table 2) (Sarker, 2021), and not all the ML models are specifically designed to deal with all types of data. Therefore, the proper identification of the type of data included in the study will result in a better decision between using a ML method or an alternative approach, and consequently, in an improved resolution of the problem.

1.2.3.4.2. ML algorithms

Despite the existence of a large variety of **types of ML algorithms**, not all of them are applicable to all studies. For example, based on genomic features, different types of learners can be used to classify a group of individuals in diseased and non-diseased, or to find different subgroups of diseased individuals. There are many factors that affect the selection of the most appropriate ML algorithm to approach a particular problem that needs to be solved in a specific dataset of study. These factors include the **type of learning**, the **input data**, and the **class of the problem**. Moreover, ML methods are defined as training-test approaches where there is a learning step (**training**) for the method to find and understand the input variables relation with the output, followed by an evaluation step (**test**). Therefore, after the selection of the most suitable group of learners to approach a genomic problem, there are different parameters that need to be adjusted inside the **ML pipeline** to obtain the best performance. These parameters include the **split** of the input data in the training and test sets, and the **hyperparameters** of the model.

Based on the **type of learning**, a ML model can be classified as **supervised**, **unsupervised**, **semi-supervised**, **reinforced**, **multitask**, **ensemble learning** or **instance-based learning** (Chicco, 2017; Dey, 2016; Greener et al., 2021; Sarker, 2021). As a matter of fact, the classification of diseases and non-diseased individuals can be approached with a supervised learner, while the creation of different subgroups of diseased individuals needs the use of unsupervised methods (Suppl. Table 3). However, the classification of ML models based on the type of learning includes a wide range of learners that can be applied to solve an extensive variety of problems. Thus, highlighting the relevance of a better characterisation of ML approaches based on the **type of problem** to be solved. The most common type of problems approached by ML algorithms are **classification**, **regression**, **clustering**, **feature engineering** and **dimensionality reduction**,

association rule learning, or **reinforcement learning techniques** (Suppl. Figure 1). Moreover, inside these last groups there are different learners. For example, there are different types of ML classifiers, which can be divided in **binary**, if there are only two classification labels, such as diseased and non-diseased, **multiclass**, when there are more than two classification labels, and **multi-label** if there is a hierarchical structure in the classification labels, so that the same object of study can belong to different classes, such as species (Sarker, 2021).

Additionally, the parameters that can be adjusted in the **ML pipeline** and the algorithm have an effect on the effectiveness of the method to solve a genomic problem. Particularly, to use a ML algorithm the first step is to **split** the input dataset in two independent subsets named **training set** and **test set**. This split needs to be done in a proportion that ensures the procurement of a large amount of observations for the training, but keeps enough data to evaluate the results in a sufficiently heterogeneous dataset. Therefore, the split can be added to the **hyperparameter adjustment** step, where the basic properties of the model are calibrated, previous to the training step, to prevent overfitting, and to obtain the best results from the analysis. In this process a **grid search** including all the possible combinations of hyperparameter values is tested using a **K-fold cross-validation (K-fold CV)** algorithm (Chicco, 2017; Greener et al., 2021). As a result, the best hyperparameters for the model are defined by those resulting in the best median global performance (**Figure 8**).

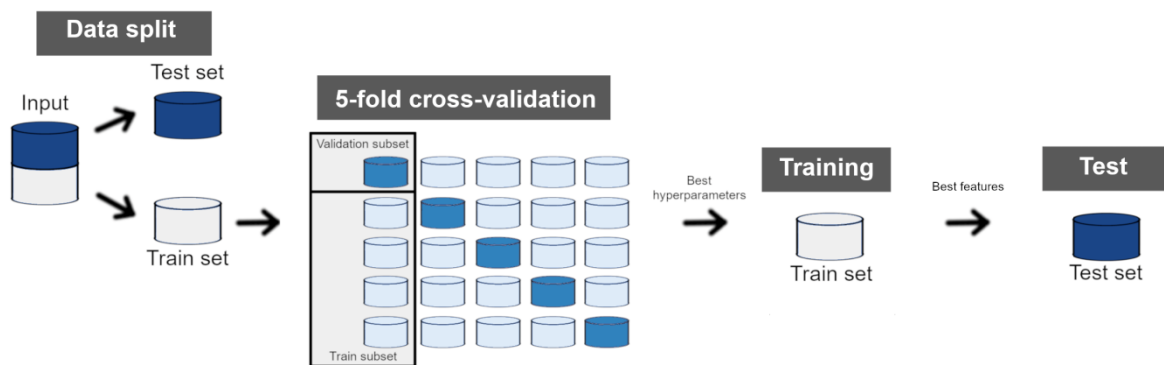


Figure 8. ML algorithm pipeline. The input dataset is divided into training (light grey) and test subsets (dark blue). The first step implies a K-fold cross-validation (K-fold CV) algorithm, which in the figure corresponds to a 5-fold CV. This K-fold CV algorithm is used to do a grid search hyperparameter adjustment, therefore, to obtain the best performance of the model, and to prevent overfitting. In the first step of the 5-fold CV, the training set is divided into 5 data subsets. Then, in each step of the 5-fold CV, these subsets are shuffled to create the corresponding train (light blue) and validation (medium blue) subsets. Each hyperparameter value combination is fitted on the train subset, and then evaluated in the validation subset. Finally, the best hyperparameters are used to fit the initial train set, in the training step, and the performance is evaluated using the test set, during the test step.

The selection of a group of similar learners and the best hyperparameters for each particular genomic problem in an specific dataset, has a direct effect on the performance, complexity, and success of the study (Greener et al., 2021). Remarkably, supervised and unsupervised learners have been broadly used ML approaches for the biology and medical community in the study of complex diseases to solve a wide diversity of problems. In particular, classification learners have been broadly applied to find the most relevant group of variables, which can be clinical or genomic, involved in the development of a disease, to classify diseased individuals into subgroups of patients, or to detect groups of genomic variants associated with disease (Ahlqvist et al., 2018; Behravan et al., 2018) (**see unpublished Epistasis**).

1.2.3.4.3. Lack of functional interpretation

The study of the genetic basis of complex disease predisposition involves the discovery of multiple disease susceptibility loci, and its functional **interpretation** to understand the underlying

molecular mechanisms to develop the disease. In this direction, the results obtained from a ML model provide the **most relevant features** for the method to solve the problem. Additionally, from the evaluation of the outcomes obtained from the model on the test set, a diverse range of measures for its **reliability** can be calculated. Finally, relying on the **model**, the interpretation of the results of ML methods is based on the comprehension of the putative relation between the features obtained as an outcome from the learner and the disease of study (**Suppl. Table 4**) (T. Chen & Guestrin, 2016; Dey, 2016; Greener et al., 2021; Sarker, 2021). For example, in the classification of a group of patients in diseased and non-diseased, which can be analysed with a binary classification learner, the outcomes of the model are the most relevant genomic variants to do the classification, each one with their corresponding associated score (**see unpublished Epistasis**). Then, as a result of the prediction on the test set, each individual can be classified as case (positive) or control (negative). Therefore, the comparison between the predicted values with the real observed values, determines if the prediction is true or false. Consequently, there are only four expected possibilities to measure the goodness of the outcomes, which correspond to **true negative (TN)**, **true positive (TP)**, **false negative (FN)**, and **false positive (FP)**. These values can be used to evaluate a **global estimate of its effectiveness** (**Figure 9; Suppl. Table 5**).

Observed	Predicted	
	Case	Control
Case	TP	FN
Control	FP	TN

True negative (TN): when a control is predicted as control.
 False positive (FP): if a control individual is misclassified as case.
 True positive (TP): when a case is predicted as case.
 False negative (FN): when a case is misclassified as control.

Figure 9. Evaluation of the results of the predictions made by a ML binary classifier. Only four possibilities can be expected from the predictions. If the value of the prediction matches the real value, it can be a True Positive (TP) or True Negative (TN) (green blocks). A TP corresponds to a diseased individual (case) which has been correctly classified. A TN corresponds to a non-diseased individual (control) properly predicted. If the prediction is incorrect, a control predicted as a case will be a False Positive (FP), and a case predicted as a control will be False Negative (FN) (red blocks).

All these outcomes, provide a global view of the performance of the model, and facilitate its interpretation in terms of the association with the disease. However, these are far from the functional interpretation, thus representing a limitation to understand the biological pathways affecting to the development of the disease. Therefore, to find the overlying molecular mechanisms of the associations found, in a similar manner than GWAS, the results obtained from the ML model need to be complemented with other related genomic, transcriptomic, and epigenetic studies.

1.2.4. Molecular basis of complex diseases and functional interpretation

The great progress made on the genetic study of complex diseases, which has involved the creation of large catalogues with thousands of variants with a putative effect on the predisposition to hundreds of complex traits and diseases, has facilitated the advance towards a better detection, prevention, and treatment protocols (Beck et al., 2014; Buniello et al., 2019; K. Watanabe et al., 2019). However, although different strategies, such as GWAS or ML methods, have been broadly applied contributing to the discovery of these variants associated with complex diseases, these methods lack of functional explanation. Therefore, evidencing the relevance of the application of complementary methodologies, which focus on the **translation of genomic variation in function**, to find new drugs and therapeutic targets.

The analysis of the effect of genomic variation on cell functions is one of the main subjects of study from the **transcriptomics** and **epigenetics** fields. Transcriptomics focus on the analysis of all

the biological processes that are related to the transcription of the DNA into RNA, and epigenetics studies the reversible modifications on a cell DNA that affect the **regulatory mechanisms of gene expression (transcription factors)**. Therefore, starting from the detection of DNA alterations associated with a complex phenotype as a result from the genomics field, a posterior transcriptomic analysis can be applied to find putative effects of these variations on genes and **gene expression**. Additionally, an epigenetic analysis can be performed to further understand if genomic variation has an effect on the regulatory mechanisms, thus possibly causing an effect on gene expression. Consequently, the integrative analysis of genomics, transcriptomics, and epigenetics, has been suggested to play a key role towards a better understanding of the biological pathways underlying genetic variability. Thus, converting the analysis of gene expression, gene expression regulatory elements, and gene expression regulatory variation, in crucial steps to find the biological underlying mechanisms involved in variant-disease associations.

1.2.4.1. Gene expression

As genes are directly related to protein production and cell specific functions, the genomic alterations with an effect on gene expression can result in cell dysfunction, and, possibly, increase the risk of developing a disease. Particularly, **gene expression** is a complex process by which the DNA information is transcribed and translated to **messenger RNA (mRNA)**. The amounts of mRNA produced in a cell during this process are used to direct protein synthesis, other post-translational processing, and modifications such as alternative splicing, which allows the same gene to code for different proteins, and therefore, leading to different biological functions. Thus, to evaluate the effects of variation of gene expression in cellular function and the phenotype it is necessary to quantify gene expression (Buccitelli & Selbach, 2020).

Gene expression analysis focuses on the study of the profile of the transcriptome to measure the relative and absolute values of the transcript. Particularly, RNA-seq and gene expression arrays technologies are used to estimate the levels of mRNA (Dalkılıç, 2009). As a result, gene expression analysis facilitates the estimation of the levels of mRNA, or expression for downstream analyses. These results facilitate the functional interpretation of disease-associated locus based on the study of its putative effect on gene expression. Thus, converting the study of gene expression variation in a crucial step to improve the understanding of complex diseases.

1.2.4.2. Gene expression regulation

During the gene expression process, the mRNA production is controlled at different levels by **regulatory proteins**, which encompass to coordinate and control the transcription and translation processes. For this reason, some alterations of the DNA sequence encoding the regulatory elements regions involved in **gene expression regulation** can result in a functional impact on gene expression, thus affecting the biological functions, and possibly mediating disease. Particularly, during gene expression regulation the **RNA polymerase**, which will transcribe the DNA to mRNA, is attracted to the promoter region of the gene located in the **transcription start site (TSS)** in **5'-UTR**. In parallel, the TFs facilitate the activation or repression of the transcription by binding to their specific **DNA-binding domains** or **motifs**. These regions are usually located in the **promoter** region or in more distant **enhancers** upstream 5'-UTR or downstream **3'-UTR**. In case of activation of the transcription, the promoter is the responsible regulatory element of allowing it to start. On the other hand, the enhancers activate or increase the rate of transcription from the target gene promoter but also can drive the transcription independent of their target promoter (T. K. Kim & Shiekhhattar, 2015; Lambert et al., 2018; Smith, Lam, Markova, Yee, & Ahituv, 2012) (**Figure 10**). Hence, playing a key role in gene expression regulation, **transcription factors (TFs)**, **epigenetic marks** and **chromatin topology**, **RNA-binding proteins**, and **non-coding RNAs** are some of the most relevant targets to evaluate the functional impact of variation in complex diseases (Buccitelli & Selbach, 2020; García-Sánchez & Marqués-García, 2016).

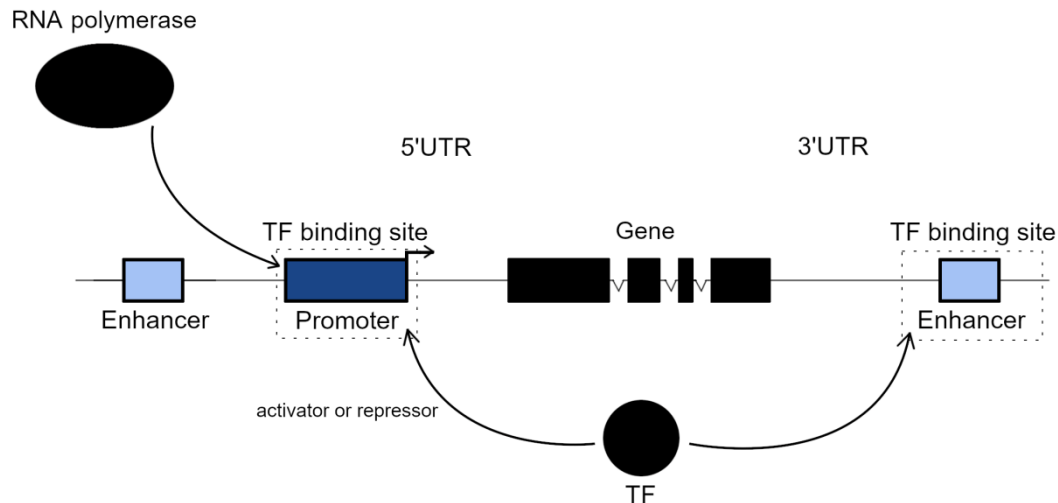


Figure 10. Gene expression regulatory process schema. The RNA polymerase is attracted by the promoter (dark blue), which will start the transcription of the gene, if the transcription factor (TF) activates it, by binding to its specific binding-site. The neighbouring enhancers (light blue) to the targeted promoter increase or activate the transcription rate.

Modifications in TFs, which define any protein involved in the transcription process or that has the ability of regulating expression, the **chromatin**, which is a substance wrapping the DNA, or **histones**, which are the major proteins in chromatin, and act as packaging elements for the DNA, can result in alterations of cell function with an effect on the phenotype (Buccitelli & Selbach, 2020; Deplancke, Alpern, & Gardeux, 2016; García-Sánchez & Marqués-García, 2016; Pope & Medzhitov, 2018). Hence, different experimental methods have been developed and used to approach the study of gene expression regulation at a genomic level. For example, as **open chromatin regions** are a potential site of TF binding, the use of **chromatin immunoprecipitation followed by sequencing (ChIP-seq)** is crucial for the identification of TF binding sites (Smith et al., 2012). Additionally, the application of **assays for transposase-accessible chromatin sequencing (ATAC-seq)** allows the identification of enhancers without any prior knowledge of TF binding and chromosome conformation capture (Buccitelli & Selbach, 2020; T. K. Kim & Shiekhhattar, 2015; Lambert et al., 2018). As a result, the use of these methodologies in genomic studies has facilitated a better comprehension of the role of chromatin and its modifications, the relationship between functional regulatory elements and features of chromatin accessibility and histone modification, their correlation with active chromatin marks such as **H3K4me1** or **H3K27ac**, and the gene silencing process occurring in **DNA methylation**, thus improving the functional interpretation of genomic variation and its potential effects on disease.

Remarkably, although gene expression can be ubiquitous or cell-type specific, some of the regulatory elements such as gene expression signatures, enhancers, and promoters are **cell-type specific** (Long, Prescott, & Wysocka, 2016; Nica & Dermitzakis, 2013; Pope & Medzhitov, 2018). Thus, suggesting the relevance of the study of disease related cell-type regulatory elements to improve the understanding of the mechanisms mediating disease (**see TIGER publication**).

1.2.4.4. *cis*-regulatory expression

The understanding of the relationship between genomic variation association results and TFs cannot always be directly inferred from the proximity of a disease association signal with a gene binding site (Deplancke et al., 2016), thus, enforcing the need of other types of gene expression analyses, such as **expression quantitative trait loci (eQTL)** or **allele-specific expression (ASE)**. Particularly, eQTL studies focus their analysis in finding the association between genetic locus with gene expression levels, and ASE assesses the allelic imbalance contribution of genetic variants to

gene expression. Consequently, these complementary analyses are fundamental to find a putative functional interpretation of GWAS signals in terms of disease susceptibility (Cleary & Seoighe, 2021; Nica & Dermitzakis, 2013) (**see TIGER publication**).

1.2.4.4.1. Expression quantitative trait loci (eQTL) studies

The connection between regulatory elements, gene expression, and disease, evidences the need of analysing the effects of genomic variation in gene expression. Particularly, **eQTL** analyses focus on the discovery of **variants statistically associated with changes in gene expression levels**. Thus, suggesting possible links between genomic variation and gene regulation (Albert & Kruglyak, 2015a; Nica & Dermitzakis, 2013). Briefly, the study of eQTL association is comparable to a quantitative GWAS, where the genotype of multiple individuals is simultaneously tested in different loci to find their association with gene expression levels. However, in contrast with GWAS, the number of individuals required in eQTL studies to obtain significant results ranges between tens to hundreds. This reduction on the sample size, which is mainly caused by the stronger effect sizes attributable to the evaluation of a quantitative trait, facilitates the inspection of the association between genomic variation and gene expression.

The gene associations captured by eQTLs are classified by their proximity to their associated genes, thus separating them on ***cis*** and ***trans***. In particular, variants with 1Mb on either side of a gene's TSS are called *cis* and those with at least 5Mb of the TSS are considered *trans*. The majority of *cis*-eQTLs have been found to act with a higher effect size (Cookson, Liang, Abecasis, Moffatt, & Lathrop, 2009). However, although with lower effects, *trans*-eQTLs are more numerous and act with more tissue specificity (Grundberg et al., 2012). Nonetheless, the possibility to capture *trans*-eQTLs is usually a computational challenge mainly due to the human genome architecture and the relatively modest effect sizes. Notably, the correlation between the discovery power and the sample size, for both *cis* and *trans* eQTLs, still represents a limitation for *trans*-eQTLs discovery (The GTEx Consortium, 2020). Therefore, although up to 70% of the variance between individuals gene expression has been attributed to *trans*-eQTLs, the multiple difficulties in their study has promoted that the vast majority of eQTL studies focus in their *cis* contribution (Umans, Battle, & Gilad, 2021).

Moreover, in terms of cell function, gene expression signatures and regulatory elements are cell-type specific, therefore suggesting that the regulatory effects of eQTL are also **tissue-dependent** (Long et al., 2016; Nica & Dermitzakis, 2013). As a consequence, to understand the effects of genetic variability on disease, the **Genotype-Tissue Expression (GTEx) project** emerged in 2017 with the large-scale initiative of generating a comprehensive public resource to facilitate the study of the effects of genomic variation in tissue-specific gene expression and regulation (The GTEx Consortium, 2017). In the last release of this project, 15,201 RNA-sequencing samples from 49 tissues of 838 post-mortem donors were analysed, thus facilitating the characterization of genetic associations for gene expression and splicing in *cis* and *trans*. This study revealed that eQTLs in tissues with higher cell specificity, such as brain, testis, lymphoblastoid cell lines, whole blood, or liver, result in stronger effect sizes and a subsequent increase in the association detection power. Nevertheless, despite this tissue-specificity condition, there is a high order of eQTL similarity between different tissues (The GTEx Consortium, 2020). Moreover, they found that the majority of genes are affected by local genetic variation, eQTLs are usually enriched in enhancers and related elements, and that although presenting differences between ancestries, common regulatory effects are largely shared between populations (Stranger et al., 2012).

As a result, the study of eQTL based on its tissue-specificity can lead to better results, in terms of power of detection based on the effect size, as well as in terms of disease interpretation. Particularly, if the regulatory signal is associated with a relevant tissue for the disease, a GWAS and eQTL correlation can be considered as a sign of a putative causal relation (**see TIGER publication**). Therefore, the integration of GWAS and eQTL signals can be used to discover target genes and

pathways underlying putative relations with the biological mechanisms mediating disease. Thus, facilitating the functional interpretation of GWAS results, but also enhancing the prioritisation of GWAS signals. Particularly, across all GTEx tissues, 43% of disease-associated loci colocalize with a known eQTL (Umans et al., 2021).

[1.2.4.4.2. Allele-specific expression \(ASE\) studies](#)

ASE emerged as a way of analysing the relation between genomic variation, gene regulatory elements, gene expression, and disease (Cleary & Seoighe, 2021). Particularly, ASE is a phenomenon that occurs, in a *cis* manner, when two alleles in the same heterozygous loci present different expression levels. Thus, creating an **allelic imbalance** where, in some cases, one of the alleles can appear totally silenced. This imbalance suggests a possible variation effect on gene expression regulation and a consequent contribution in human phenotypes and complex disease susceptibility. Particularly, ASE can contribute to disease susceptibility when the prioritisation of expression is towards the disease allele instead of the functional allele (Lee, Kang, Gandal, Eskin, & Geschwind, 2019; Luft, Young, Meynert, & Taylor, 2020). In contrast, it can protect from disease by compensating variation through a higher expression of the functional allele (N. de Klein et al., 2020).

ASE analyses are usually performed at the level of the individual, therefore complementing the results obtained in other expression studies such as eQTL, by capturing signals that can be masked by the group analyses. Particularly, ASE pipelines have three steps involving the detection of heterozygous positions, a filtering to improve the accuracy of the identified heterozygous loci, and a final estimation of the regulatory effects of variation (Cleary & Seoighe, 2021). For this reason, these types of studies require individual high coverage sequencing, mapping, and alignment, to detect the heterozygous loci. Thus, deriving in many complications mostly related to the accuracy to detect the heterozygous positions. However, many strategies have been developed to improve these tasks, like the use of genotyping array data to remove false positive heterozygous positions (Van De Geijn, Mcvicker, Gilad, & Pritchard, 2015). Therefore, ASE studies result in the association between allelic imbalance and expression, where the haplotypes of multiple expressed heterozygous SNPs are simultaneously tested for unequal representation of the two alleles (**see TIGER publication**).

Interestingly, the different advances made in the genomic field have opened the possibility to improve this individual analysis (Cleary & Seoighe, 2021). For example, the availability of population-based phasing facilitates the inspection of other regulatory variants present in the same region. This information can be used to identify the association between the imbalance and nearby putative regulatory variants (**see TIGER publication**). In addition, the availability of multiple individuals' information can be used to extend the expression imbalance analysis to find correlations with the allele at the regulatory variant. In this case, allelic imbalance can be combined with an overlapping or colocalizing eQTL to confirm its *cis* effect on the gene. As a result, ASE results can be used to facilitate GWAS interpretation by fine-mapping functional genetic variants, or to prioritise the results by including variants enriched in active regions in the genome.

[1.2.4.5. Public genomic functional interpretation databases](#)

The remarkable progress made by the genomic, transcriptomic, and epigenetic fields to understand the underlying molecular mechanisms of genetic variability and complex diseases, has promoted the generation of **publicly available databases** containing this valuable resource of information. Complementary to the Human Genome Project database (International Human Genome Sequencing Consortium, 2004), these large databases aim to provide the community with powerful tools that facilitate the functional assessment and interpretation of the genomic outcomes of GWAS (**see TIGER publication**). Particularly, the catalogue of resources include databases that contribute, among others, with **genes** and **isoforms** description and categorization, gene and **gene products** functional descriptions, **protein** and **macromolecular complexes** roles, lists of **TFs** with annotated elements and **binding interfaces**, lists of TFs and their corresponding **regulatory interactions**,

global and tissue-specific gene expression regulators, or **epigenetic feature profiles** (Suppl. Table 6). The use of the annotations provided by these projects has facilitated the functional interpretation of a large proportion of disease-associated variants. However, there is still a fraction of variants, which have not been captured in these analyses, that remains with missing explanation. Thus, opening a new avenue to further explore the molecular mechanisms underlying complex diseases.

1.3. The study of type 2 diabetes and the relevance of pancreatic islets

The advances made in the genomics field, combined with transcriptomics and epigenetics have facilitated the study of different complex diseases. This is the case of **Type 2 Diabetes (T2D)**, where the parallel efforts done in its study from a large diversity of complementary scopes, such as the clinical, biological, genomic, and pharmacologic, has led to a better understanding of its aetiology, as well as, to the development of different treatments. However, the complexity and the heterogeneity of this common disorder, which affects over 463 million individuals worldwide, needs further analysis to have a complete explanation of its heritability, and to enhance the early detection in clinics. Therefore, a better understanding of the **metabolic, genomic, and epidemiological mechanisms** underlying the disease, the **environmental factors** related to this disorder, as well as an improved comprehension on the **genetic heterogeneity** of T2D, is essential for the advance in its study towards personalised medicine.

1.3.1. Metabolic pathophysiology

T2D is a **complex metabolic disorder** usually observed as a result of a dysfunction in the regulation and use of glucose due to defects on the insulin signalling pathway. **Glucose** is the primary energy resource for our body and consequently, one of the main reasons for food intake. Particularly, glucose is ingested during digestion, entering the blood system, and activating the different mechanisms that promote the **glucose uptake process**. However, glucose cannot be directly uptaken by our organism. Indeed, **insulin**, which is a hormone generated by the pancreas, needs to be secreted to activate the **glucose uptake mechanisms**. Thus, in a common scenario, once insulin has been secreted proportionally to blood glucose concentrations, the glucose uptake from different organs is facilitated (DeFronzo, 2009; Galicia-Garcia et al., 2020). There are many organs involved in the glucose uptake process, including the **stomach, pancreas, liver, gut, primary muscle, kidneys, adipose tissue, and brain** (Kaku, 2010). More specifically, the **beta-cells** present in the **pancreatic Langerhans Islets** are the primary insulin secretors of our body, thus allowing glucose homeostasis.

As a result of the diversity of organs and mechanisms involved in the glucose uptake process, there are different ways of dysfunction that can lead to the development of T2D. For example, dysfunctions in beta-cells can result in a decrease in glucose responsiveness and an **insulin secretion (IS)** impairment (Galicia-Garcia et al., 2020; Kaku, 2010). Moreover, blood insulin concentrations can be exceptionally insufficient to activate the major target organs. This condition, which is referred to as **insulin resistance (IR)**, is promoted by different mechanisms and can result in several regulatory problems. Consequently, T2D is known to be a common multifactorial metabolic disorder related to pancreatic beta-cell IS dysfunctions, and usually surrounded by a background of IR (Bartolomé, 2022; Del Guerra et al., 2005; Eizirik, Pasquali, & Cnop, 2020; Gloyn et al., 2022).

In addition, dysfunctions in each of the main organs during the glucose uptake process, including the adipose tissue, skeletal muscle, liver, gut, the pancreatic beta and alpha cells, kidney and brain, derive different consequences for the disease (**Suppl. Table 7**) (Cnop et al., 2005; Cornell, 2015; DeFronzo, 2009; Del Guerra et al., 2005; Eizirik et al., 2020; Galicia-Garcia et al., 2020; Gilon, 2020; Rhodes, 2005). However, despite the many differences in the mechanisms of the main organs that are involved in the glucose uptake and IS process, it is reasonable to find a straight relation between them. Particularly, the connections between their consequences on dysfunctionality such as

glucotoxicity, lipotoxicity, IIS and hypoglycemia, share an important role in the development of T2D. Therefore, converting the transcriptomic analysis of these tissues into a great resource to find a functional explanation of T2D susceptibility loci (see **TIGER publication**).

1.3.2. Genetics

The synchronised multi organ behaviour, and the large variety of functions surrounding the metabolic pathophysiology of T2D, has a direct reflection in the **polygenic** nature of the disease. More specifically, T2D is a complex disease where multiple variants affecting different genes with small effects, contribute to the disease progress (McCarthy et al., 2008). As a result, although more than 3 thousand genes have been found associated with diabetes, there are some **well-known genes** which are particularly involved in the glucose uptake and IS process (**Table 1**) (Cornell, 2015; DeFronzo, 2009; Eizirik et al., 2020; Galicia-Garcia et al., 2020; Rhodes, 2005; Rouillard et al., 2016; Stelzer et al., 2016).

Table 1. Genes involved in the glucose uptake and IS process.

GENE NAME	DESCRIPTION	FUNCTION	PROBLEM
<i>IAPP</i>	Islet Amyloid Polypeptide	Co-secreted with insulin	<i>IAPP</i> hypersecretion can lead to progressive beta-cell failure
<i>GLUT4</i> or <i>SLC2A4</i>	Glucose Transporter Type 4 or Solute Carrier Family 2 Member 4	Major transporter involved in the uptake of glucose into skeletal muscle	Mutation can result in an under expression of <i>GLUT4</i> , and defects in its pathway reduce the glucose intake and can lead to hyperglycemia
<i>GLP-1</i> or <i>GLP1R</i>, and <i>GIP</i>	Glucagon Like Peptide 1 Receptor and Gastric Inhibitory Polypeptide	Stimulate the release of insulin and the IS after food intake. <i>GLP-1</i> is also involved in the regulation of satiety, gastric emptying, and glucagon secretion	Deficiencies in <i>GLP-1</i> contribute to T2D progression and beta-cell resistance to <i>GIP</i> , thus inducing glucotoxicity
<i>SGLT2</i>, and <i>GLUT2</i> or <i>SLC2A2</i>	Solute Carrier Family 5 Member 2, and Glucose Transporter Type 2 or Solute Carrier Family 2 Member 2	Glucose reabsorption and transport in the kidneys	Mutations can lead to reabsorption excess and hyperglycemia

These findings highlight the relevance of the **glucose transport** and **exocytosis** of the insulin granules pathways in the study of T2D, as well as, the importance of a good understanding of the genetic basis of the disease to a better explanation of its pathophysiology (Del Guerra et al., 2005).

Some studies have revealed that T2D and beta-cell dysfunction cluster in families, thus suggesting a putative **genetic predisposition** in some individuals to develop the disease (Cornell, 2015; DeFronzo, 2009; Kaku, 2010). This suggested genetic predisposition has converted the genomic study of T2D into a major motivation towards the early detection and prevention of the disease. In this direction, first-degree relatives familial and twin studies have revealed that T2D has an estimated **heritability** from 0.3 to 0.72 in monozygotic twins (Newman et al., 1987; R. M. Watanabe et al., 1999; Willemssen et al., 2015). Therefore, to discover the genomic variants that can predispose to develop the disease, the genetic component of T2D has been broadly analysed during the last decades through GWAS and large GWAS meta-analyses. This type of studies have played a central role in the discovery of **more than 700 signals** associated with T2D and related glycemic traits (Bonàs-Guarch et al., 2018; J. Chen et al., 2021; Mahajan, Taliun, et al., 2018; Scott et al., 2017; The DIAGRAM Consortium, The AGEN-T2D Consortium, The SAT2D Consortium, The MAT2D Consortium, & The

T2D-GENES Consortium, 2014; Vujkovic et al., 2020). Most of these GWAS signals correspond to common variants with a low effect on the disease which **only explain a 20% of T2D heritability** (DeForest & Majithia, 2022), and which although combined in a polygenic score result in a good prediction (AUC=0.901), still cannot be used at a clinical level for the early detection of the disease, and most importantly, **do not improve the prediction based on clinical variables** (Collins, Doudna, Lander, & Rotimi, 2021; Kullo et al., 2022; Kumuthini et al., 2022; Liu, Zhuang, Wang, Huang, & Liu, 2021; McGuire et al., 2020; Padilla-Martínez, Collin, Kwasniewski, & Kretowski, 2020). (**see unpublished Epistasis**).

Remarkably, despite the lack of **functional interpretation** of GWAS, diverse transcriptomic and epigenetic studies have led to the generation of large lists of putative candidate genes, causal variants, and regulatory elements, which have facilitated the better understanding of the disease (Akerman et al., 2017; Miguel-Escalada et al., 2019; Morán et al., 2012; Pasquali et al., 2014; Solimena et al., 2018; Thurner et al., 2018; van de Bunt et al., 2015). Remarkably, the generation of **publicly available resources**, which integrate these large-scale genetic data, has been crucial to facilitate the access to this valuable resource of information and to promote the study of T2D (Flannick & Florez, 2016; Flannick, Johansson, & Njølstad, 2016) (**see TIGER publication**).

1.3.3. Environmental factors

As a complex disease, T2D is characterised by the effect of multiple genetic and **environmental factors** with a straight connection between them and the metabolic pathways affecting the development of the disease. Thus, converting the study of environmental factors that rely on the mechanisms involved in the glucose uptake process into a major interest for the better understanding of this metabolic disorder. For example, **obesity** and **age** have been proved to play an important role in terms of disease development and treatment. However, in addition to age, **physical activity** and **food intake**, there are many other environmental factors related to diet and lifestyle that have been suggested to have a direct effect on the disease. Thus, the study of obesity, **overeating**, **lack of exercise**, **smoking**, **stress**, **alcohol intake**, **nervous and endocrine systems disorders**, and ageing, are of special interest to gain a better comprehension of this complex metabolic disorder (Galicia-Garcia et al., 2020; Kaku, 2010).

In particular, the fact that **one-third of obese individuals develop T2D**, defines obesity as one of the main factors driving the development of the disease (Rhodes, 2005). This relation between obesity and T2D is usually associated with liver and muscle IR, which can result in a progressive beta-cell failure. More explicitly, the liver and muscle IR generate an increased metabolic load demand for insulin, which is usually impossible to cover by the beta-cells, thus causing its failure (Cnop et al., 2005; Cornell, 2015; Rhodes, 2005). Moreover, different studies have revealed the important role of age in the **progressive beta-cell failure** (DeFronzo, 2009). Particularly, the effect of age in the beta-cell mass decrease directly affects the beta-cell function and, consequently, the IS (Cornell, 2015; Rhodes, 2005). Thus, converting the study of pancreatic islets into a relevant tissue to find a functional explanation of T2D and other related traits susceptibility loci (**see Polymorphic Inversions and TIGER publication**).

1.3.4. Epidemiology and Treatments

The combination of the environmental and genetic factors that affect the metabolic pathways involved in the glucose uptake process favours the development of T2D, and confers its **high incidence** and **prevalence** in the population. Particularly, according to the Epidemiological International Diabetes Federation, over 463 million adults suffered from T2D worldwide in 2019. Moreover, T2D prevalence is projected to increase by 25% in 2030 and 51% in 2045 globally, independently of the different ethnic predispositions to develop the disease (Galicia-Garcia et al., 2020; Saeedi et al., 2019). Additionally to the high prevalence and incidence of this disease, there are

several **comorbidities** associated with T2D. Particularly, T2D is a major risk factor for the development of **cardiovascular disease, hypertension, chronic kidney disease, renal disease, depression, thyroid gland diseases, chronic obstructive pulmonary disease, lower limb amputations, and blindness** (Cornell, 2015; Defronzo, 2009; Nowakowska et al., 2019). As a result, T2D has led to over 4.2 million deaths in 2019. Consequently, T2D is considered a major global health problem which has been further discussed in terms of its prevention and treatment.

Particularly, the effect of diet and exercise has been broadly studied in T2D in terms of predisposition, prevention but also, as a treatment for the disease (Hu, 2011; Magkos, Hjorth, & Astrup, 2020). For example, the inclusion of **healthy lifestyle** changes benefits the prevention and delay of T2D (Nathan et al., 2009). More specifically it has been proved that **weight loss** (~15 Kg) and **fitness** can contribute to a remission of the disease in over 80% of the patients, thus reducing its prevalence worldwide. As a consequence, although the most severe diabetic cases still need to be treated with **insulin replacement therapies**, to maintain the glycemic control, different therapies are still being proposed to prevent and treat this complex disorder (Nathan et al., 2009; Rhodes, 2005). More specifically, it is known that the **reduction of islet cell oxidative stress** can partly reverse the functional impairment of diabetic islets (Del Guerra et al., 2005). Additionally, **beta-cell transplantation** and **regeneration** therapies are currently being proposed as promising to treat and even to cure insulin-deficient diabetes (Ji, Lu, Xie, Yuan, & Chen, 2022). However, the many different challenges surrounding these methods still prevent its broad application in the clinics.

1.3.5. Disease heterogeneity

The diversity of factors affecting T2D defines its heterogeneous nature. Particularly, the understanding of this disease heterogeneity is fundamental to improve its prognosis and treatment under the scope of personalised medicine. In this direction, many efforts have been conducted to generate a **classification of T2D patients** in subgroups based on a wide range of clinical and genomic measures. As a result, this disease heterogeneity has been associated by recent studies with the heterogeneous contribution of **different processes and pathways** to the disease (Ahlqvist et al., 2018; McCarthy, 2017), the major **clinical parameters** involved in the development of the disease (Ahlqvist et al., 2020), and by clustering the **genomic variants** shared by diabetic individuals (Ahlqvist et al., 2018, 2020; Dimas et al., 2014; H. Kim et al., 2022; Mahajan, Wessel, et al., 2018; Mansour Aly et al., 2021; Scott et al., 2017; Udler et al., 2018). These last classifications are the promise of the future steps to a better understanding of the disease pathogenicity and towards precision medicine.

HYPOTHESIS AND OBJECTIVES

2. Hypothesis and Objectives

Complex diseases are a global major health problem that affects millions of individuals worldwide. Therefore, as the understanding of the effects of genetic variation in the development of complex diseases can contribute to a better disease prognosis, its genomic study has been one of the major goals of Biomedicine. More specifically, the knowledge obtained from the study of disease predisposition can facilitate the early detection, prevention and posterior treatment. In this direction, a more detailed explanation of the underlying genetic and molecular mechanisms of complex diseases is known to be a crucial step towards precision medicine. Particularly, in the case of T2D, its genomic study through GWAS has identified more than 700 genetic variants associated with this complex disease. However, although T2D heritability is estimated to be around 70%, these findings still only explain approximately 20% of it, do not enhance the early detection of the disease when compared with clinical measures, and most of them still lack of functional interpretation. For these reasons, there is still room to improve the identification of new associated variants, as well as to determine their functional mechanisms. This thesis aims at directly contributing to these two fronts. On one side, we searched for candidate variant interactions that are associated with disease (T2D), contributing with new genes for the generation of polygenic risk scores, as well as with insights and functional interpretation of the potential functional interaction. On the other side, we also aimed at generating resources to facilitate and improve the functional interpretation of associated variants, which constitutes one of the major bottlenecks with the study of complex diseases.

Accordingly, our hypothesis can be summarised as:

- Hypothesis 1: part of the explanation of the missing heritability of complex diseases is attributable to methodological limitations of GWAS. These limitations can be partially overcome by considering potential variant-variant interactions.
- Hypothesis 2: the integration of homogenized gene expression variation results from pancreatic islets with other functional databases into a comprehensive resource can benefit the functional interpretation of T2D disease-susceptibility loci.

To disentangle these hypotheses the main objectives of this thesis are:

- Objective 1: to provide new variants associated with T2D by exploring the variant-variant interaction space with Machine Learning approaches combined with statistical methods.
- Objective 2: to infer potential functional interpretation of the candidate variant interactions identified as linked to the disease.
- Objective 3: to contribute to the functional interpretation of variants through the performance of a large-scale expression analysis on pancreatic islet samples and the integration of the results with comprehensive data on T2D association studies.
- Objective 4: to generate a comprehensive database and a web portal for the entire community that grants the efficient access and interpretation of pancreatic islets expression regulatory variation.

REPORT FROM THE DIRECTOR

3. Report from the director

The director of this thesis, Dr. David Torrents Arenales informs that:

Lorena Alonso Parrilla is presenting her PhD. Thesis entitled “From the discovery of epistatic events in Type 2 Diabetes Mellitus to the related study of gene expression regulatory variation”, which has been developed at the Barcelona Supercomputing Center (BSC). During her PhD., Lorena has contributed to two studies, including one published as a co-first author. These studies represent the main work of her thesis. Additionally, she has coordinated and pushed a review on GWAS methodology, and has also participated in an external collaboration that ended up in another publication. These two last studies are included in the Appendix of this thesis. In general, Lorena’s contribution to the studies has consisted in the performance of bioinformatic analysis to respond to biological questions, to provide the community with a web platform to enable and facilitate the interpretation of Type 2 Diabetes gene expression variation, and to explain the mathematical models underlying GWAS.

Here below, you can find the scientific contribution made by the PhD. Student in each of the studies, as well as the impact factor of the journals.

3.1. Epistasis (Unpublished)

3.1.1. Title

The role of epistasis to improve the missing heritability explanation and to refine the predictions in Type 2 Diabetes

3.1.2. Authors

Lorena Alonso, Ignasi Morán, and David Torrents.

3.1.3. Contribution

An important research line in our group is the analysis and the discovery of epistasis in the genetics of complex diseases at genome wide level. This line is composed of two different fronts, each using different methodologies. Whereas Dr. Moran is coordinating the overall line, Lorena Alonso is responsible for one of these fronts, which is based on the analysis of Epistatic events using Machine learning approaches. Because this study corresponds to the last activity during Lorena’s thesis, it is still not published, although the results obtained so far are promising and already pointing towards the submission of a potential publication soon.

3.2. TIGER publication

3.2.1. Title

TIGER: The gene expression regulatory variation landscape of human pancreatic islets

3.2.2. Authors

Lorena Alonso*, Anthony Piron*, Ignasi Morán*, Marta Guindo-Martínez, Sílvia Bonàs-Guarch, Goutham Atla, Irene Miguel-Escalada, Romina Royo, Montserrat Puiggròs, Xavier Garcia-Hurtado, Mara Suleiman, Lorella Marselli, Jonathan L.S. Esguerra, Jena-Valéry Turatsinze, Jason M. Torres, Vibe Nylander, Ji Chen, Lena Eliasson, Matthieu Defrance, Ramon Amela, MAGIC, Hindrik Mulder,

Anna L. Gloyn, Leif Groop, Piero Marchetti, Decio L. Eizirik, Jorge Ferrer, Josep M. Mercader[#], Miriam Cnop[#], and David Torrents[#].

* These authors contributed equally

[#] These senior authors contributed equally

3.2.3. Journal

Cell Reports, 2021

Impact factor (Scientific Journal Rankings 2021): 4.845 (Q1)

Citations (Google scholar): 15

3.2.4. Contribution

This study emerges within the Horizon 2020 T2DSystems project, which was devoted to study Type 2 Diabetes from a genetic and clinical point of view, focusing on Pancreatic Islets. Lorena joined the group at the moment when this project started, and she soon became very active at different levels.

In particular, Lorena's contribution to that project was focused on the construction of the Translational human pancreatic Islet Genotype-tissue Expression Resource (TIGER), and the creation of a publicly accessible portal to facilitate the access to this valuable resource of information. Her work can be summarised in three blocks which consist of: 1) the preparation of the data, pipelines, and the analytical environments to obtain islet gene expression regulatory variation results, 2) the creation of a database to collate the genomic, transcriptomic, and epigenetic results obtained from different analyses conducted in human pancreatic islets, as well as other relevant publicly available genomic information, and 3) the population of the database and the creation of the TIGER Portal to make this valuable resource of information accessible for the research community.

In terms of data collection, she was granted access to all the available human islet samples, from the different groups participating in the project, that were planned to be included and analysed in the study. That consisted of 514 samples from pancreatic islet donors distributed in 5 cohorts. From each sample the RNA-seq paired reads, genotyping array information and metadata was gathered. Then, she used different tools for sample harmonisation and quality control processes to ensure the quality of the samples, avoid samples presenting contamination or mismatching samples, and to ensure the good quality of the genotyping array data at the level of the individual and at the level of the sample. All this work was done under the direct guidance and supervision of Dr. Mercader.

After this process, to perform islet gene expression regulatory variation analyses, first she prepared the genotyping array data to recover the individual's genotype using phasing tools. Then, she used these haplotypes to increase the number of variants included in the expression analyses to a genome-wide level by using imputation tools, separately for each cohort. She merged the imputation results into a single cohort containing over 22 million variants ready for the eQTL (mainly done by Anthony Piron) and cASE (mainly done by Ignasi Morán) expression analyses. She used the ASE pipeline, under the guidance and supervision of Dr. Morán, to analyse the RNA-seq of all the human islets included in the project, and prepared the results for cASE analysis. Moreover, she used the RNA-seq data to obtain the gene expression counts needed to calculate eQTL, cASE, and homogeneous pancreatic islet expression. The eQTL and cASE analyses were performed by Anthony Piron and Dr. Morán, respectively. All this work was guided and supervised by Dr. Mercader. She homogenised and normalised the TPM expression counts to obtain homogeneous pancreatic islet gene expression, and then scaled them to be comparable with the 54 tissue expression counts from the GTEx. This work was guided and supervised by Dr. Morán.

Finally, to create a comprehensive islet expression publicly available database, she was granted access to the results from epigenetic and transcriptomic studies from human islets, from the different groups participating in the project. Moreover, she downloaded publicly available T2D GWAS meta-analysis results, variant, gene, pathway, disease association and functional impact genomic databases. She collated all this information with the results obtained from eQTL, cASE, islets gene expression, and GTEx tissues expression. She used this data to populate an Elasticsearch database, which was made accessible through an ICGC-code based web portal. She prepared the environments to install the database and the portal under the supervision of Romina Royo. She adapted the website, and provided it with different graphical and visualisation tools to facilitate the integration and interpretability of this wealth of data. The quality control of the portal was supervised by Dr. Mercader, and correspondingly, by any of the co-authors providing the data.

3.3. Polymorphic Inversions publication

3.3.1. Title

Polymorphic Inversions Underlie the Shared Genetic Susceptibility of Obesity-Related Diseases

3.3.2. Authors

Juan R. González, Carlos Ruiz-Arenas, Alejandro Cáceres, Ignasi Morán, Marcos López-Sánchez, **Lorena Alonso**, Ignacio Tolosana, Marta Guindo-Martínez, Josep M. Mercader, Tonu Esko, David Torrents, Josefa González, and Luis A. Pérez-Jurado.

3.3.3. Journal

The American Journal of Human Genetics, 2020

Impact factor (Scientific Journal Rankings 2021): 5.042 (Q1)

Citations (Google scholar): 6

3.3.4. Contribution

This project was part of a long trajectory of collaborations with Dr. González (ISGlobal, Barcelona) resulting in a publication in 2020.

Lorena's contribution to this study was focused on the provision of support for the functional interpretation of the resulting inversions obtained from obesity-diabetic associations. Particularly, she provided the genotype probabilities for different sets of variants, located in obesity-diabetes inversion regions, from the 70KforT2D diabetes cohorts. This information was used by Dr. González to perform association analysis and support the findings obtained from the UKB, which suggested that obesity-diabetes associated inversions can explain a fraction of T2D shared susceptibility that cannot be explained by single variants. Moreover, she calculated and provided the gene expression counts, normalised TPM and genotyping array data from 207 pancreatic islets donors. This data was then used by Dr. González to understand the possible transcriptomic effects of obesity-diabetes associations with inversions.

3.4. Genome Wide Association Studies review

3.4.1. Title

In Search of Complex Disease Risk through Genome Wide Association Studies

3.4.2. Authors

Lorena Alonso, Ignasi Morán, Cecilia Salvoró, and David Torrents.

3.4.3. Journal

Mathematics, 2021

Impact factor (Scientific Journal Rankings 2021): 0.538 (Q2)

Citations (Google scholar): -

3.4.4. Contribution

This project was an invitation to a review, which started in 2020 and continued until its publication in 2021.

Lorena pushed and coordinated this review exercise. She gathered and reviewed the information and coordinated the writing, which also involved Dr. Salvoró and Dr. Morán. This review presents an overview of the past and current statistical methods used in GWAS field, discuss current practises and their main limitations, and describes the remaining open challenges.

In particular, she collected and read multiple GWAS publications and information to have an overview of the state-of-the-art of the methodology. She studied the different statistical approaches currently applied in GWAS and their limitations from a mathematical point of view. Then, she further explored the most common GWAS complementary methods which are broadly applied to overcome its limitations. Particularly, all this information was used to write a methodological review briefly detailing the mathematical models used in GWAS and summarising its current limitations and available complementary analyses. All this work was revised by Dr. Morán and Dr. Salvoró, and supervised by Dr. Torrents.

SUMMARY OF THE STUDIES

4. Summary of the studies

4.1. Epistasis (Unpublished)

Complex diseases develop as a result of the combination of the simultaneous effect of multiple environmental and genomic factors (Manolio et al., 2008). Particularly, at the genomic level, despite the large amount of variants that have been discovered during the last decades associated with complex diseases, these findings only explain a small fraction of disease heritability (Génin, 2020). Moreover, the utilisation of this knowledge to improve the prediction of the risk of developing common diseases is still far from being usable within the clinical field (Kullo et al., 2022; Kumuthini et al., 2022; Lambert et al., 2019). This is, in part, because of the limitations and simplifications of Genome Wide Association Studies (GWAS) strategies (Alonso, Morán, et al., 2021; Tam et al., 2019). For example, due to computing limitations, GWAS considers within the analysis the effect and role of each single variant as independent within the disease, which is actually far from reality. In complex diseases, many loci (and therefore many variants) are expected to contribute to the risk and development cooperatively, both additively and in a synergic dependent manner (epistasis). In this direction, although still incomplete, polygenic risk scores have enhanced the prediction and prevention of complex diseases, by using the additive model to combine the GWAS effects of multiple variants (Kullo et al., 2022; Kumuthini et al., 2022; Lambert et al., 2019). Additionally, the analysis of the epistatic interaction between variants has been crucial towards a better understanding of complex diseases (Manduchi et al., 2018; Josep Maria Mercader et al., 2008). It is therefore necessary to incorporate the interaction of variants within association studies to broaden the study of complex diseases, to analyse not only the effect of single variants, but also of pairs, trios, and bigger groups. Nevertheless, the genome wide study of these interactions using classical statistical frames still represents a computational challenge, as the analysis of combinations of variants increases by several fold the computational demands (Marchini et al., 2005).

In order to overcome these limitations, we designed a study that is focused on the analysis of epistasis in Type 2 Diabetes (T2D), by using machine learning (ML) models, in combination with statistical approaches. More specifically, to find groups of candidate variants associated with T2D, we have used XGBoost (T. Chen & Guestrin, 2016), a ML classifier based on random forest. Although XGboost can be used as a predictor, we are only focusing on the groups of variants associated with the disease that have been identified by the method. As an input for the method, we have used a subset from the 70KforT2D (Bonàs-Guarch et al., 2018), a large T2D dataset which was previously generated and analysed within the group using GWAS strategies. In particular, after a quality control on the individuals to ensure a good performance of the model, a group of 22,802 individuals, where 11,401 are diabetic and 11,401 are non-diabetic were selected. Moreover, to deal with some ML models limitations, we have reduced the number of the initial set of variants, starting from 105,896 variants which have some degree of association with diabetes ($-\log_{10}(p\text{-value}) > 2$). Under this background, XGBoost is used to find individual variants and groups of 2, 3, and 4 variants which are synergically associated with diabetes.

Among the thousands of groups obtained in this preliminary analysis, there are some groups which contain variants that can contribute additively to the disease, and other groups from which the effect of variants on the disease derives from the interaction. Because our initial goal is to identify examples of epistasis, the effect of the interaction is evaluated under a logistic regression model. We only kept the groups containing an interaction statistically associated with T2D ($\alpha = 0.05$), thus resulting into 10 pairs, 1 triplet, and 1 quadruplet. Under the premise that the effect of the sum of each variant separately should be smaller than the effect of the variants together, we have performed logistic regression analysis to demonstrate that, certainly, the variants show epistatic effect. From these analyses we have also observed some differences in the marginal effects of the variants when evaluated synergically. Remarkably, some of these variations can result in a change of sign in the

effect, thus involving an inverse effect of the variant; for example, changing from being protective to representing a risk on the predisposition to disease. Finally, we have functionally inspected the interactions using the summary statistic annotations of diverse large T2D GWAS meta-analyses, glycemic traits GWAS meta-analysis, and regulatory expression variation from pancreatic islets, which is a disease-related tissue (Alonso, Piron, et al., 2021; Bonàs-Guarch et al., 2018; J. Chen et al., 2021; Mahajan, Taliun, et al., 2018; Miguel-Escalada et al., 2019; Pasquali et al., 2014; Scott et al., 2017; The DIAGRAM Consortium et al., 2014). The results suggest that the interactions between the underlying regulatory mechanisms of the variants inside the groups, as well as the connections of the gene pathways affected can be one of the causes to explain disease predisposition.

Overall, the results obtained from this study show the relevance of including epistasis in current association studies to improve the explanation of the heritability of complex diseases, to enhance current detection and prevention protocols, and to gain insight of complex diseases pathophysiology.

4.2. TIGER publication

The simultaneous effect of multiple genomic and environmental factors affects the development of complex diseases (Manolio et al., 2008). At the genomic level, one of the most relevant and challenging parts of the study of the genetic architecture of complex diseases is the functional interpretation of the variants found to be statistically associated with the trait from GWAS studies. Particularly, to improve the comprehension of the pathophysiology of this type of disorders, it is necessary to find and understand which are the diverse underlying molecular mechanisms of disease-associated loci, usually in the form of identifying the affected gene and protein. This knowledge enhances the discovery of new drugs, and promotes the creation and the improvement of protocols for disease treatment. However, the outcomes of current association methods (i.e. GWAS) are limited to the provision of a list of disease susceptibility loci, and their contribution to the risk of disease development (Alonso, Morán, et al., 2021). Therefore, the understanding of their functional mechanisms requires additional approaches and efforts (Cano-Gamez & Trynka, 2020; Lichou & Trynka, 2020; Manolio, 2013).

During the last decades, gene expression variation and regulatory regions analyses have promoted the development of large databases, listing numerous associations between variants (loci) and their change in gene expression, such as expression quantitative trait loci (eQTL), and cataloguing disease-related regulatory elements, such as enhancers and promoters, thus enabling a better understanding of complex diseases (Han et al., 2015; Jiang, Xuan, Zhao, & Zhang, 2007; Papatheodorou et al., 2020; The GTEx Consortium, 2020). Remarkably, as the expression analysis is linked to tissue-specific functions, despite the many difficulties derived from the analysis of specific tissues, the study of disease-related tissues has improved the functional interpretation of these signals. However, this information only covers the regions of the genome that have been analysed in these studies, thus leaving some signals without a functional explanation, and more importantly, some tissues or groups of cells have been disregarded or still need to be further inspected.

To improve the genomic understanding of diseases related to pancreatic islets dysfunction, this study is focused on the analysis of genomic variation and its effect on gene expression in human pancreatic islets. In particular, it is accepted that Type 2 Diabetes (T2D) is mainly caused by dysfunctions within the beta cells of the pancreas, making this tissue a key target for the study of the disease (Bartolomé, 2022; Del Guerra et al., 2005; Eizirik et al., 2020; Gloyn et al., 2022). Within the context of the T2DSysTems, a European Project, we developed the Translational human pancreatic Islets Genotype-tissue Expression Resource (TIGER), a large human islet regulatory expression database (<http://tiger.bsc.es/>). This database integrates, in a unique platform, the results obtained from the performance of extensive expression, eQTL, and combined allele-specific expression

analyses (cASE), with publicly available summary statistics results from islets analyses, including expression array, regulatory elements, and other gene, variant, and disease functional information (Akerman et al., 2017; Alonso, Piron, et al., 2021; Bonàs-Guarch et al., 2018; Buniello et al., 2019; Frankish et al., 2019; Jassal et al., 2020; Karczewski et al., 2020; Mahajan, Taliun, et al., 2018; McLaren et al., 2016; Miguel-Escalada et al., 2019; Pasquali et al., 2014; Piñero et al., 2017; Scott et al., 2017; Solimena et al., 2018; The DIAGRAM Consortium et al., 2014; The Gene Ontology Consortium, 2000; The GTEx Consortium, 2020; Thurner et al., 2018).

As a first effort to generate this platform, the genotypes and phenotypes (RNA-seq) of 514 human pancreatic islets samples from mostly non-diabetic individuals were collected. As the inspection of expression in islets requires to know not only the expression of any given gene but also its comparison with the rest of the genes in the genome, we calculated, harmonised, and homogenised gene expression among all the non-diabetic individuals. Then, to facilitate the comparison between islets expression and other reference tissues, we aggregated and scaled the gene expression measures from islets and the Genotype-Tissue Expression project (GTEx) (The GTEx Consortium, 2020). In addition, to promote an exhaustive inspection of the effects of genomic variation in islets gene expression, the imputed genotypes of more than 22 million variants were prepared for cASE, and eQTL *cis*-regulatory expression analyses, including a 10% of Indels and Structural Variants, more than 1.05 million variants in the chromosome X, and more than 14 million rare and low-frequency variants. This resulted in over 1.11 million eQTLs and 256,981 cASE variants. Next, to facilitate the assessment of the overlap between variation and islet regulatory elements and open chromatin regions, diverse DNA methylation, human islet regulome, long non-coding RNA, ATAC-seq and CHIP-seq results were collated (Akerman et al., 2017; Miguel-Escalada et al., 2019; Pasquali et al., 2014; Thurner et al., 2018). Finally, to enhance the interpretation of the potential functional impact of variants, we collated the variants with the GWAS Catalog and T2D GWAS meta-analyses summary statistics, and with their functional impact on genes (Bonàs-Guarch et al., 2018; Buniello et al., 2019; Mahajan, Taliun, et al., 2018; McLaren et al., 2016; Scott et al., 2017; The DIAGRAM Consortium et al., 2014).

As a result, this platform contains information for more than 27 million variants and 59,625 genes and facilitates the search at the level of the gene and at the level of the variant. It encloses tools for visualising, querying, and downloading human islet data enhancing the study of T2D and other islet-related diseases pathophysiology. It includes eQTL and cASE results, and associations with T2D and other complex diseases from the GWAS Catalog, thus simplifying the analysis of colocalisation. Moreover, it integrates graphs to enhance the inspection of gene expression in pancreatic islets and its comparison with other tissues, and a genomic browser to explore the genomic context information.

In summary, the database generated in this study represents a unique and formidable resource to interrogate the molecular aetiology of beta-cell failure.

4.3. Polymorphic Inversions publication

The development of a complex disease is attributed to the combined effect of different genetic and environmental factors (Manolio et al., 2008). Genetically, despite the large catalogue of variants that have been discovered during the last decades associated with complex diseases, only a small fraction of their heritability has been explained (Génin, 2020). Thus, resulting in an impact on the effectiveness of current detection and prevention protocols (Kullo et al., 2022; Kumuthini et al., 2022; Lambert et al., 2019). This lack of explanation is usually attributed to the limitations surrounding the methodology used in association studies. Among others, the inclusion of structural variants in this type of studies, and a better control on the effect related to the presence of covariates are two of the causes to lose information (Génin, 2020). In particular, inversions can affect the gene function if they

overlap to the inversion breakpoints, thus suggesting a putative pathway to disease predisposition. Moreover, the analysis of disease association, under the presence of covariates and comorbidities, reduces the results obtained to just recover the most common variants shared between diseases or related conditions. Therefore, some variants with a lower frequency, which are already known to be associated with a certain disease, can be masking a shared susceptibility for other related diseases (González et al., 2020).

To improve the genomic understanding of co-occurrent traits, this study evaluates the effect of inversions in the shared susceptibility between obesity and other related complex diseases and traits. 21 common inversions are assayed to test their association with 8 comorbidities and 17 related conditions, including obesity, hypertension, asthma, diabetes, and some mental diseases. As a result, 3 of the inversions are found associated with different diseases. Particularly, inversions 8p23.1 and 16q11.2 showed a shared susceptibility between obesity with diabetes, hypertension, asthma, and depression. In contrast, inversion 11q13.2 shares susceptibility between obesity with diabetes and hypertension. Remarkably, the effect of the co-occurrent association is found greater when compared to the individual association with the diseases.

Then, the genetic relevance of these inversions is explored in the 70KforT2D (Bonàs-Guarch et al., 2018), an independent dataset from the discovery data. Particularly, the genotype of the SNPs located in the same region of the inversions are inspected to test their association with obesity and T2D in different subgroups of individuals. As a result, none of the SNPs overlapping the inversion were found significantly associated. Thus, suggesting that single variants are not driving the association. Additionally, at a transcriptomic level, being a disease-related tissue for Type 2 Diabetes, human pancreatic islets genotypes and gene expression are used to reveal any possible relation between the inversions and changes in expression (Alonso, Piron, et al., 2021; van de Bunt et al., 2015). These analyses allowed the identification of some associations between inversions 8p23.1 and 16p11.2 and the deregulation of some well-known genes for diabetes.

In brief, this study provided evidence for the presence of polymorphic inversions associated with several related diseases, and provides preliminary functional interpretation of these signals.

4.4. GWAS review

The combination of the effects of multiple environmental and genetic factors can result in the development of a complex disease (Manolio et al., 2008). Consequently, at a genetic level, the discovery of the genomic variants associated with the risk of developing complex diseases, as well as its functional interpretation, are one of the major goals in Biomedicine. Indeed, the knowledge about the variants that predispose to disease development is crucial to improve the detection and prevention protocols. Therefore, to facilitate the prediction of novel variants associated with complex diseases, a wide diversity of methods and bioinformatic tools have been developed. In particular, during the last two decades, Genome Wide Association Studies (GWAS) have emerged as the key to explore disease and trait associations at a genome-wide level (R. J. Klein et al., 2005). However, despite the great advances made, thanks to the use of current statistical frames, the discovery of novel variants associated with a disease and its interpretation are still one of the big challenges in Biomedicine. To encourage the mathematical community to get involved in this fundamental question and to provide more adjusted and powerful statistical frames, in this review, we inspect the current status in GWAS, detailing the underlying mathematical models, the possibilities, and the limitations (Alonso, Morán, et al., 2021).

The many limitations surrounding these methods, mostly resulting in a lack of statistical power, represent the current boundaries in the discovery. Under a background of genetic heterogeneity, where multiple variants each one with a small effect on the disease are needed to its

development (McCarthy et al., 2008), the common way to gain discovery power is to increase the sample size and the number of genomic variants analysed. As a result, the creation of genetic biobanks and large cohorts, the use of meta-analysis approaches, or the application of imputation techniques has enhanced the discovery (Alonso, Piron, et al., 2021; Lo, 2014; Marchini, 2019; Panagiotou, Willer, Hirschhorn, & Ioannidis, 2013; Swede et al., 2007). However, variants with less presence in the population, such as rare variants, variants from specific or isolated populations, from specific subgroups of individuals, and even variants which present difficulties to be included in the analysis, such as structural variants, despite their relevance, are still difficult to capture (Ahlqvist et al., 2020; J. Chen et al., 2021; Génin, 2020; González et al., 2020). Moreover, current statistical frames only test the independent effects of each variant, while the nature of complex diseases is defined by the synergic effect of multiple variants and environmental factors. To tackle this problem, genomic interaction studies, and environment-wide association studies (EWAS), although still challenging, have emerged promoting the advance towards the discovery of the effect of environmental and genetic interactions in complex diseases.

Additionally, to improve the understanding of disease pathophysiology, it is also necessary to find the relation between variation and the underlying mechanisms of genomic variation. Particularly, this knowledge is crucial for the development of new drugs and to improve the treatments. However, the outcomes obtained from the application of GWAS methodology are reduced to a list of disease-associated variants, their effect on the disease, and a measure of reliability for the association test. As a result, the interpretation of the results obtained from a GWAS is merely reduced to the statistical level, thus resulting in a lack of biological explanation, and making it necessary to use complementary approaches to improve the understanding of GWAS results. In particular, the combination of genomics with other omic layers such as transcriptomics and epigenetics is a valuable tool to translate genomic variation into function. Consequently, expression analyses, gene, pathway, regulatory elements and epigenetic marks enrichment are broadly used methodologies to find the molecular mechanisms underlying complex diseases (Cano-Gamez & Trynka, 2020; Lichou & Trynka, 2020; Manolio, 2013). Moreover, although still in development, other tools such as Polygenic Risk Scores are planned to be applied to convert genomic associations into predictions that can be applied in the clinics (Kullo et al., 2022; Kumuthini et al., 2022; Lambert et al., 2019).

In brief, this review details the current statistical models surrounding GWAS to promote the creation of new frameworks that can facilitate and improve the study of complex diseases.

EPISTASIS

UNPUBLISHED

5. Epistasis (Unpublished)

The role of epistasis to improve the understanding and to refine the predictions in Type 2 Diabetes

Tables and Figures list

Figures

Figure 1. General strategy.

Figure 2 Evaluation of the effect of the candidate epistatic groups on the risk of developing T2D.

Figure 3. Percentage of significant annotations overlap.

Figure 4. Some examples of epistatic variants with a well-known functional interpretation in terms of disease.

Supplemental Figure 1. Evaluation of the performance of XGBoost under case-control imbalance.

Supplemental Figure 2. Evaluation of the performance of XGBoost in terms of randomness.

Supplemental Figure 3. Evaluation of the performance of XGBoost in terms of variable explanation.

Supplemental Figure 4. Evaluation of the performance of XGBoost in terms of variable redundancy.

Supplemental Figure 5. Evaluation of the performance of XGBoost in terms of missingness.

Supplemental Figure 6. Evaluation of the performance of XGBoost in terms of data availability.

Supplemental Figure 7. Evaluation of the performance of XGBoost in terms of overfitting.

Supplemental Figure 8. Minor Allele Frequencies from the variants included in the groups obtained from the different scenarios (50, 100, 250, and 500 trees).

Supplemental Figure 9. Annotations overlap.

Tables

Table 1. Groups of epistatic variants and their effect in T2D.

Table 2. Percentage of unique variants significantly annotated with T2D and glycemetic traits GWAS meta-analyses, islets expression, functional impact annotations, and epigenetic marks.

Supplemental Table 1. Evaluation of the performance of different machine learning methods in a subset of the discovery dataset (1,667 GWAS significant features, 11,401 cases, 11,401 controls).

Supplemental Table 2. Evaluation of the performance of XGBoost under case-control imbalance.

Supplemental Table 3. Evaluation of the performance of XGBoost in terms of randomness.

Supplemental Table 4. Evaluation of the performance of XGBoost in terms of variable explanation.

Supplemental Table 5. Evaluation of the performance of XGBoost in terms of variable redundancy.

Supplemental Table 6. Evaluation of the performance of XGBoost in terms of missingness.

Supplemental Table 7. Evaluation of the performance of XGBoost in terms of data availability.

Supplemental Table 8. Evaluation of the performance of XGBoost in terms of overfitting.

Supplemental Table 9. Evaluation of the relation between candidate epistatic groups of variants by depth and by tree.

Supplemental Table 10. Evaluation of the differences between the marginal effects in the additive logistic regression model and the model including interactions.

Supplemental Table 11. Logistic regression coefficients of 3 examples of variant interaction with a change in variants effect on T2D.

Abstract

Complex diseases are affected by the combination of the simultaneous effect of multiple variants and environmental factors. However, the numerous statistical and computational challenges surrounding the classical approaches used in association studies, has reduced the discovery to a limited group of variants which are associated with common diseases in a single independent manner. As a result, the effect of multiple variants interactions or epistasis has been pointed as one of the causes to explain the missing heritability of complex diseases, as well as for improving the prediction power of the genetic signal towards the use of detection protocols in the clinics. To find groups of epistatic variants that are cooperatively statistically associated with Type 2 Diabetes (T2D), in this study, we have explored the potential of a machine learning strategy, XGBoost, combined with statistical approaches to analyse a group of 11,401 diabetic and 11,401 non-diabetic individuals, and a subset of 105,896 T2D nearly nominally associated variants ($-\log_{10}(p\text{-value}) > 2$) derived from previous GWAS studies in the group. Among the different groups obtained by XGBoost statistically associated with T2D (pairs, triplets, and quadruplets), there are groups which affect the disease in an additive manner, and other groups which include variants which synergically contribute to the disease (epistasis). To find epistatic variants we applied a logistic regression to the results obtained from the machine learning approach, resulting in a group of 10 pairwise variant interactions, 1 variant triplet, and 1 variant quadruplet from which the association is epistatic. In agreement with the definition of epistatic interactions, we validated these results and found that the effect of the interaction is significantly stronger than the sum of the effects of each variant separately. Moreover, although 75% of the epistatic groups contain new susceptibility loci for T2D, the analysis of the overlap of these interactions with T2D and related glycemic traits GWAS meta-analyses, and islet gene expression regulatory variation, reveals multiple gene interaction and islet regulatory elements as the underlying molecular mechanisms mediating the association with T2D. Despite many improvements having to be applied to enhance the detection possibilities, these preliminary results evidence the potential of using machine learning approaches to study epistasis in complex diseases and to gain insight of their genetic pathophysiology, and consequently, to improve its prognosis and treatment.

Introduction

Complex diseases such as diabetes, asthma, or Alzheimer's disease, are known to be affected by the combination of multiple genetic and environmental factors (Manolio et al., 2008). Particularly, during the last decades, the study of the genetic component of complex diseases based on Genome Wide Association Studies (GWAS), has led to the discovery of thousands of variants associated with different complex traits or diseases (Beck et al., 2014; Buniello et al., 2019; K. Watanabe et al., 2019). However, despite GWAS having revealed a large catalogue of disease-associated variants, only a small fraction of the heritability of complex diseases has been uncovered (Génin, 2020). Moreover, regardless of the complex disease nature, where the combination of multiple genetic and environmental factors predispose the individual to develop the disease, these variants have been found associated with the disease in a single independent manner, thus, limiting current detection and prevention protocols and, therefore, distancing the translation of the results into the clinics (Alonso, Morán, et al., 2021; Tam et al., 2019; Uffelmann et al., 2021). This is the case, for example, of Type 2 Diabetes (T2D), a complex metabolic disorder which affects over 465 million people worldwide.

Particularly, the study of the genetic component of T2D during the last decades, based on GWAS results, has led to the discovery of more than 700 independent signals associated with the disease (Bonàs-Guarch et al., 2018; J. Chen et al., 2021; Mahajan, Taliun, et al., 2018; Scott et al., 2017; The DIAGRAM Consortium et al., 2014; Vujkovic et al., 2020). However, despite the extensive T2D genomic knowledge that these GWAS findings represent, there is still a lack of explanation for its complete genetic heritability. More specifically, the heritability of T2D based on twin studies has been estimated to range from 0.3 to 0.72 (Newman et al., 1987; R. M. Watanabe et al., 1999; Willemsen et al., 2015). However, the contribution of these loci to its missing heritability explanation is approximately 0.2 (DeForest & Majithia, 2022). And, more importantly, there is not enough information to be able to efficiently predict for a particular individual the real risk of developing the disease. In other words, the results obtained with classical GWAS, despite uncovering a number of genetic determinants and resulting in a good prediction (AUC=0.901), still cannot be used at a clinical level for prevention or for patient stratification, and most importantly, do not improve the prediction that can be obtained from the use of clinical variables (Collins et al., 2021; Kullo et al., 2022; Kumuthini et al., 2022; Liu et al., 2021; McGuire et al., 2020; Padilla-Martínez et al., 2020). Therefore, highlighting not only the relevance of the genomic study of T2D but also of its current limitations.

Overall, the factors contributing to the development of the disease are genomic variants, with the presence of gene-environment interactions and gene-gene interactions (Génin, 2020; Herzig, Clerget-Darpoux, & Génin, 2022). In the last case, the study of interactions in complex diseases can be tackled through the analysis of the non-independent effect of specific groups of variants, beyond the simple addition of their effects separately (Mackay, 2014). This type of phenomena, which is known as epistasis, has its biological basis on the known networks between regulatory elements and interconnected pathways, where the change (variation) of the function and impact of one gene (i.e. protein) can enhance the change of function in another gene, converging cooperatively into a synergic effect. Multiple methodologies have been applied to the study of epistasis in complex diseases, ranging from statistical to artificial intelligence approaches (Niel et al., 2015). In short, these methods are able to detect which are the groups of variants that synergically contribute to the development of the disease. Particularly, these methods have been applied to study epistasis in small groups of variants, only including loci functionally related with the disease, or selecting some variants using dimensionality reduction techniques (Behravan et al., 2018; Y. M. Cho et al., 2004; Kirino et al., 2013; Manduchi et al., 2018). The success obtained from these reductions to find variants with an increased effect jointly (Cordell, 2009; Kirino et al., 2013; Monir & Zhu, 2017), suggests the potential of epistasis to improve the knowledge about complex diseases, opening an avenue to cover the analysis of variant interactions at a genome-wide level. However, the numerous difficulties related to the detection power, and other computational problems, have limited the discovery. Specifically for T2D,

although certain studies have addressed some of the limitations surrounding the genome-wide analysis of epistasis, still no credible evidence for interaction effects has been found (Nag, McCarthy, & Mahajan, 2020).

In order to overcome these limitations and to obtain more integrated results, we explored the use of machine learning approaches for the identification of groups of variants that are cooperatively associated with the risk of developing complex diseases. From this analysis we aim to provide clear examples of epistatic interactions that can improve disease risk prediction. Following previous research and experience built in the group this exploratory analysis is focused on T2D. Particularly, this study first targets the discovery of groups of interacting variants, which can contribute additively to T2D, or that have an effect produced by the dependency (i.e. epistatic) relation between the variants. For this, we analyse with XGBoost (T. Chen & Guestrin, 2016), a machine learning classifier based on random forest, a subset from the 70KforT2D (Bonàs-Guarch et al., 2018), a large T2D multi-cohort dataset. This subset contains genotypes and basic phenotypic information for 22,802 European individuals (11,401 cases and 11,401 controls) and 105,896 variants. The results obtained here will contribute to gain understanding of the effect of epistasis in complex diseases, to improve the explanation of the missing heritability of T2D and, ultimately, to clinically predict the risk of developing this disease.

Results

Overall strategy

To enhance the detection of genetic factors that can improve the prediction of the risk of developing a complex disease, we first focused on the discovery of candidate groups of epistatic variants. There are different methods that can be applied with the purpose of finding groups of variants which synergically contribute to the risk of developing the disease, ranging from the most classical statistical approaches to the application of machine learning techniques. To avoid the limitations related to the discovery power derived from the use of statistical methodologies, we applied machine learning methods (Nag et al., 2020; Niel et al., 2015). In this first preliminary approach, we decided to explore the effects of these synergies in T2D using a supervised machine learning classifier. Supervised classifiers start from a group of observations that can be separated into different categories, for example cases and controls, to learn which are the most informative variables to generate each category. Therefore, the results that we expected from the analysis consisted of a classification of the individuals in groups of diabetics and non-diabetics, a list of the most relevant variants and groups of variants that were used to do the classification, and their corresponding scores. The groups of variants found by the method can contain variants that have an effect on the disease only in an additive manner, as well as variants that act synergistically. Following our goal of identifying epistatic events in T2D, i.e. groups of variants where their combined effect was higher, or lower than the effect of the sum of their corresponding effects obtained independently, we then applied logistic regression analysis (**Figure 1**).

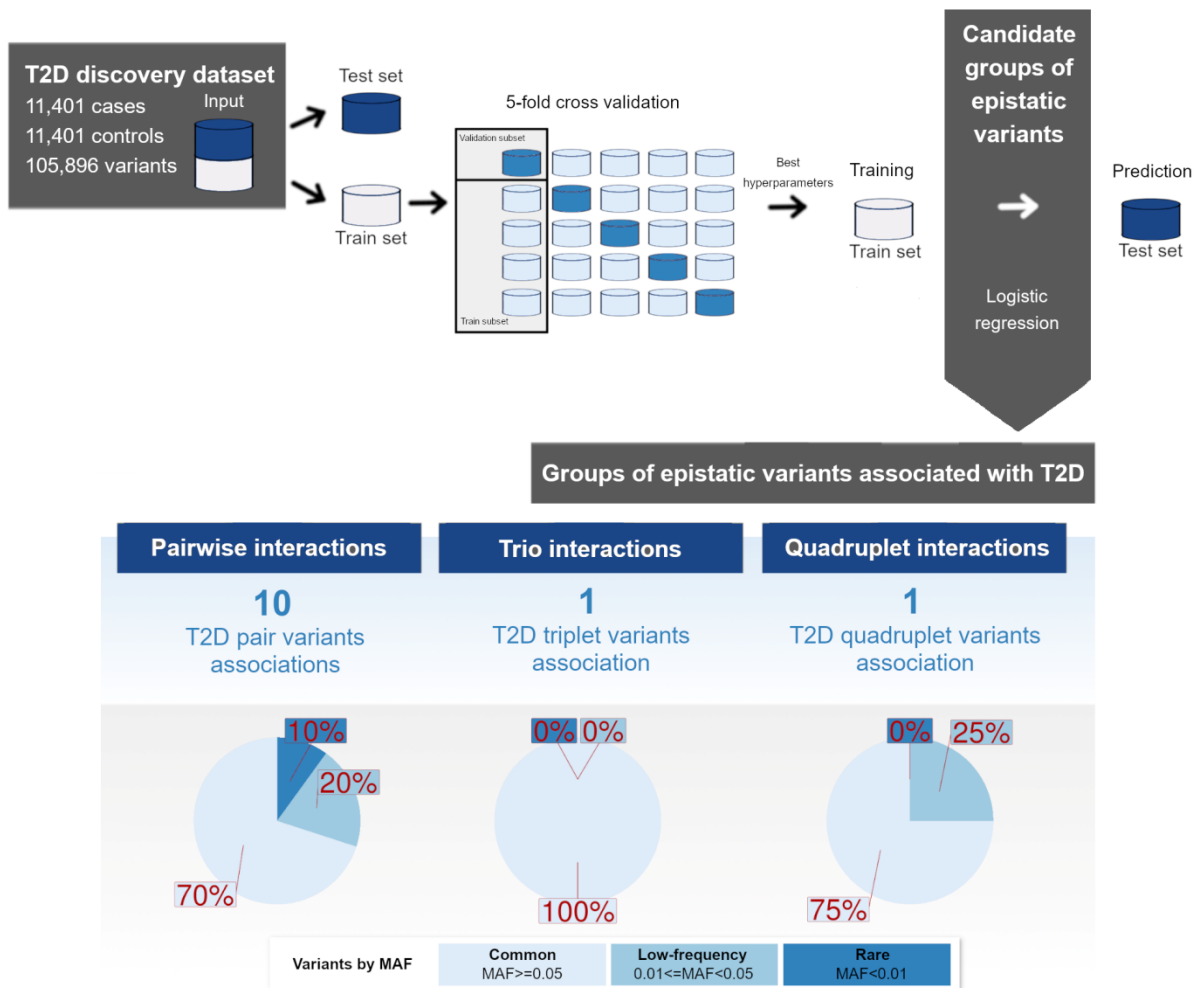


Figure 1. General strategy. The T2D discovery dataset, which contains the imputed genotype of 105,896 variants ($-\log_{10}(p\text{-value}) > 2$) for 22,802 individuals (11,401 diabetic and 11,401 non-diabetic), was divided into a

training and a test subsets. To ensure the best performance of the model, and to prevent overfitting a 5-fold cross-validation algorithm was implemented to do a grid search hyperparameter adjustment. The best hyperparameters were used to fit the train set, in the training step, and the performance was evaluated using the test set, during the test step. As a result, several groups of single variants, pairs, triplets and quadruplets of candidate epistatic variants associated with T2D were obtained. From these, only the groups that presented a significant association between the interaction and T2D, under a logistic regression model, were kept. The pie charts show, for each group, the percentage of variants classified by Minor Allele Frequency (MAF). Common variants ($MAF \geq 0.05$) are represented in light blue, low-frequency ($0.01 \leq MAF < 0.05$) in medium blue, and rare variants ($MAF < 0.01$) in dark blue.

In terms of the input data, there are many factors that can affect the performance of a machine learning approach, which include the number of observations that are available to do the training, the presence of missing values and outliers, redundancy or the existence of any type of imbalance which can result in trend decisions for the method (Chicco, 2017; Dey, 2016; Greener et al., 2021; Sarker, 2021). Therefore, a good previous knowledge of the type of the data included in the analysis is crucial for the correct preparation of the input dataset that will be used by the classifier; this will facilitate the creation of a trustworthy model to solve our problem. In our case, to find groups of candidate epistatic variants in T2D, we decided to use the 70KforT2D study, a large T2D genome-wide association studies (GWAS) meta-analysed cohort within our group (Bonàs-Guarch et al., 2018). The data used for our analysis includes the imputed genotypes of the individuals from the five cohorts analysed in the 70KforT2D project, the metadata of these individuals, and the GWAS summary statistics (**Methods T2D case-control dataset**). After merging the individual genotypes in a unique cohort, to avoid any of the above mentioned problems and other computational limitations we did a data pre-processing step (**Methods Dataset preparation; Suppl. Figures 1-6; Suppl. Tables 2-7**). First, as our data is completely imbalanced for the case-control proportion (12,926 diabetic and 57,191 controls), we kept the maximum number of diabetic individuals from the 70KforT2D that pair with a control sharing the same measures of (body-mass index) bmi, age, and sex (**Methods Data imbalance**). This resulted in a dataset with 22,802 individuals from which 11,401 diabetic and 11,401 non-diabetic. Despite this is a large number of observations, our aim of performing the analysis at a genome-wide level, thus involving the inclusion of 15,131,345 imputed variants, results in an overfitting problem (**Methods Maximisation of variables explanation**). To prevent overfitting and to make our analysis possible from the computational point of view, we limited our study to only kept the variants with higher levels of association with T2D (i.e. with $-\log_{10}(p\text{-value}) > 2$), as we expect them to be enriched in functionally relevant interacting groups. As a result, our discovery dataset consisted of 22,802 individuals with their imputed genotypes in 105,896 genomic variants.

There are different types of supervised machine learning classifiers that can be applied to find groups of synergic variants associated with a disease. However, the performance obtained from each classifier varies based on their underlying statistical models and, therefore, on the way the method evaluates the input data. These many factors related with the input data range from the ability of the method to understand and manage the type of data included in the input, to the capacity of working with missing values or duplicate observations (Dey, 2016; Greener et al., 2021). In our case, although all the individuals are of European ancestry, and despite working with imputed genotypes, the heterogeneity in the population of study generates missing values in some of the genotypes. Moreover, the genetic background of linkage disequilibrium results in correlated variants, which can be interpreted as duplicates by the model. Therefore, among the multiple supervised machine learning classifiers that were assayed, we selected the method that was better prepared to work with the genomic information present in our dataset, which include missing values, and correlated data. These methods were evaluated, in terms of precision and computational time, in a subset of the discovery dataset, only including the imputed genotype for 1,667 GWAS significant signals ($-\log_{10}(p\text{-value}) > 7$) (**Methods Method selection**). From all the methods assayed, XGBoost was the one which performed better in the classification of individuals (T. Chen & Guestrin, 2016; Pedregosa et al., 2011) (**Suppl. Table 1**).

One common problem derived from the use of machine learning methods is overfitting, which results in linking the results obtained to the dataset of analysis and, therefore, not allowing the extension to other independent datasets. To prevent the overfitting and to ensure the procurement of the best outcomes, we prepared a test-train model with a previous hyperparameter adjustment using a 5-fold cross-validation algorithm (Chicco, 2017; Greener et al., 2021). To obtain the best performance, the most relevant parameters to adjust in our model were the split, the learning rate, the number of trees, and the depth of the tree. The split corresponds to the percentage of individuals that are kept in the test set to do the final prediction, once the model is trained. Consequently, to ensure that a good proportion of individuals were used to train the model, we allowed the adjustment between 0.2 and 0.3. The learning rate corresponds to the minimum contribution score that is required for a new feature to be included in the final model. Therefore, although a smaller learning rate is computationally expensive, as it can result in more steps for the algorithm to decide which are the best features, we tested different small learning rates including 0.01, 0.04, 0.07 and 0.1. As XGBoost is a method based on decision trees, the number of trees corresponds to the number of decision trees that the method will include in the resulting model after the training, and the depth corresponds to the maximum level of features that will result in a decision for a tree, which in our case correspond to the dimension of the groups of synergic variants. An increase in these two parameters involve the generation of a higher number of trees or more dense trees, respectively, thus resulting in more tests and the inclusion of more features in the final model, which is more computationally expensive and can result in overfitting. To maximise the performance but prevent overfitting, we tested the results obtained by the generation of 50, 100, 250, and 500 trees, and allowed the combination of variant in pairs, trios, and quadruplets (depth ≤ 4) (**Suppl. Figure 7; Suppl. Table 8**) (**Methods Algorithm preparation; Hyperparameters adjustment**).

Once the machine learning method was selected (i.e. XGBoost), the data was pre-processed and the complete machine learning pipeline was prepared, we finally executed the method. As a result, after adjusting for the optimal learning rate and split in each case, we explored the different groups of singletons, pairs, trios, and quadruplets obtained as an outcome from the different scenarios when varying the number of trees. In particular, we compared the Minor Allele Frequency (MAF) of the variants, and studied the possible relations, in terms of inclusion of variants, between the groups obtained in each scenario (**Methods Candidate epistatic groups base genomics**). After the inspection of the MAF we observed that a higher percentage of rare ($MAF < 0.01$) and low-frequency ($0.01 \leq MAF < 0.05$) variants were captured in the scenarios with more groups (**Suppl. Figure 8**). This can be interpreted as a possible indicator of overfitting, however, as the hyperparameters adjustment ensured that our models were not overfitted (**Suppl. Figure 7; Suppl. Table 8**), in this case it indicates that a deeper search of interactions enhances the ability of the method to improve the capture of disease heterogeneity, and broadens the study with the inclusion of rare and low-frequency variants. Then, to find any possible relation between the variants inside the different groups obtained in each scenario, we analysed their linkage disequilibrium (LD) correlation. From the inspection of the inclusion of smaller groups in bigger groups, we observed that between 10.98-57.22% of the variants are preserved through the groups in the same scenario ($r^2 \geq 20$) (**Suppl. Table 9**). However, none of the groups were completely kept. Similar results were obtained from the analysis when only the number of final trees was changed, therefore, comparing groups with the same number of variants between the distinct scenarios, where none of the groups was replicated but some of the loci were preserved for single associations (5.49-9.37%), and also were retained when increasing the size of the group (5.49-83.63%). Therefore, the prevalence of the loci included in the different epistatic groups suggests the relevance of that particular genomic region in terms of association with the disease, while their unique way to group highlights the importance of their interconnections. More specifically, the scenario with 500 trees was the most inclusive, allowing the inspection of more groups of synergic variants, and retaining more disease-associated loci. Moreover, although doing a prediction with XGBoost is far from our preliminary objectives, this scenario resulted in a better

classification of the individuals (60.52% precision) (**Suppl. Table 8**). The improvement on the prediction can be related to the ability of the method to capture more synergic loci, including rare and low-frequency variants, which are expected to have a higher effect on the disease. For all these reasons, we decided to keep this scenario for downstream analyses, thus accounting for 367 single variants, 980 pairs, 1,952 triplets, and 3,607 quadruplets.

The groups of variants obtained from applying our machine learning strategy include: 1) genetic markers which can contribute to the risk of developing the disease independently, and therefore in an additive manner, and 2) groups where there is a dependency relation between the variants that drive the effect on the disease. Although new disease-susceptibility loci can be found from the exploration of the genomic markers included in both groups, thus contributing with a better explanation of the missing heritability fraction and improving the prediction, in this first preliminary approach, we have focused on the study of the second group, which corresponds to epistatic variants. Therefore, to keep only the epistatic groups of variants, we used a logistic regression model adjusted by bmi, age, sex, and the first 7 PCs, to evaluate the effect of the interaction in the disease. As a result, we only preserved the candidate groups with a significant association with the disease driven by the interaction ($\alpha = 0.05$ with the corresponding Bonferroni adjustment for each group size) (**Methods Logistic regression epistasis**). From the complete set of groups of synergic variants obtained with XGBoost, at most 1.02% included epistatic variants. Thus, resulting in 10 pairs, 1 triplet, and 1 quadruplet of statistical interactions, containing 20, 3, and 4 unique variants, respectively (**Figure 1; Table 1**).

Table 1. Groups of epistatic variants and their effect in T2D.

Depth	Variant 1 (Effect_Ref)	Variant 2 (Effect_Ref)	Variant 3 (Effect_Ref)	Variant 4 (Effect_Ref)	Interaction Effect	p-value
2	chr5:157545791 (CAT_C)	chr4:168037835 (T_TAC)			-0.2922 (OR~1.33)	4.79x10 ⁻⁵
	chr17:76790279 (T_C)	chr6:12027402 (A_G)			0.3210 (OR~1.37)	1.18x10 ⁻⁶
	chr9:89501123 (T_G)	chr21:25168622 (C_T)			0.3862 (OR~1.47)	7.88x10 ⁻⁶
	chr11:3385759 (A_G)	chr11:123906346 (G_A)			0.6448 (OR~1.90)	7.01x10 ⁻⁶
	chr2:180203761 (T_C)	chr7:36373191 (A_AG)			0.9698 (OR~2.63)	1.51x10 ⁻⁵
	chr2:107596627 (T_G)	chr22:26957284 (C_A)			0.7055 (OR~2.02)	3.75x10 ⁻⁵
	chr4:96761220 (G_A)	chr1:206513621 (C_CCT)			1.3485 (OR~3.85)	2.49x10 ⁻⁷
	chr3:35766559 (C_T)	chr8:98754889 (A_C)			1.4534 (OR~4.27)	2.14x10 ⁻⁵
	chr4:104128410 (G_A)	chr6:111759237 (A_T)			0.2776 (OR~1.31)	4.47x10 ⁻⁵
	chr10:101881887 (G_A)	chr17:70463870 (C_T)			-0.2977 (OR~0.74)	1.16x10 ⁻⁶
3	chr20:30314136 (C_CTTT)	chr10:108835343 (G_A)	chr5:55861786 (C_T)		0.7647 (OR~2.14)	2.28x10 ⁻⁵
4	chr1:104373712 (CT_C)	chr1:147362531 (G_GT)	chr2:147085498 (G_A)	chr11:97009227 (G_T)	-2.0809 (OR~0.12)	4.13x10 ⁻⁶

Measuring the effect of epistasis in T2D

Current complex disease genomic predictors only rely on the addition of the effects of GWAS variants, thus, disregarding not only the effect of epistasis but also the possible changes in the marginal effects of variants derived from their synergies. To assess the impact of the interactions found in this study, a logistic regression was performed under two models. The first logistic regression model only evaluated the additive marginal effects of the groups of variants (pairwise, trio,

quadruplet), and the second model also included the interaction terms (full model) (**Methods Logistic regression epistasis**). The results obtained from the two models were compared to find significant differences for each of the terms included in the regression, thus involving the additive effect of the variants, and the effect of the interactions (**Figure 2; Suppl. Table 10**).

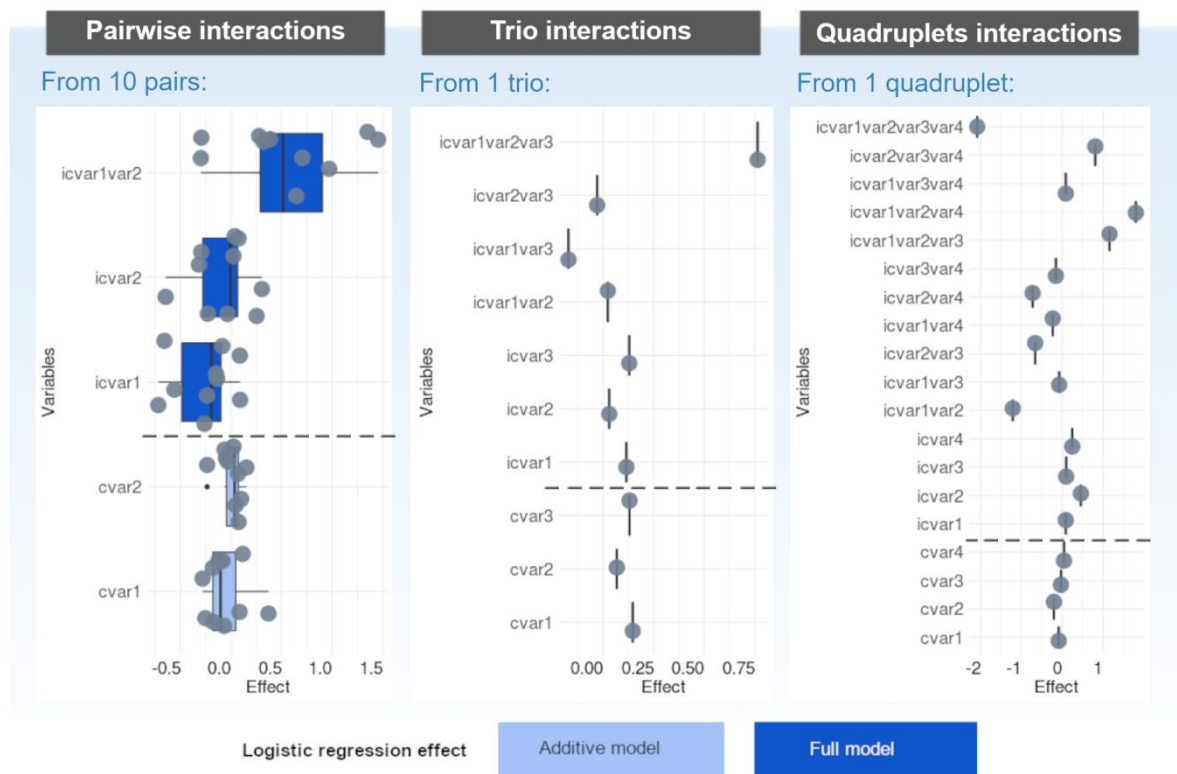


Figure 2. Evaluation of the effect of the candidate epistatic groups on the risk of developing T2D. The groups of candidate interactions (pairwise, trio, quadruplets) were evaluated for T2D associations under two logistic regression models: one considering only the marginal effects in an additive manner (light blue), and the other also including the interaction terms (dark blue). The box plots represent the distribution of the effects (x axis) for each of the terms included in the logistic regression models (y axis). The dots correspond to the effects for the different groups. The effects captured by the full model are represented above the dashed line.

First, after the inspection of pairwise interactions, few significant differences were observed between the mean and median marginal effects of the variants under the two logistic regression models. However, 50% of the pairs (5 pairs) included at least one variant which presented a change in the sign of the effect when adding the interaction term. Moreover, in one of these pairs, both variants changed their sign. In the case of the quadruplet, only one variant preserves the sign (**Suppl. Table 11**). All these cases are of particular interest, given that changes in the sign of the effect involve changes in the risk of disease development, thus for example transforming a protective variant into a risk locus, or vice versa. Additionally, the effect in the disease of any of the terms in the full model is greater in module when compared with marginal effects. Particularly, the variants under the additive model have a modest marginal effect (OR between 0.755-1.448) compared with the effects in the full model (OR between 0.488-1.579), where we observed more extreme effects. Last, there is a considerable effect on the disease derived from the full interactions, which ranges from -2.08 to 1.45 (OR between 0.13-4.27).

The epistatic variants functional impact and its association with T2D

The loci found in these epistatic associations, as well as their effect, can be used to improve T2D detection and prevention protocols. However, to find new potential drugs and to improve the treatments, it is also necessary to understand the putative molecular mechanisms underlying the

associations. Particularly, as some alterations at the genomic level can result in changes in cell function and enhance the predisposition to the development of the disease, it is crucial to find the genomic pathways underlying the associations between T2D and our epistatic groups of variants. For this reason, we analysed the genomic, transcriptomic and epigenetic context of the variants included in the epistatic groups.

First, to find any hint of the relation between the disease and the epistatic groups, at the genomic and transcriptomic levels, we explored the variants inside the groups analysing their overlap with T2D and related traits annotations, and with associated changes in the expression of pancreatic islets, a disease-related tissue. Therefore, we annotated the variants inside the epistatic groups with the summary statistics resulting from different T2D GWAS meta-analyses (Bonàs-Guarch et al., 2018; Mahajan, Taliun, et al., 2018; Scott et al., 2017; The DIAGRAM Consortium et al., 2014), glycemic traits GWAS meta-analyses (J. Chen et al., 2021), pancreatic islets expression quantitative trait loci (eQTL), and islets combined allelic specific expression (cASE) analyses (Alonso, Piron, et al., 2021) (**Methods Resources; Annotations overlap; Suppl. Figure 9**). As a result, we observed that 25% of the groups (3 groups) contain at least one variant already known to be significantly associated with T2D, glycemic traits or expression in islets (p -value $<5 \times 10^{-8}$; 5% False Discovery Rate (FDR)).

To further inspect the putative mechanisms underlying the associations with the disease, we extended our functional analysis to also cover the functional impact of variants in genes, and their overlap with human islets epigenetic marks and regulatory elements (Alonso, Piron, et al., 2021; McLaren et al., 2016; Miguel-Escalada et al., 2019; Pasquali et al., 2014). Additionally, to ensure that the functional relations found were not stochastic, we compared the genomic, transcriptomic, and epigenetic overlap obtained from our groups of candidate epistatic variants with control groups of variants randomly generated from the discovery dataset. These randomly generated groups included the same number of variants as the epistatic groups, and shared the same allelic frequency distribution (**Methods Functional annotations enrichment**).

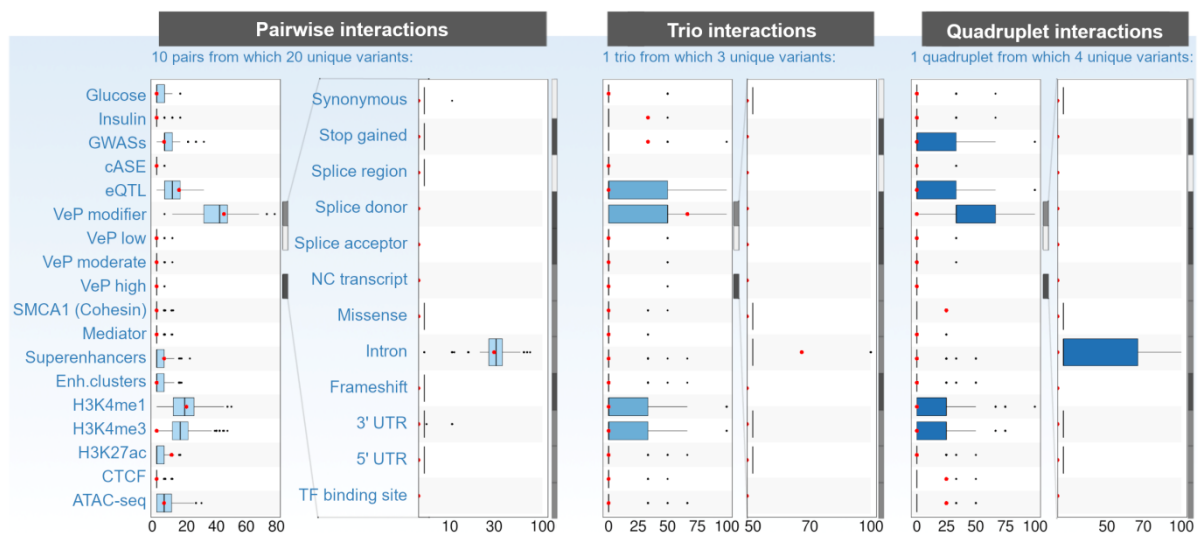


Figure 3. Percentage of significant annotations overlap. The unique list of variants present in each group of candidate epistatic variants (pairwise, trio, and quadruplet) were annotated with significant summary statistics results from T2D GWAS meta-analyses (Bonàs-Guarch et al., 2018; Mahajan, Taliun, et al., 2018; Scott et al., 2017; The DIAGRAM Consortium et al., 2014), glycemic traits GWAS meta-analyses (J. Chen et al., 2021), pancreatic islets expression analyses (Alonso, Piron, et al., 2021), islet regulatory elements (Miguel-Escalada et al., 2019; Pasquali et al., 2014), and functional impact annotations (McLaren et al., 2016). These annotations were compared with a control distribution to assess the differences. The boxplots show the distribution of the percentage (x axis) of variants with significant annotations (y axis) in the control distribution. The red dots display the proportion of epistatic candidate variants with significant annotations.

As a result from the comparison between the annotations from the epistatic groups and the annotations from the control groups, we observed that the variants included in the pairs and trio of epistatic variants are significantly more associated with T2D, related glycemic traits, islet expression variation, and more likely to fall in an islet regulatory region, than control variants (**Figure 3; Table 2**). Particularly, between 5-33.33% of these variants were significantly associated with T2D or glycemic traits. Interestingly, for the pairs, half of the GWAS signals were also an eQTL in pancreatic human islets. Moreover, the analysis of the results obtained from the annotation of the epistatic variants included in the pairs in terms of islet regulatory regions, revealed that 20% fall in H3K4me1 regions and 10% fall in H3K27ac regions. In contrast, the overlap with H3K4me3 is significantly higher in the control set (15.79%). Finally, although a more significant gene functional impact explanation was found in controls than in the candidate epistatic variants, it was mostly attributable to intronic regions. In the case of the quadruplet we observed a higher overlap with pancreatic islets cohesin, CTCF, and ATAC-seq regions (25%) when compared with controls (0%).

Table 2. Percentage of unique variants significantly annotated with T2D and glycemic traits GWAS meta-analyses, islets expression, functional impact annotations, and epigenetic marks.

Annotation	Pairwise		Trio		Quadruplet	
	discovery	control median	discovery	control median	discovery	control median
Glucose	0	0	0	0	0	0
Insulin	0	0	33.33	0*	0	0
T2D GWAS	5	5.26**	33.33	0*	0	0
eQTL	15	10.53*	0	0	0	0
cASE	0	0	0	0	0	0
VeP modifier	45	42.10*	66.66	50*	0	33.33**
VeP low	0	0	0	0	0	0
VeP moderate	0	0	0	0	0	0
VeP high	0	0	0	0	0	0
Cohesin	0	0	0	0	25	0*
Mediator	0	0	0	0	0	0
Superenhan.	0	0	0	0	0	0
Enh.cluster	0	0	0	0	0	0
H3K4me1	20	18.75*	0	0	0	25**
H3K4me3	0	15.79**	0	0	0	25**
H3K27ac	10	0*	0	0	0	0
CTCF	0	0	0	0	25	0*
ATAC-seq	5	5**	0	0	25	0*

* mean control random set annotations overlap percentage lower than discovery set results annotations overlap percentage (5% significance level)

** mean control random set annotations overlap percentage greater than discovery set results annotations overlap percentage (5% significance level)

To improve the understanding of the underlying biological mechanisms mediating the interactions, we inspected some of the most relevant epistatic groups in terms of disease explanation. In particular, the simultaneous association of a locus with disease and regulatory expression in a disease-related tissue suggests the deregulation of the gene affected as one of the putative underlying mechanisms to mediate the disease. Two of the epistatic groups (16.66%) include a T2D GWAS significant signal, from which one has at least one variant simultaneously associated with T2D and *cis*-regulatory islet expression. This is the case of the pairwise interaction of variants rs6821617 (chr4:104128410_G_A, MAF=0.395) and rs12215743 (chr6:111759237_A_T, MAF=0.1476) which contribute positively to the risk of disease development (interaction effect=0.283110 (OR~1.32), p -value=3.91x10⁻⁵) (**Figure 4.A**). The T2D GWAS variant rs6821617 (OR=0.965 p -value=3.4x10⁻⁸), although being an intergenic variant, is an islet eQTL for *BDH2* (score=-6.359, p -value=2.03x10⁻¹⁰ 1FDR), *MANBA* (score=-3.876, p -value=1.06x10⁻⁴ 5FDR), and *NFKB1* (score=-3.752, p -value=1.75x10⁻⁴ 5FDR), of which some of them have been suggested to play an important role in T2D

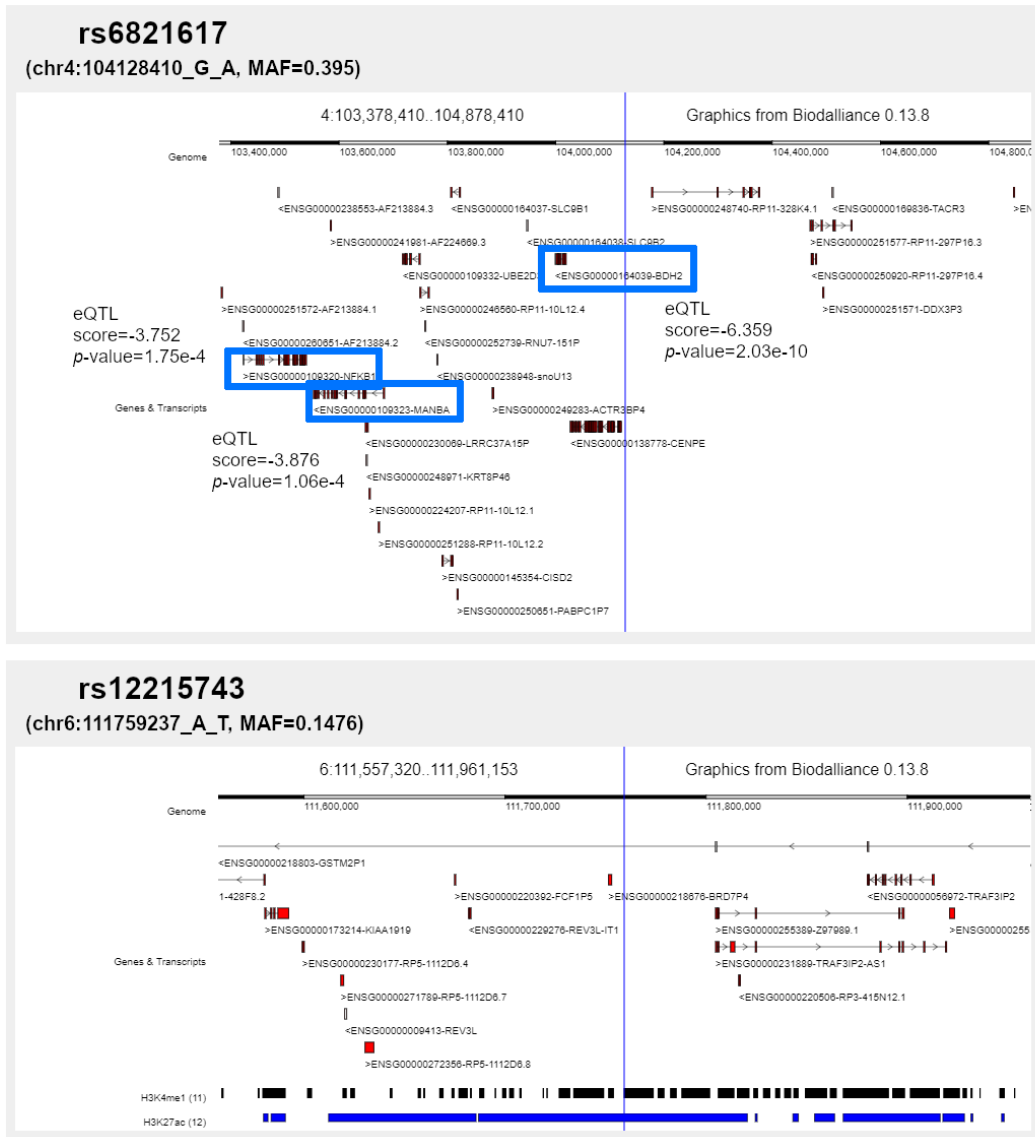
(Alonso, Piron, et al., 2021; Mahajan, Taliun, et al., 2018; McLaren et al., 2016). Particularly, the downregulation of *BDH2* has been associated with iron homeostasis thus possibly mediating its relation with diabetes (Zughaier, Stauffer, & McCarty, 2014). Moreover, *NFK1B* is lately emerging as a novel potential target for the development of therapeutic strategies to treat or prevent diabetes (Meyerovich, Ortis, & Cardozo, 2018). Additionally, rs12215743, which is a nonsense-mediated decay (NMD) mRNA variants for *REV3L*, has been associated with expression changes in *MFSD4B* (score=-4.828, p -value= 1.38×10^{-6} 1FDR), a gene involved in glucose and fructose transport in rat kidney (Alonso, Piron, et al., 2021; Horiba et al., 2003; McLaren et al., 2016).

Additionally, although variants with a modifier effect on a gene are usually non-coding or affect a non-coding gene, with no evidence of impact in the protein function, their effect can be mediated through gene expression. Thus, suggesting gene expression again as a putative mechanism to mediate the association with the disease. Particularly, 5 of the epistatic groups (41.66%) include a variant with a modifier effect on a gene and, from these, 1 group (20%) has all variants acting as a modifier for the gene. This is the case of the pair made by variants rs8073626 (chr17:76790279_C_T, MAF=0.4631) and rs17697699 (chr6:12027402_A_G, MAF=0.3219) which represent a risk for the development of the disease (interaction effect=0.321087 (OR~1.37), p -value= 1.18×10^{-6}) (**Figure 4.B**). rs8073626 is a NMD transcript variant, which falls in an intronic region of *USP36*, and has been detected as a human islet eQTL for the same gene (score=-5.625, p -value= 1.85×10^{-8} 1FDR) (Alonso, Piron, et al., 2021; McLaren et al., 2016). This gene has been suggested to participate in the pathogenesis of diabetic kidney disease, thus providing potential intervening targets (Zhu et al., 2021). Additionally, rs17697699 is an intronic *HIVEP1* variant, which falls in a human islet H3K4me1 region (McLaren et al., 2016; Pasquali et al., 2014). Interestingly, the insulin treatment induce expression of this gene, and blocking autocrine TGF-beta signalling with SB431542 substantially reduce its expression (Budi, Hoffman, Gao, Zhang, & Derynck, 2019). Moreover, *HIVEP1* has also been related to the effect of maternal diabetes and obesity in the fetal epigenome of Hispanic population (Rizzo et al., 2020).

Moreover, to gain a better insight about the decisions made to find the groups of epistatic variants, as well as their synergies, we performed an exhaustive analysis of the models generated by the machine learning method. Particularly, as XGBoost is a tree-based method, we scrutinised the decisions made in each tree to generate the different candidate groups of variants (**Methods Model outcomes interpretation**). From this analysis we found of particular interest the groups where the decision of creating the group is based on, at least, one variant having an alternate allele (heterozygous or alternate homozygous). There are 5 of these groups between our results (41.66%). This is the case of the very rare variant rs142378541 (chr4:96761220_G_A, MAF=0.007039), which couples with rs199607206 (chr1:206513621_C_CCT, MAF=0.4819) in case of being heterozygous or alternate homozygous (interaction effect=1.352184 (OR~3.86), p -value= 2.30×10^{-7}), but couples with the low-frequency variant rs76334393 (chr5:173320206_T_C, MAF=0.03595) when being homozygous reference (interaction effect=-0.274208 (OR~0.76), p -value= 3.30×10^{-1}) (**Figure 4.C**). rs142378541, which is located in an inactive open chromatin region overlapping an ATAC-seq peak from the human islet regulome, is an upstream gene variant for *PDHA2*, a gene involved in glucose metabolism for which beta-cell-specific deficiency has been related to the impairment of the glucose-stimulated insulin secretion in mouse (McLaren et al., 2016; Miguel-Escalada et al., 2019; Srinivasan et al., 2010). In case of being heterozygous, or alternate homozygous couples with rs199607206, which lays in a H3K4me1 islet region upstream of *SRGAP2*, a gene recently related to diabetic kidney disease (Levi, Myakala, & Wang, 2018; McLaren et al., 2016). In contrast, when rs142378541 is homozygous reference, it couples with rs76334393, a downstream non-coding transcript variant for *CPEB4* which falls in an islet H3K4me3, H3K27ac regions (McLaren et al., 2016; Miguel-Escalada et al., 2019). This gene, which protects against diet-induced obesity, has been associated with measures of insulin sensitivity and insulin resistance (Orozco et al., 2018; Pell et al., 2021). Particularly, this low-frequency variant is correlated ($r^2=0.1333$, p -value<0.0001) with the OGTT

fasting and plasma insulin *cis*-eQTL rs72812818 for *CPEB4* (Machiela & Chanock, 2015). Remarkably, although the interaction effect is only significant for the case of rs142378541 presenting a heterozygous or alternate homozygous genotype, it results in a high risk of developing the disease for the first couple, and a protective effect in the second couple.

A)



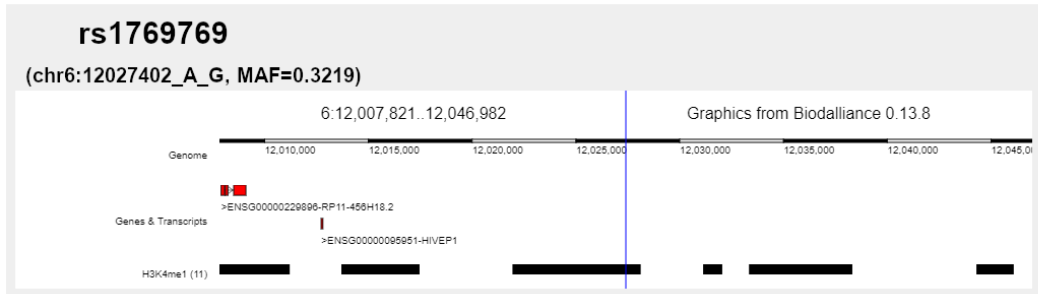
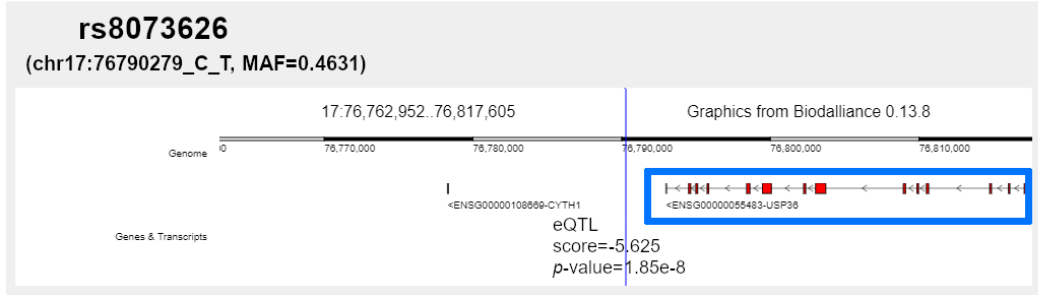
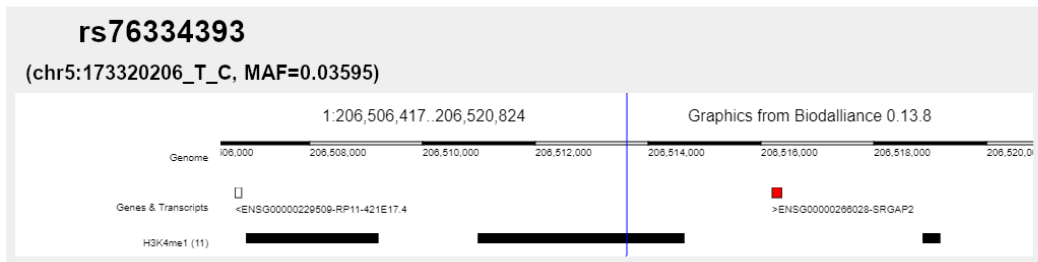
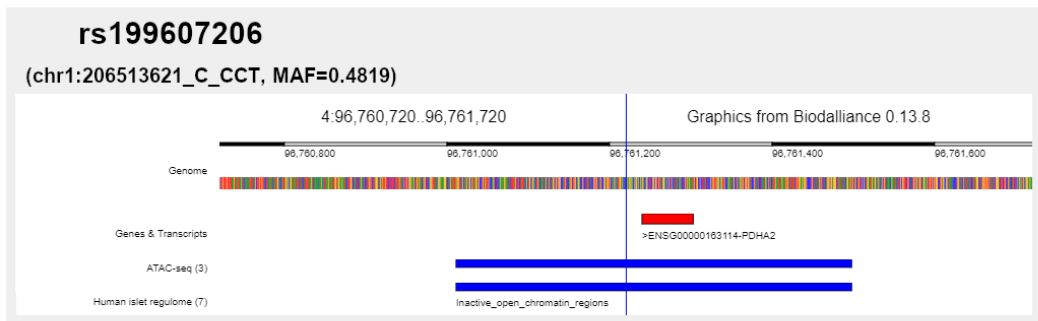
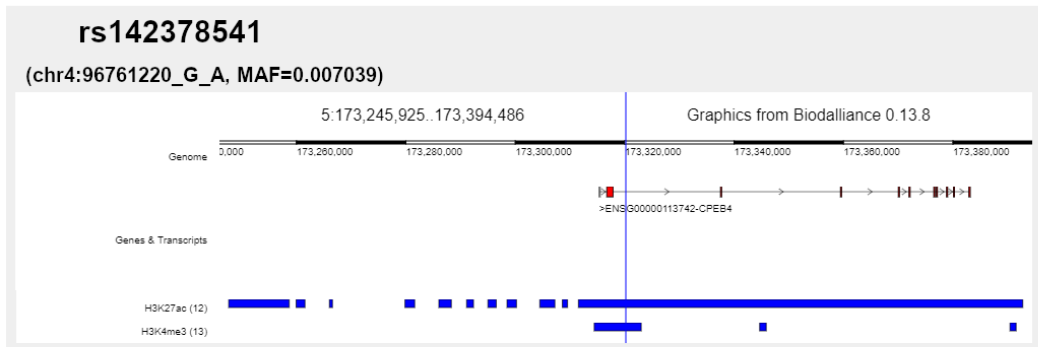
B)**C)**

Figure 4. Some examples of epistatic variants with a well-known functional interpretation in terms of disease. Each panel shows the human islet genomic context of the variant with diverse tracks for genes and transcripts and different islet regulatory regions including superenhancers, enhancer clusters, H3K4me1, H3K4me3, and H3K27ac (Miguel-Escalada et al., 2019; Pasquali et al., 2014). For each variant examined the graph only displays the tracks containing relevant functional information. The panels represent:

- A) Genomic context for the pairwise interaction between variants rs6821617 and rs12215743.
- B) Genomic context for the pairwise interaction between variants rs8073626 and rs17697699.
- C) Genomic context for the pairwise interaction between rs142378541, which couples with variant rs199607206 in case of not being reference homozygous, and with variant rs76334393 when reference homozygous.

Discussion

The analysis of a large cohort of T2D to find groups of epistatic variants affecting the disease, has led us to find 10 pairs, 1 triplet and 1 quadruplet with an interaction effect statistically associated with T2D. Interestingly, although some studies have suggested that a smaller effect on the disease is expected from interactions (Tam et al., 2019), our results showed that the effect of the interaction terms appears to be greater in module when compared with marginal effects (OR between 0.13-4.27). Furthermore, despite current polygenic predictive models of the risk to develop the disease are based on the sum of the marginal effects of GWAS variants (Alonso, Morán, et al., 2021), we have observed that these effects can vary in the presence of variant synergies. Particularly, we have found some variants changing their effect from being protective to represent a risk for the disease, thus supporting the relevance of including variant interactions in future disease association models to obtain refined measures of the effects on the disease, and to detect novel regions that in terms of interaction, both in an additive and multiplicative manner, have a higher impact on the risk of developing the disease. Therefore, our epistatic groups can represent a step forward for the genomic understanding of T2D in terms of disease predisposition, and to complement and improve the prediction scores that are currently applied to the clinics.

After studying the possible functional explanations that mediate the statistical associations between the interactions and the disease (Siemiatycki & Thomas, 1981), using genomic, transcriptomic, and epigenetic information (Cano-Gamez & Trynka, 2020; Lichou & Trynka, 2020; Manolio, 2013), we observed that 25% of the groups contain at least one variant already known to be significantly associated with T2D, glycemic traits or expression in islets ($p\text{-value} < 5 \times 10^{-8}$; 5% FDR) (Alonso, Piron, et al., 2021; Bonàs-Guarch et al., 2018; J. Chen et al., 2021; Mahajan, Taliun, et al., 2018; Scott et al., 2017; The DIAGRAM Consortium et al., 2014). These results provide support to the associations found in our study and, furthermore, although most of the epistatic groups found include some variants from which previous knowledge in T2D is unknown, the overlaps suggest the potential relation between islet regulatory variation with the disease as the underlying molecular mechanisms of the associations. Remarkably, despite this functional relation can be attributed to the selection of the discovery dataset, which was enriched in variants with higher levels of association with T2D ($-\log_{10}(p\text{-value}) > 2$), we discarded this option by comparing the results with a control set. In fact, after this comparison we conclude that islets expression regulatory variation can be suggested as some of the possible mechanisms underlying the epistatic associations. More specifically, we observed that the variants included in the groups obtained of candidate epistatic variants were significantly more associated with T2D and related glycemic traits, and more likely to fall in an islet regulatory region, than control variants (Miguel-Escalada et al., 2019; Pasquali et al., 2014), thus, suggesting that the combined effect of variation in different genomic regions and its effects on gene regulatory expression can be one of the putative mechanisms to mediate the disease. Remarkably, 16.66% of the groups include variants that have been both associated with T2D and with changes in human pancreatic islet expression. Additionally, 41.66% of the groups of candidate epistatic variants are composed of variants which present a modifier effect on genes and, although lying on non-coding regions, have an effect on islet expression or overlap an islet regulatory region.

Finally, although some of the single independent variants previously known to be associated with T2D can be thought as driving the effect of the interaction in these groups under the additive model (Hemani et al., 2021), we have proved that the interaction term was the one significantly associated with the disease through the comparison between an additive and a full logistic regression model. In particular, this is not possible for the groups that do not contain variants previously associated with the disease in a single independent manner, which correspond to the (quadruplet and 80% of the pairs). Thus, evidencing again the relevance of including the interaction of variants in association and prediction analyses to gain insight of the genomic effect of variation in the development of complex diseases.

However, despite the promising results presented, there are some limitations surrounding this study that can be improved in future epistatic analyses. First, this study focuses on the analysis of European ancestry individuals. Thus, affecting the possible extension of the results obtained to non-European populations, and limiting its projection to those loci that are shared between ancestries (Josep Maria Mercader & Florez, 2017; Spracklen et al., 2020; Vujkovic et al., 2020). Second, the high computational power required to analyse millions of variants simultaneously represents a burden for the discovery, thus limiting our study to those variants with a higher probability to be associated with T2D. Third, the number of individuals included in the study also represented a methodological limitation for the application of a ML technique. Particularly, it is recommended that the number of variants do not exceed the 10% of individuals (Chicco, 2017; Dey, 2016; Greener et al., 2021; Sarker, 2021). Although this is especially difficult in the genomics field, where the number of variants rises to millions, the number of individuals is increasing in current studies. Therefore, in the future, better results can be obtained by using this type of approaches to improve the genetic understanding of complex diseases.

Moreover, in terms of accomplishing our goals to publish the study of the effects of epistasis in T2D, all these results are still preliminary. For example, to align with other association studies we need proof of replication in a completely independent dataset to ensure that our results can be extended to the European population, and that the same methodology can be applied to analyse the effect of epistasis in other complex diseases and other populations. For this reason, we have planned to assay the replication of the epistatic variants groups obtained in the UKBiobank (UKB). Particularly, we are in the process of being granted permission to access the UKB data, and to start the analyses in this direction.

Additionally, given the increase in the availability of genomic data, current genomic studies are emerging where disease heterogeneity is being considered to find a better explanation of the disease towards personalised medicine (Ahlqvist et al., 2018, 2020; Dimas et al., 2014; H. Kim et al., 2022; Mahajan, Wessel, et al., 2018; Mansour Aly et al., 2021; McCarthy, 2017; Scott et al., 2017; Udler et al., 2018). Although in this first study we have focused our analysis in finding epistatic groups of variants in T2D, after replication and publication of the results presented in this thesis, we have also planned to expand the study analysing T2D subgroups to reveal the groups of epistatic variants shared between these subgroups of diabetic individuals, and the exclusive interactions in each subgroup of patients. Additionally, it will be also interesting to apply the same methodology to other complex diseases, and to improve our analytical frameworks to analyse the epistatic problem at a genome wide level. Therefore, including in the study all the variants that we have discarded for computational limits reasons. Moreover, all the approaches presented in this study were implemented under the additive model. However, although most variants follow this genetic pattern, the remaining variants under non-additive models can escape from the discovery (Guindo-Martínez et al., 2021). For this reason, it will be of particular interest to extend our models to cover all the possible inheritance patterns, and therefore improve the explanation of T2D heritability. Finally, the analysis of chromosome X has been proved of particular relevance in terms of disease explanation (Bonàs-Guarch et al., 2018). Particularly, although we have included this chromosome in the study, further efforts need to be applied to improve the analysis based on its particularities, therefore enhancing the complete inspection of its epistatic effects on the disease.

Methods

Discovery dataset

T2D case-control dataset

The 70KforT2D is a T2D case-control dataset which includes data from 12,926 diabetic and 57,191 non-diabetic individuals of European ancestry (Bonàs-Guarch et al., 2018). The individuals included in this dataset belong to 5 studies: Resource for Genetic Epidemiology Research on Aging (GERA), Finland-United States Investigation of NIDDM Genetics (FUSION), Wellcome Trust Case Control Consortium (WTCCC), Gene Environment Association Studies initiative (GENEVA), Northwestern University NUgene project (NUgene) (Burton et al., 2007; Colditz & Hankinson, 2005; Ghosh et al., 2000; Gottesman et al., 2013; Kvale et al., 2015). The genetic information is publicly available through the dbGaP platform for FUSION (phs000867.v1.p1), GENEVA (phs000091.v2.p1), NUgene (phs000237.v1.p1), GERA (phs000788.v2.p3), and the Sanger platform for WTCCC. The available metadata for each individual corresponds to measures of body-mass index, sex, age and diabetic type. Nonetheless, there is no available information from NUgene individuals' age, neither for WTCCC individuals' age and bmi. The genotype of the individuals included in each of the 5 cohorts that comprehend the 70KforT2D dataset, passed a quality control, and were imputed by Sílvia Bonàs-Guarch, to reach genome-wide level, combining the power of two reference panels (Bonàs-Guarch et al., 2018).

Dataset preparation

To ensure the good quality of the genotype information included for downstream analysis, only the variants with an imputation INFO score >0.7 were kept from the panel with the best imputation quality, thus consisting on 15,131,345 variants. To avoid many factors that affect the performance of the machine learning method such as data type, the amount of available data, data imbalance, the presence of outliers, and data missingness, many preliminary analyses were performed (Chicco, 2017; Dey, 2016; Greener et al., 2021; Sarker, 2021). As a result, first, to prevent trend decisions based on case-control imbalance, only paired metadata individuals were included from each cohort, where 547 case-control pairs were included from FUSION, 1,883 from GENEVA, 6,743 from GERA, 334 from NUgene and 1,894 from WTCCC (**Methods Data imbalance**). Then, the datasets were merged with qctool (Band & Marchini, 2018), therefore creating a dataset which consists of 15,131,345 variants and 22,802 individuals (11,401 diabetic and 11,401 non-diabetic). The genotype probabilities were converted into hardcalls (hardcall-threshold 0.9) with PLINK (Chang et al., 2015). Finally, to ensure the good performance of the model, avoid computational problems, and prevent overfitting, the number of variants included in the analysis was reduced by keeping only those variants with a $-\log_{10}(p\text{-value}) > 2$ from the 70KforT2D GWAS summary statistics (**Methods Randomness assessment, Maximisation of variables explanation, Variables redundancy, Missingness, Data availability**). Thus, the discovery dataset included only 105,896 genomic variants.

Machine learning approaches

Method selection

Different supervised machine learning classifiers from the scikit-learn library in python (Pedregosa et al., 2011) were applied to evaluate their performance in a reference subset of the data (1,667 GWAS significant ($-\log_{10}(p\text{-value}) > 7$) features and 22,802 individuals) (Dey, 2016; Greener et al., 2021). The methods evaluated were Nearest Neighbours, Linear SVM, RBF SVM, Gaussian Process, Decision Trees, Random Forest, Neural Networks, AdaBoost, Naive Bayes, QDA and XGBoost. The results obtained by each method were evaluated in terms of computing time, precision, and the data type accepted by the method (**Suppl. Table 1**). The unique learners prepared to work

with missing data were Gaussian Process and XGBoost, however as missingness over a 10% is present in less than a 26% of the genomic variants included in our dataset, the rest of the methods were tested by assigning a new class to the missing genotypes. As a result, the best performance was obtained in terms of computation time and precision by XGBoost (T. Chen & Guestrin, 2016).

Algorithm preparation

A basic train-test algorithm was prepared first splitting the discovery dataset in two independent datasets: a train set and a test set (Chicco, 2017; Greener et al., 2021). Then, the train set was used by the XGBoost algorithm (T. Chen & Guestrin, 2016) to learn and the test set to evaluate the results. To prevent the overfitting of the model and to obtain the best performance, a grid search hyperparameter adjustment was applied under a 5-fold cross-validation algorithm. The hyperparameters adjusted were split (0.2, 0.3), learning rate (0.01, 0.04, 0.07, 0.1), number of trees (50, 100, 250, 500), and depth (1, 2, 3, 4).

Data imbalance

The presence of data imbalance can affect the performance of the model resulting in trend decisions (Chicco, 2017; Dey, 2016; Greener et al., 2021; Sarker, 2021). To ensure the best performance of the model in the presence or absence of data imbalance, a subset of 1,667 GWAS significant variants ($-\log_{10}(p\text{-value}) > 7$ (Bonàs-Guarch et al., 2018)) was created. Two datasets of individuals were prepared, one containing all the individuals from the 70KforT2D (12,926 diabetic and 57,191 non-diabetic individuals) (Bonàs-Guarch et al., 2018), and a paired-metadata dataset, where each diabetic individual was paired with a non-diabetic individual sharing bmi and age range, and same sex. As a result, the paired-metadata dataset included the genotype information from 11,401 diabetic and 11,401 non-diabetic individuals. Each individual included in these datasets was provided with the corresponding genotype information. Each dataset was used as an input to train and test the XGBoost model (T. Chen & Guestrin, 2016) under a 5-fold cross-validation with hyperparameters adjustment (split (0.2, 0.3), learning rate (0.01, 0.04, 0.07, 0.1), number of trees (50, 100, 250, 500), and depth (1, 2, 3, 4)). The best hyperparameters were then used to train and test the model, and therefore, to evaluate the outcomes in terms of precision, accuracy, Recall, F1-score, and Matthews Correlation Coefficient (MCC). The remaining cases and controls from the 70KforT2D (Bonàs-Guarch et al., 2018) were also used as an independent test set for the paired-metadata dataset (**Suppl. Table 2, Suppl. Figure 1**). The results obtained evidenced a best performance of the method in the absence of data imbalance. Therefore, the subsequent analyses were performed with the paired-metadata dataset.

Randomness assessment

To assess the effects of randomness at the genotype and phenotype level in the results obtained by the method, the performance of the 1,667 GWAS significant ($-\log_{10}(p\text{-value}) > 7$ (Bonàs-Guarch et al., 2018)) dataset with 22,802 individuals (reference dataset), was compared to two random control datasets with the same number of features and individuals. The first control dataset has the individual's genotype randomly assigned, and the second dataset has the individual's phenotype randomly assigned, maintaining the proportion of cases and controls. Each of these datasets was generated 1,000 times. Each dataset was used as an input to train and validate the XGBoost model (T. Chen & Guestrin, 2016) in the 5-fold cross-validation with hyperparameters adjustment (split (0.2, 0.3), learning rate (0.01, 0.04, 0.07, 0.1), number of trees (50, 100, 250, 500), and depth (1, 2, 3, 4)). The results were evaluated in terms of precision for each depth and tree pairs. After assessing the normality of the distribution of the precision, using a Shapiro-Wilks test ($3 < N < 5000$) or Anderson-Darling test ($N \geq 5000$), we performed a t-test (normal distribution), sign test (not normal, non-symmetric distribution), or Wilcoxon signed-rank test (not normal, symmetric distribution), to check if the mean precision was significantly better than random (5% significance).

level) (**Suppl. Table 3, Suppl. Figure 2**). As a result, we discarded randomness as affecting the outcomes obtained by the method in the reference dataset.

Maximisation of variables explanation

The amount of available data can affect the performance of the model, particularly, the ideal machine learning situation is having at least ten times the number of features in the number of observations (Chicco, 2017; Dey, 2016; Greener et al., 2021; Sarker, 2021). However, although 1,667 GWAS significant variants ($-\log_{10}(p\text{-value}) > 7$ (Bonàs-Guarch et al., 2018)) evaluated in 22,802 individuals (reference dataset) represents 7.3% of the number of observations, the effect of including different groups of variants without limiting the discovery dataset to only GWAS significant variants is not clear. To ensure that the inclusion of more variants can lead to a better classification of diabetic patients, a comparison was performed between the reference dataset with a random control dataset. This control dataset was created including 1,667 variants for each individual from a subset of the 70KforT2D (105,896 variants, $-\log_{10}(p\text{-value}) > 2$ (Bonàs-Guarch et al., 2018)). Each random control dataset was generated 1,000 times. Each dataset was used as an input to train and validate the XGBoost model (T. Chen & Guestrin, 2016) in the 5-fold cross-validation with hyperparameters adjustment (split (0.2, 0.3), learning rate (0.01, 0.04, 0.07, 0.1), number of trees (50, 100, 250, 500), and depth (1, 2, 3, 4)). The results were evaluated in terms of precision for each depth and tree pairs. After assessing the normality of the distribution of the precision, using a Shapiro-Wilks test ($3 < N < 5000$) or Anderson-Darling test ($N \geq 5000$), we performed a t-test (normal distribution), or Wilcoxon Mann-Whitney test (not normal), to check if the mean precision was significantly greater in the random dataset (5% significance level). Moreover, the precision results obtained for the best hyperparameters were compared between the datasets (**Suppl. Table 4, Suppl. Figure 3**). As the number of features was a limitation to improve the discovery, and better precision results were obtained for the random datasets, the decision was to include in the discovery dataset as many variants as possible. However, computational and methodological limitations reduced the discovery dataset to a subset of 105,896 variants ($-\log_{10}(p\text{-value}) > 2$).

Variables redundancy

To ensure that variables redundancy was not affecting the model (Chicco, 2017; Dey, 2016; Greener et al., 2021; Sarker, 2021), we assessed the comparison between the results obtained by the model between the complete discovery dataset (105,896 variants, $-\log_{10}(p\text{-value}) > 2$ (Bonàs-Guarch et al., 2018), 22,802 individuals), and the same dataset after doing a linkage disequilibrium clumping with PLINK (Chang et al., 2015) ($r^2 = 0.2$, 250kb, $p\text{-value} = 0.5$, 70KforT2D summary statistics (Bonàs-Guarch et al., 2018)). The results obtained by the 5-fold cross-validation algorithm with hyperparameters adjustment (split (0.2, 0.3), learning rate (0.01, 0.04, 0.07, 0.1), number of trees (50, 100, 250, 500), and depth (1, 2, 3, 4)) were compared in terms of precision for each depth and tree pairs. After assessing the normality of the distribution of the precision, using a Shapiro-Wilks test, we performed a t-test (normal distribution), or Wilcoxon Mann-Whitney test (not normal), to check if the mean precision was significantly greater in the discovery dataset than in the clumped dataset (5% significance level). Moreover, the precision results obtained for the best hyperparameters were compared between the datasets (**Suppl. Table 5, Suppl. Figure 4**). No significant differences were observed between the datasets in terms of the mean, median, or best precision obtained. As a result, we prioritised the use of the complete discovery dataset to include the maximum number of signals, and to prevent hidden causal variants driving the effect of the interaction (Hemani et al., 2021).

Missingness

Although XGBoost (T. Chen & Guestrin, 2016) is a method prepared to work with missing values, it is known that the presence of missing values in the dataset can affect the performance of the model (Chicco, 2017; Dey, 2016; Greener et al., 2021; Sarker, 2021). Particularly, the 26% of the variants included in the discovery dataset (105,896 variants, $-\log_{10}(p\text{-value}) > 2$ (Bonàs-Guarch et al.,

2018), 22,802 individuals) present over a 10% of missing values. Consequently, to ensure the good performance of the model with this proportion of missingness, a comparison was made between the complete discovery dataset and the same dataset reducing the number of variants to be analysed to those with less than a 10% of missingness. The results obtained were compared under the 5-fold cross-validation algorithm with hyperparameters adjustment (split (0.2, 0.3), learning rate (0.01, 0.04, 0.07, 0.1), number of trees (50, 100, 250, 500), and depth (1, 2, 3, 4)) in terms of performance. After assessing the normality of the distribution of the precision, using a Shapiro-Wilks test, we performed a t-test (normal distribution), or Wilcoxon Mann-Whitney test (not normal), to check if the mean precision was significantly greater in the discovery dataset than in the dataset with less missingness (5% significance level). Moreover, the precision results obtained for the best hyperparameters were compared between the datasets (**Suppl. Table 6, Suppl. Figure 5**). No significant differences were observed between the datasets in terms of the mean, median, or best precision obtained. As a result, we prioritised the use of the complete discovery dataset to include the maximum number of signals, and to prevent hidden causal variants driving the effect of the interaction (Hemani et al., 2021).

Data availability

The amount of available data can affect the performance of the model, particularly, the ideal machine learning situation is having at least ten times the number of features in the number of observations (Chicco, 2017; Dey, 2016; Greener et al., 2021; Sarker, 2021). Given that the discovery dataset (105,896 variants, $-\log_{10}(p\text{-value}) > 2$ (Bonàs-Guarch et al., 2018), 22,802 individuals) was not accomplishing this rule, the results between the discovery dataset and applying PCA multidimensionality reduction with scikit-learn library in python (Pedregosa et al., 2011) were compared. For this reason, two datasets were created keeping the PCs explaining the 95% of variability (PCA), and just keeping the first 2,200 PCs (10% of the number of observations; PCA10). The performance of the algorithm was evaluated under a 5-fold cross-validation algorithm with hyperparameters adjustment (split (0.2, 0.3), learning rate (0.01, 0.04, 0.07, 0.1), number of trees (50, 100, 250, 500), and depth (1, 2, 3, 4)). The results were compared in terms of precision for each depth and tree pairs. After assessing the normality of the distribution of the precision, using a Shapiro-Wilks test, we performed a t-test (normal distribution), or Wilcoxon Mann-Whitney test (not normal), to check if the mean precision was significantly greater in the discovery dataset than in the PCs datasets (5% significance level). Moreover, the precision results obtained for the best hyperparameters were compared between the datasets (**Suppl. Table 7, Suppl. Figure 6**). Although a significantly better precision was obtained with both PCs datasets compared with the discovery dataset, the small benefit in terms of precision (<2%) produced by the use of these datasets, in contrast with the loss of biological and genetic explanation caused by the PCA transformation, the discovery dataset was the one selected to continue the analysis.

Hyperparameters adjustment

After ensuring a good performance of the complete method with the discovery dataset (105,896 variants, $-\log_{10}(p\text{-value}) > 2$ (Bonàs-Guarch et al., 2018), 22,802 individuals), the hyperparameters were adjusted to prevent overfitting and to obtain the best results from the model (Greener et al., 2021). The parameters under evaluation were the *split* of the dataset in training and test (0.2, 0.3), the *learning rate* needed to create a new tree (0.01, 0.04, 0.07, 0.1), the *number of trees* that the method will construct (50, 100, 250, 500), and the *depth* of each tree (1, 2, 3, 4). Particularly, the number of trees and the depth of the tree were detected as the main causes leading to overfitting during the 5-fold cross-validation. For this reason, the maximum depth of a tree was limited to 4. To ensure that the overfitting observed during the 5-fold cross-validation was not extended to the prediction, the performance of the algorithm in terms of precision was compared between the validation set and the test set. After assessing the normality of the distribution of the precision, using a Shapiro-Wilks test, we performed a t-test (normal distribution), or Wilcoxon Mann-Whitney test (not normal), to check if the mean precision was significantly different in the validation dataset than in the test dataset (5% significance level) (**Suppl. Table 8, Suppl. Figure 7**). No

significant differences were detected. Therefore, the rest of the hyperparameters (split and learning rate) were defined by taking those that lead to the best median precision in the validation set under the 5-fold cross-validation. All the scenarios were kept for downstream analysis based on the number of trees expected as an outcome. The best hyperparameters for each scenario correspond to:

50 trees

- Depth 1: split = 0.2, learning rate = 0.1
- Depth 2: split = 0.2, learning rate = 0.1
- Depth 3: split = 0.2, learning rate = 0.1
- Depth 4: split = 0.2, learning rate = 0.1

100 trees

- Depth 1: split = 0.2, learning rate = 0.1
- Depth 2: split = 0.2, learning rate = 0.1
- Depth 3: split = 0.2, learning rate = 0.1
- Depth 4: split = 0.2, learning rate = 0.07

250 trees

- Depth 1: split = 0.2, learning rate = 0.1
- Depth 2: split = 0.2, learning rate = 0.1
- Depth 3: split = 0.3, learning rate = 0.07
- Depth 4: split = 0.2, learning rate = 0.07

500 trees

- Depth 1: split = 0.3, learning rate = 0.1
- Depth 2: split = 0.3, learning rate = 0.1
- Depth 3: split = 0.3, learning rate = 0.1
- Depth 4: split = 0.2, learning rate = 0.07

Genomic inspection of the results

Model outcomes interpretation

The XGBoost method (T. Chen & Guestrin, 2016) is based on extreme gradient boosting trees, therefore the resulting model obtained after the training is composed of a list of the most relevant variants for the method to do the classification with their corresponding scores, and the complete set of final decision trees including the decisions. After the test step, the method provides a list with the predictions and the real observed values. The list of variants can be scored using two different measures, the weight, which is related to the number of times that the variant has been used to make a decision, or the gain, which corresponds to the accuracy value after adding the variant to the final model. The trees obtained represent at least one group of candidate interacting variants, where the leaves are the variants in each group, and the branches are the decisions made by the method. Particularly, the analysis of the complete set of decisions made during the training correspond to find differences between the variants genotype, thus responding to questions such as the variant being reference homozygous or alternate homozygous for a particular individual. Therefore, a list of all the unique groups of candidate epistatic variants was created, based on the decisions made by each of the trees, to facilitate a better genomic comprehension of the epistatic groups obtained as an outcome of the model, and to simplify their downstream functional assessment.

Candidate epistatic groups base genomics

To have a preliminary overview of the variants included in the groups of candidate epistatic variants for each scenario (50, 100, 250, and 500 trees), we first classified them by their minor allele frequency calculated with PLINK (Chang et al., 2015) (**Suppl. Figure 8**). Second, to understand the relation between the candidate variants by tree and by depth, we calculated the linkage disequilibrium (LD) between all the pairs of variants resulting from our analyses in terms of r^2 with PLINK (Chang et al., 2015) (--r2 --ld-window-kb 500). Then we evaluated the percentage of variants in strong LD ($r^2 > 0.8$) and in weak LD ($r^2 > 0.2$) that were included in each group of candidate epistatic variants. This last analysis was performed considering all the variants in the first group being in LD with the variants

in the second group, and also accepting that only some variants from the first group were in LD with the variants in the second group (**Suppl. Table 9**).

Genomic, transcriptomic, and epigenetic functional assessment

Resources

The TIGER browser (<http://tiger.bsc.es>) and its database (Alonso, Piron, et al., 2021) was used to inspect and extensively annotate the different loci included in each group of epistatic variants. To prepare the annotations, the genomic information from T2D GWAS meta-analysis summary statistics from the 70KforT2D (Bonàs-Guarch et al., 2018), DIAGRAM DIAMANTE (Mahajan, Taliun, et al., 2018), DIAGRAM trans-ethnic (The DIAGRAM Consortium et al., 2014), DIAGRAM 1000G (Scott et al., 2017), transcriptome expression results from human pancreatic islets expression quantitative trait loci (eQTL) and combined allelic specific expression (cASE) (Alonso, Piron, et al., 2021), islets epigenetic marks (Miguel-Escalada et al., 2019; Pasquali et al., 2014), and variant effect predictor annotations (McLaren et al., 2016), were downloaded from the TIGER resource. Moreover, MAGIC trans-ancestry and single-ancestry meta-analyses on glycemic traits (fasting glucose, 2h glucose levels, and fasting insulin levels) summary statistics were gathered (J. Chen et al., 2021).

Annotations overlap

To assess the overlap between the candidate epistatic variants obtained in each group (pairs, trios, and quadruplets), each variant was annotated using the summary statistics from different T2D GWAS meta-analysis (Bonàs-Guarch et al., 2018; Mahajan, Taliun, et al., 2018; Scott et al., 2017; The DIAGRAM Consortium et al., 2014), European ancestry glycemic traits meta-analysis (J. Chen et al., 2021), and human pancreatic islets eQTL and cASE (Alonso, Piron, et al., 2021). Only the significant annotations were kept, therefore only allowing the inclusion of the annotations with a p -value $< 5 \times 10^{-8}$ for T2D and glycemic traits GWAS meta-analyses, p -value $< 3.453 \times 10^{-5}$ for eQTL (5% FDR), and 5% FDR for cASE. The results obtained were used to calculate the proportion of epistatic variants which overlap with already known significant variants associated with T2D or glycemic traits. The same calculation was applied with eQTL and cASE to see the proportion of variants included in the epistatic groups which have an already known functional interpretation in terms of pancreatic islet expression (**Suppl. Figure 9**).

Functional annotations enrichment

To analyse the functional annotations enrichment of the list of epistatic variants obtained in each group (pairs, trios, and quadruplets) the summary statistics and available annotations from T2D GWAS meta-analysis (Bonàs-Guarch et al., 2018; Mahajan, Taliun, et al., 2018; Scott et al., 2017; The DIAGRAM Consortium et al., 2014), European ancestry glycemic traits meta-analysis (J. Chen et al., 2021), human pancreatic islets expression (Alonso, Piron, et al., 2021), islet regulatory elements (Miguel-Escalada et al., 2019; Pasquali et al., 2014), and gene functional impact (McLaren et al., 2016), were downloaded. For the pancreatic islets expression, only the significant annotations (5% FDR) of eQTL and cASE were kept. In the same manner, only the significant annotations for T2D and glycemic traits GWAS meta-analyses (p -value $< 5 \times 10^{-8}$) were evaluated in the epistatic groups. For each group of epistatic variants a null distribution of control variants from the discovery dataset was generated. The control group included the same number of variants as the epistatic group, with the same MAF distribution. Therefore, first, the MAF distribution by decile was calculated on the discovery set. Second, a recount of epistatic variants included in each decile was performed to then randomly select 1,000 times the same amount of variants from the corresponding MAF decile in the discovery dataset. All the variants in the sets were annotated using GWAS and islet significant annotations. The proportion of annotated variants in the groups was finally compared. After assessing the normality of the distribution of the percentage of variants annotated, using a Shapiro-Wilks test, we performed a t-test (normal distribution), or Wilcoxon Mann-Whitney test (not normal), to check if the mean precision

was significantly different in the results dataset than in the control dataset (5% significance level) (**Figure 3; Table 2**).

Statistical assessment

Logistic regression epistasis

To validate the results obtained from the machine learning algorithm, a logistic regression was performed in the discovery dataset (22,802 individuals, candidate groups of epistatic variants). The regression was applied for two statistical models, where the first (additive model) only considered the additive marginal effect of the variants

$$\ln\left(\frac{P(\text{T2D})}{P(\text{control})}\right) = \sum_{i=1}^N \beta_i \text{variant}_i, N \in \{2,3,4\},$$

and the second (interaction model) combined the additive marginal effect of the variants with all their possible interactions

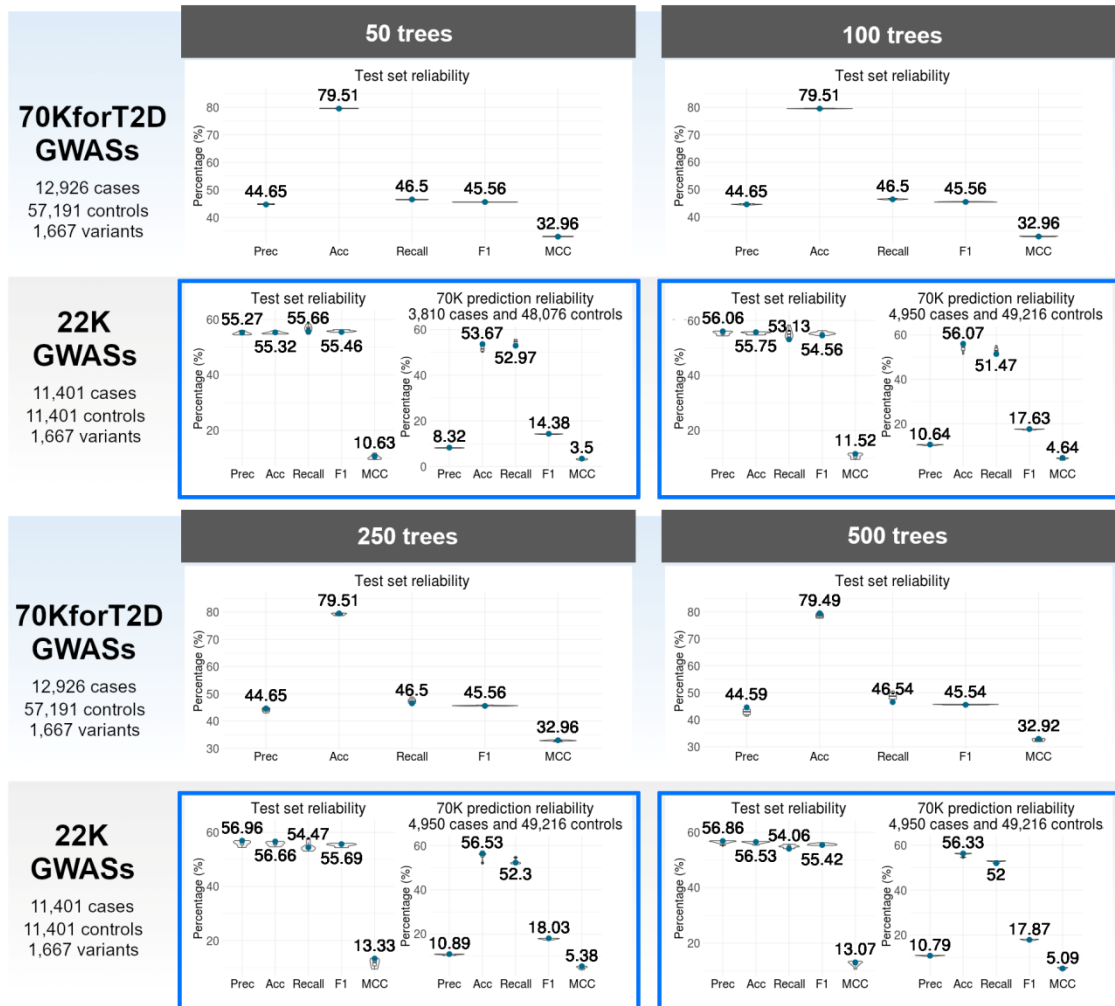
$$\ln\left(\frac{P(\text{T2D})}{P(\text{control})}\right) = \sum_{i=1}^N \beta_i \text{variant}_i + \sum_{\substack{i,j=1 \\ i \neq j}}^N \beta_{ij} \text{variant}_i \text{variant}_j + \sum_{\substack{i,j,k=1 \\ i \neq j \neq k}}^N \beta_{ijk} \text{variant}_i \text{variant}_j \text{variant}_k + \sum_{\substack{i,j,k,l=1 \\ i \neq j \neq k \neq l}}^N \beta_{ijkl} \text{variant}_i \text{variant}_j \text{variant}_k \text{variant}_l, N \in \{2,3,4\}.$$

Each of the models was adjusted to capture the effect of bmi, age, sex, and the first 7 PCs. The PCs were calculated using PLINK (Purcell et al., 2007) multidimensional-scaling method (MDS) to account for the population structure. The results obtained from the machine learning algorithm with a non-significant Bonferroni p -value association ($\alpha = 0.05$) in the interaction term, adjusted for multiple testing correction for each group size, were filtered. Moreover, the effect of epistasis in the candidate groups of epistatic variants was measured and compared between the two models. After assessing the normality of the distribution of the effect, using a Shapiro-Wilks test, we performed a t-test (normal distribution), or Wilcoxon Mann-Whitney test (not normal), to check if the mean effect was significantly different in the additive model than in the interaction model for the marginal effects (5% significance level) (**Figure 2**). Some significant differences were detected in the pairwise interactions. Moreover, to check if there were significant differences in the distribution of the marginal effects between the two models, a Kolmogorov-Smirnov test was performed (**Suppl. Table 10**). No significant differences were detected. Additionally, we calculated the proportion of changes observed in the sign of the variants marginal effects between the two models.

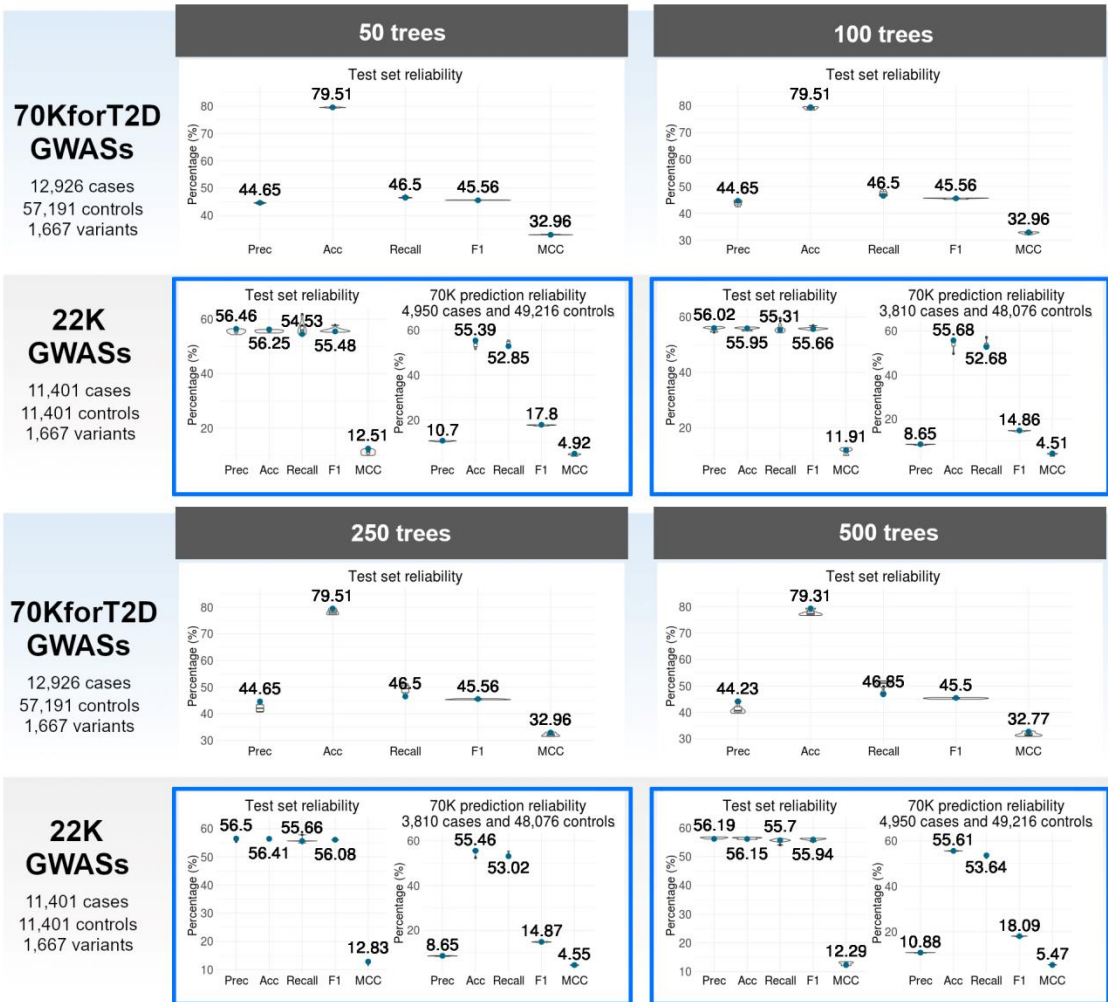
Supplemental Materials

Supplemental Figure 1. Evaluation of the performance of XGBoost under case-control imbalance.

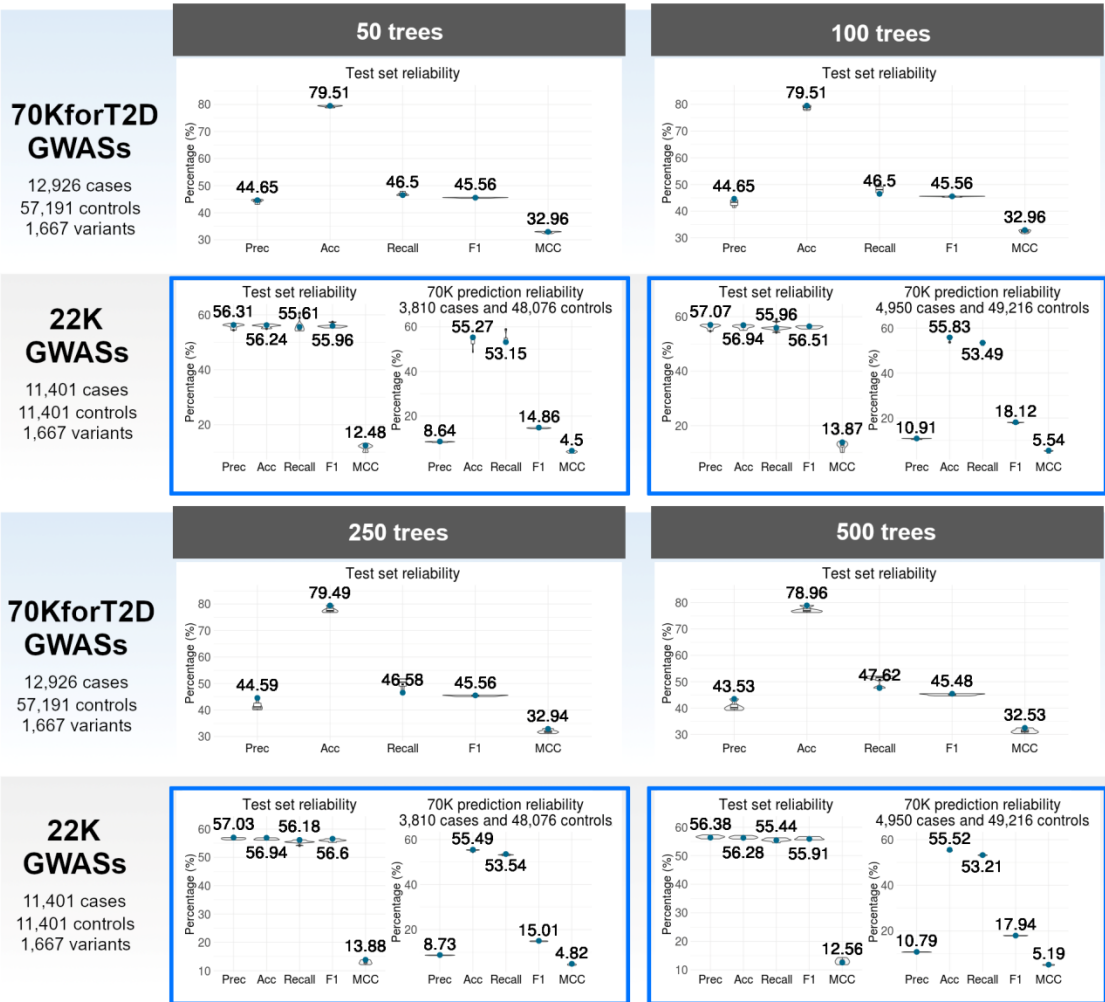
A)



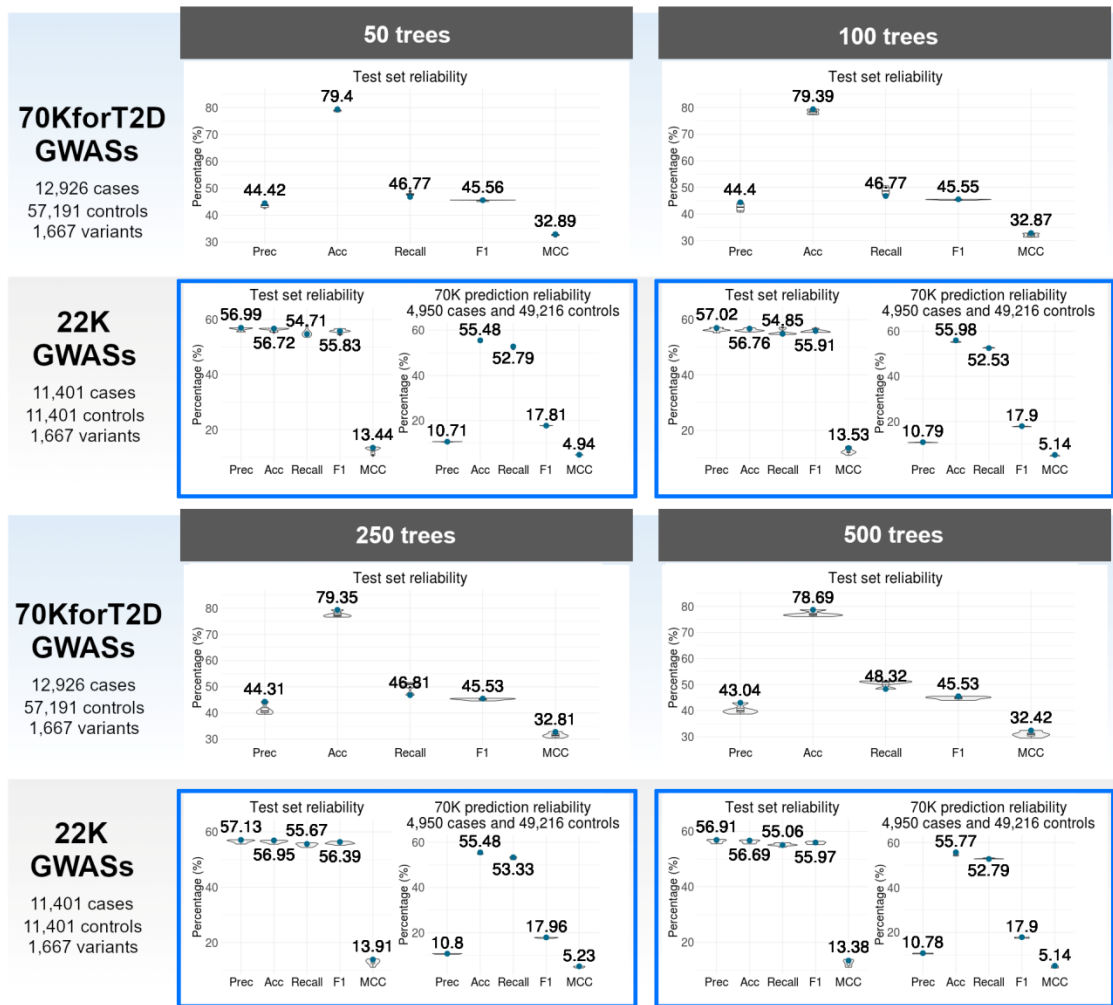
B)



c)



D)



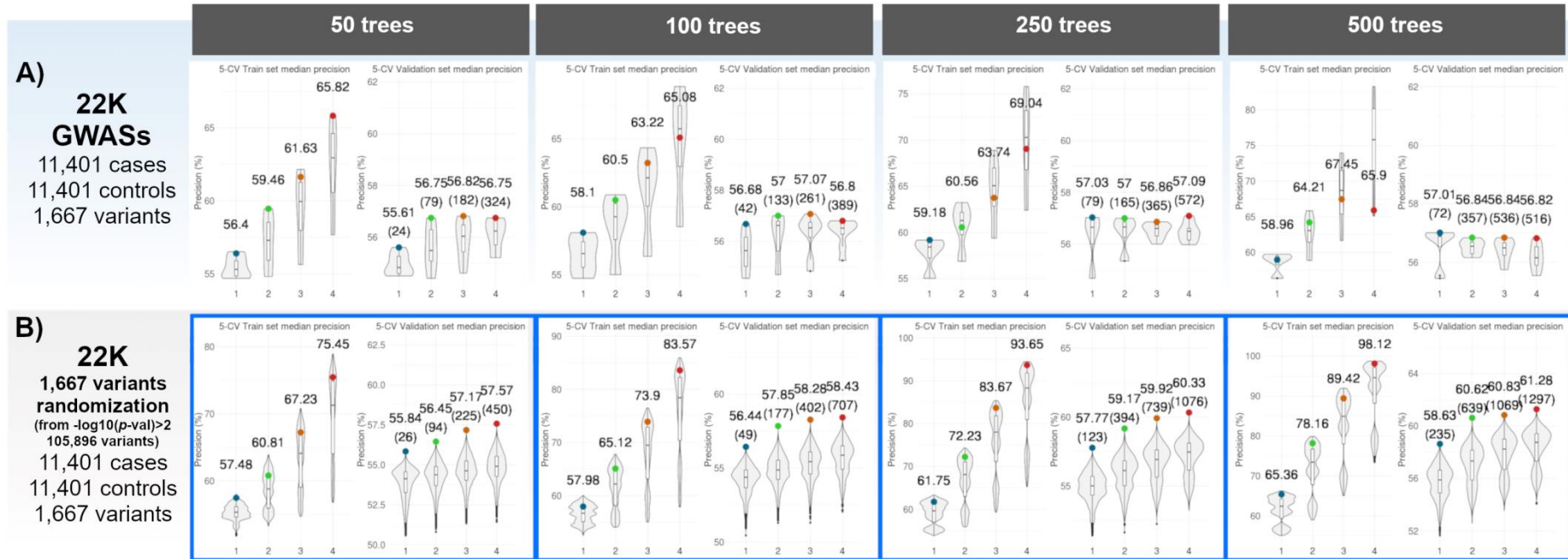
The performance of the method was evaluated with case-control imbalanced data 70KforT2D (12,926 diabetic and 57,191 non-diabetic individuals, 1,667 variants, $-\log_{10}(p\text{-value}) > 7$ (Bonàs-Guarch et al., 2018)), and balanced data 22K (11,401 diabetic and 11,401 non-diabetic paired metadata individuals, 1,667 variants, $-\log_{10}(p\text{-value}) > 7$ (Bonàs-Guarch et al., 2018)), in terms of precision (Prec), accuracy (Acc), recall (Recall), F1-score (F1), and Matthews Correlation Coefficient (MCC). After a 5-fold cross-validation with hyperparameters adjustment (split (0.2, 0.3), learning rate (0.01, 0.04, 0.07, 0.1), number of trees (50, 100, 250, 500), and depth (1, 2, 3, 4)), the best hyperparameters were used to test the results for each dataset. For each depth A) 1, B) 2, C) 3, D) 4, and number of trees (columns), each row displays the percentage obtained (y axis) for each reliability measure (x axis) evaluated in the imbalanced (top) and balanced (bottom) datasets. Each violin plot represents the distribution of the reliability measures obtained for each hyperparameter combination. The coloured dots correspond to the results obtained with the best hyperparameters. The squared data encapsulates the best results. For the balanced dataset the results on the test set (left), and a prediction on the remaining 70KforT2D (right) are provided.

Supplemental Figure 2. Evaluation of the performance of XGBoost in terms of randomness.



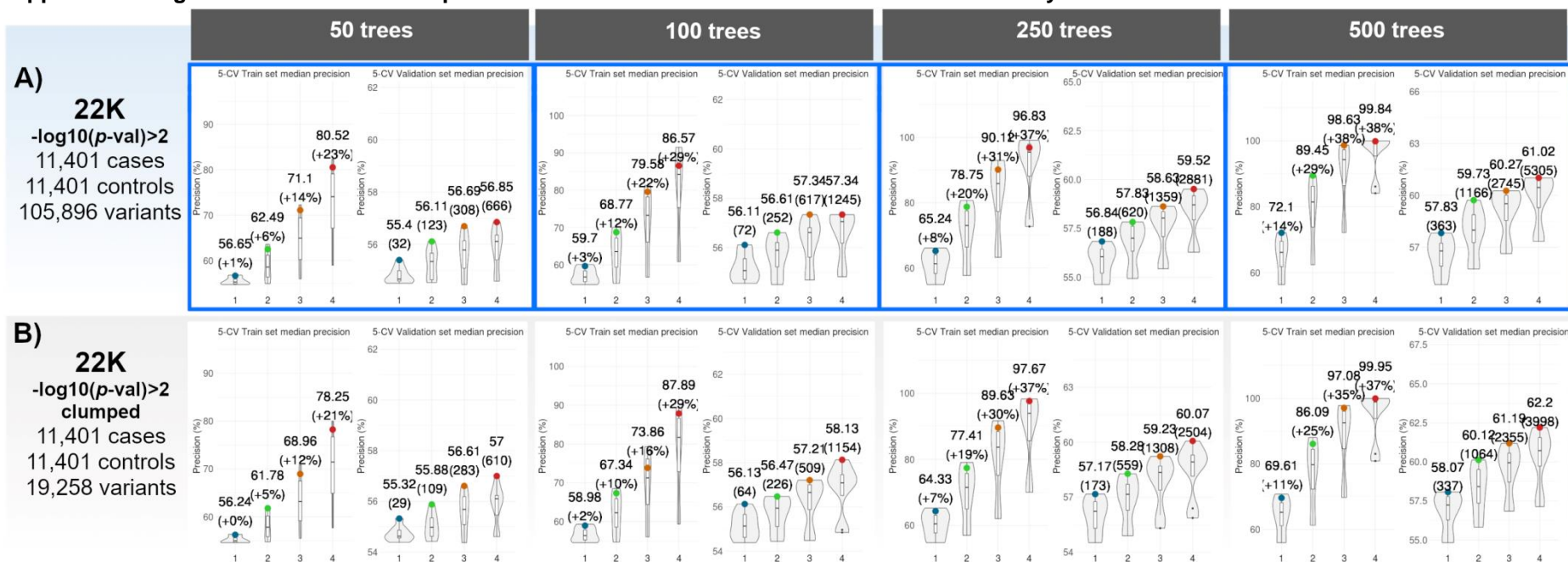
The effect of randomness in the prediction was evaluated based on a comparison between A) the 22K GWAS significant dataset (11,401 diabetic, 11,401 non-diabetic paired metadata individuals, 1,667 variants, $-\log_{10}(p\text{-value}) > 7$ (Bonàs-Guarch et al., 2018)), and 1,000 control randomizations of B) the genotype, and C) the phenotype. Each row shows the results obtained for the precision during the train (left) and validation (right) steps of the 5-fold cross-validation with hyperparameters adjustment (split (0.2, 0.3), learning rate (0.01, 0.04, 0.07, 0.1), number of trees (50, 100, 250, 500), and depth (1, 2, 3, 4)) for each scenario. The violin plots display the distribution of the percentage of precision (y axis) obtained by each combination of hyperparameters in terms of depth (x axis), and number of trees (columns), for each dataset. The coloured dots represent the results obtained with the best hyperparameters. The numbers inside the parentheses correspond to the median number of candidate interacting variants obtained during the training for the best hyperparameters. The squared data encapsulates the best results.

Supplemental Figure 3. Evaluation of the performance of XGBoost in terms of variable explanation.



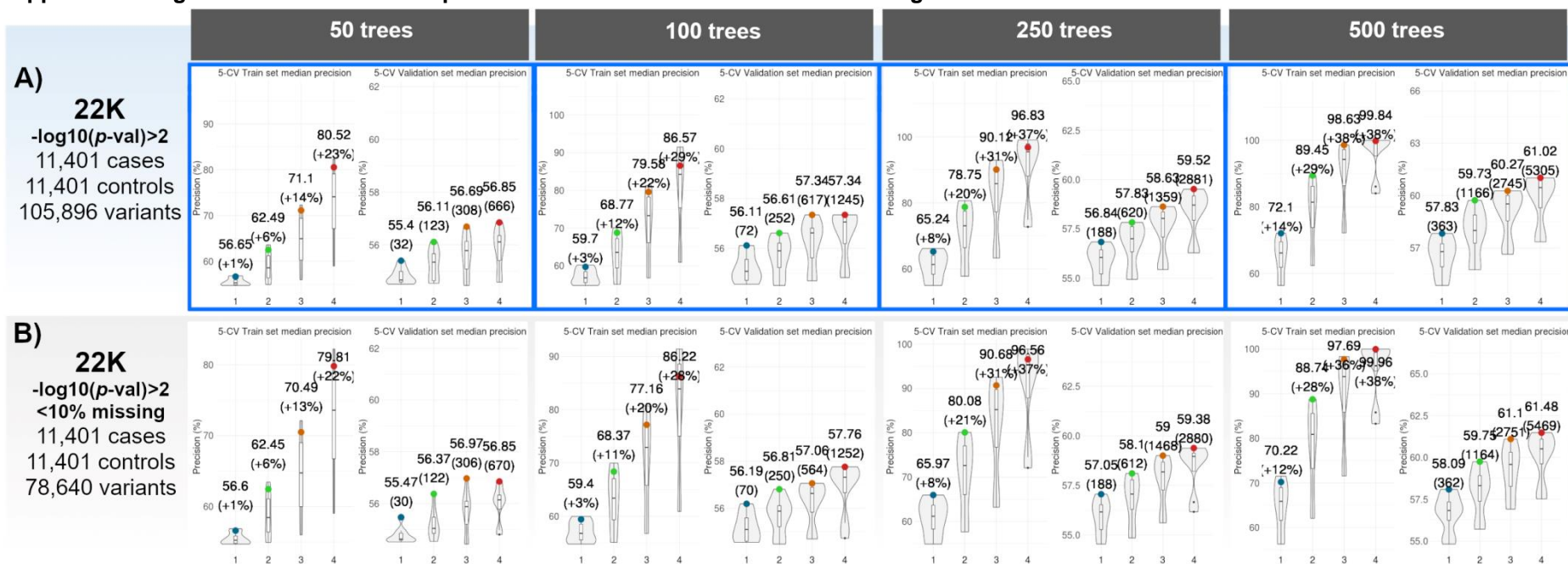
The effect of variable explanation in the prediction was evaluated based on the comparison between A) the 22K GWAS significant dataset (11,401 diabetic, 11,401 non-diabetic individuals, 1,667 GWAS significant variants, $-\log_{10}(p\text{-value}) > 7$ (Bonàs-Guarch et al., 2018)), and B) 1,000 control randomizations of the variants included in the discovery dataset (1,667 variants, $-\log_{10}(p\text{-value}) > 2$ (Bonàs-Guarch et al., 2018)). Each row shows the results obtained for the percentage of precision during the train (left) and validation (right) steps of the 5-fold cross-validation with hyperparameters adjustment (split (0.2, 0.3), learning rate (0.01, 0.04, 0.07, 0.1), number of trees (50, 100, 250, 500), and depth (1, 2, 3, 4)) for each dataset. The violin plots display the distribution of the percentage of precision (y axis) obtained by each combination of hyperparameters in terms of depth (x axis), and number of trees (columns), for each dataset. The coloured dots represent the results obtained with the best hyperparameters. The numbers inside the parentheses correspond to the median number of candidate interacting variants obtained during the training for the best hyperparameters. The squared data encapsulates the best results.

Supplemental Figure 4. Evaluation of the performance of XGBoost in terms of variable redundancy.



The effect of variable redundancy in the prediction was evaluated based on the comparison between A) the 22K discovery dataset (11,401 diabetic, 11,401 non-diabetic individuals, 105,896 variants, $-\log_{10}(p\text{-value}) > 2$ (Bonàs-Guarch et al., 2018)), and B) the discovery dataset after LD clumping ($r^2 = 0.2$, 250kb, $p\text{-value} = 0.5$, 70KforT2D summary statistics (Bonàs-Guarch et al., 2018)). Each row shows the results obtained for the percentage of precision during the train (left) and validation (right) steps of the 5-fold cross-validation with hyperparameters adjustment (split (0.2, 0.3), learning rate (0.01, 0.04, 0.07, 0.1), number of trees (50, 100, 250, 500), and depth (1, 2, 3, 4)) for each dataset. The violin plots display the distribution of the percentage of precision (y axis) obtained by each combination of hyperparameters in terms of depth (x axis), and number of trees (columns), for each dataset. The coloured dots represent the results obtained with the best hyperparameters. The numbers inside the parentheses in the 5-fold cross-validation training step correspond to the difference between the precision of the training and the validation for the best hyperparameters. The numbers inside the parentheses in the 5-fold cross-validation validation step correspond to the median number of candidate interacting variants obtained during the training for the best hyperparameters. The squared data encapsulates the best results.

Supplemental Figure 5. Evaluation of the performance of XGBoost in terms of missingness.



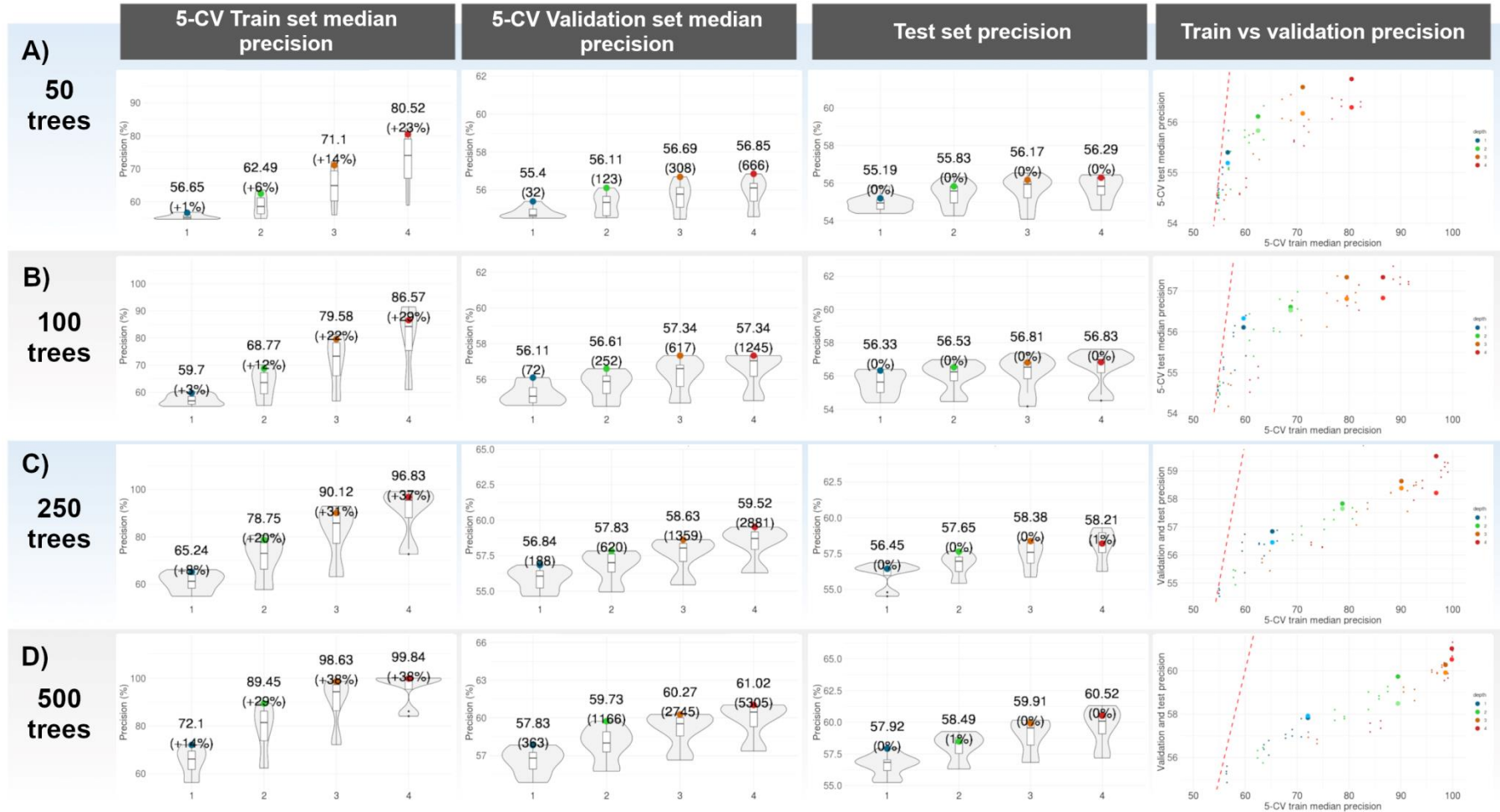
The effect of missingness in the prediction was evaluated based on the comparison between A) the 22K discovery dataset (11,401 diabetic, 11,401 non-diabetic individuals, 105,896 variants, $-\log_{10}(p\text{-value}) > 2$ (Bonàs-Guarch et al., 2018)), and B) the discovery dataset after filtering variants with over 10% of missing genotypes. Each row shows the results obtained for the percentage of precision during the train (left) and validation (right) steps of the 5-fold cross-validation with hyperparameters adjustment (split (0.2, 0.3), learning rate (0.01, 0.04, 0.07, 0.1), number of trees (50, 100, 250, 500), and depth (1, 2, 3, 4)) for each dataset. The violin plots display the distribution of the percentage of precision (y axis) obtained by each combination of hyperparameters in terms of depth (x axis), and number of trees (columns), for each dataset (rows). The coloured dots represent the results obtained with the best hyperparameters. The numbers inside the parentheses in the 5-fold cross-validation training step correspond to the difference between the precision of the training and the validation for the best hyperparameters. The numbers inside the parentheses in the 5-fold cross-validation validation step correspond to the median number of candidate interacting variants obtained during the training for the best hyperparameters. The squared data encapsulates the best results.

Supplemental Figure 6. Evaluation of the performance of XGBoost in terms of data availability.



The effect of data availability in the prediction was evaluated based on the comparison between A) the 22K discovery dataset (11,401 diabetic, 11,401 non-diabetic individuals, 105,896 variants, $-\log_{10}(p\text{-value}) > 2$ (Bonàs-Guarch et al., 2018), B) the PCs of the discovery dataset explaining a 95% of the variance, and C) the PCs of the discovery dataset representing the 10% of the number of observations (2,200 first PCs). Each row shows the results obtained for the percentage of precision during the train (left) and validation (right) steps of the 5-fold cross-validation with hyperparameters adjustment (split (0.2, 0.3), learning rate (0.01, 0.04, 0.07, 0.1), number of trees (50, 100, 250, 500), and depth (1, 2, 3, 4)) for each dataset. The violin plots display the distribution of the percentage of precision (y axis) obtained by each combination of hyperparameters in terms of depth (x axis), and number of trees (columns), for each dataset. The coloured dots represent the results obtained with the best hyperparameters. The numbers inside the parentheses in the 5-fold cross-validation training step correspond to the difference between the precision of the training and the validation for the best hyperparameters. The numbers inside the parentheses in the 5-fold cross-validation validation step correspond to the median number of candidate interacting variants obtained during the training for the best hyperparameters. The squared data encapsulates the best results.

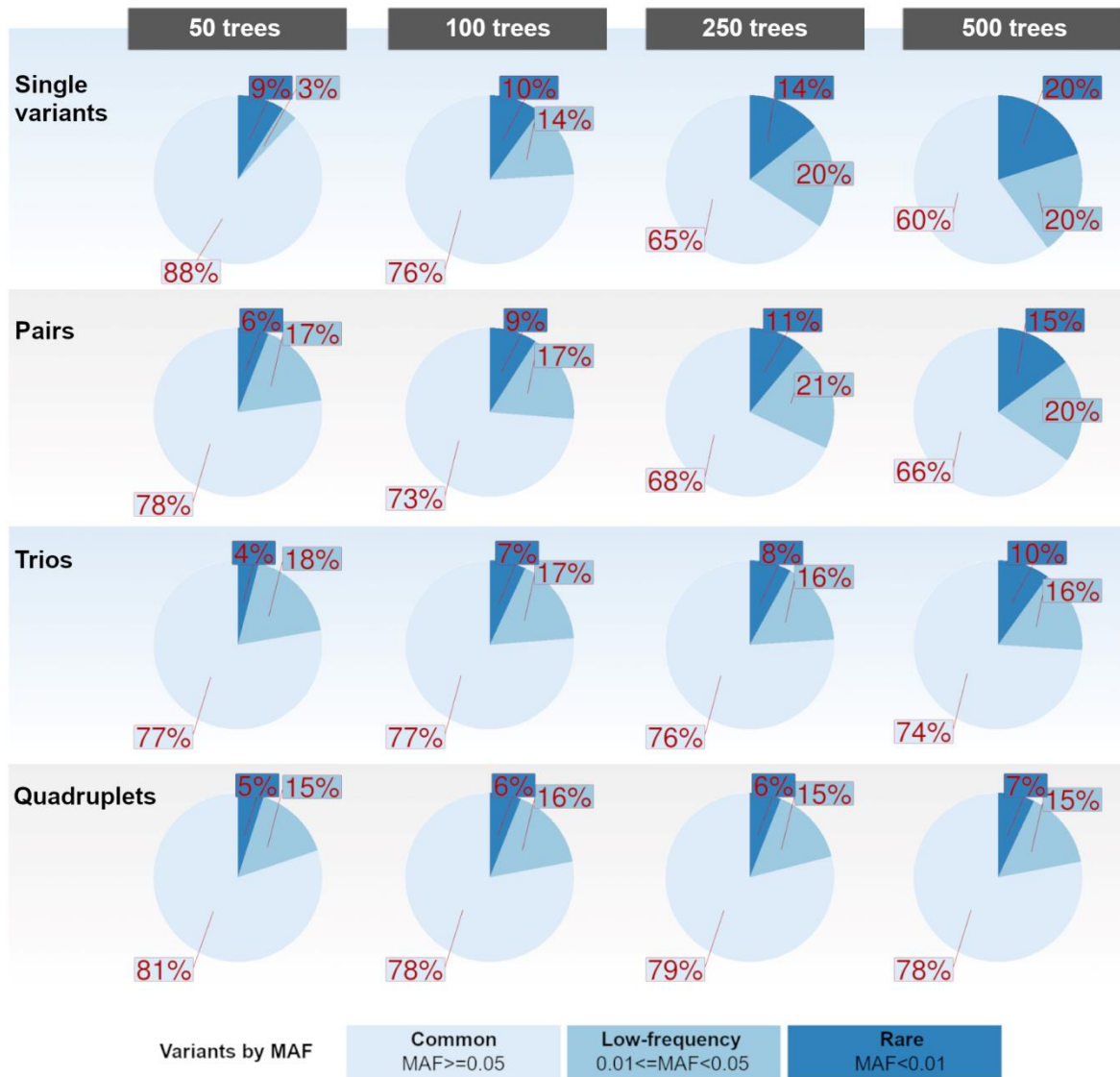
Supplemental Figure 7. Evaluation of the performance of XGBoost in terms of overfitting.



The effect of overfitting was evaluated based on the comparison between the 22K discovery dataset (11,401 diabetic, 11,401 non-diabetic individuals, 105,896 variants, $-\log_{10}(p\text{-value}) > 2$ (Bonàs-Guarch et al., 2018)) in the validation set under the 5-fold cross-validation and the test set. Each row shows the results obtained in terms of precision during the 5-fold cross-validation with hyperparameters adjustment (split (0.2, 0.3), learning rate (0.01, 0.04, 0.07, 0.1), number of trees (50, 100, 250, 500), and depth (1, 2, 3, 4) and test steps for different numbers of trees: A) 50, B) 100, C) 250, D) 500. The violin plots display the distribution of the percentage of precision (y axis), obtained by each combination of hyperparameters in terms of depth (x axis), and number of trees (row), in the training step (left) and the validation step (middle) of the 5-fold cross-validation,

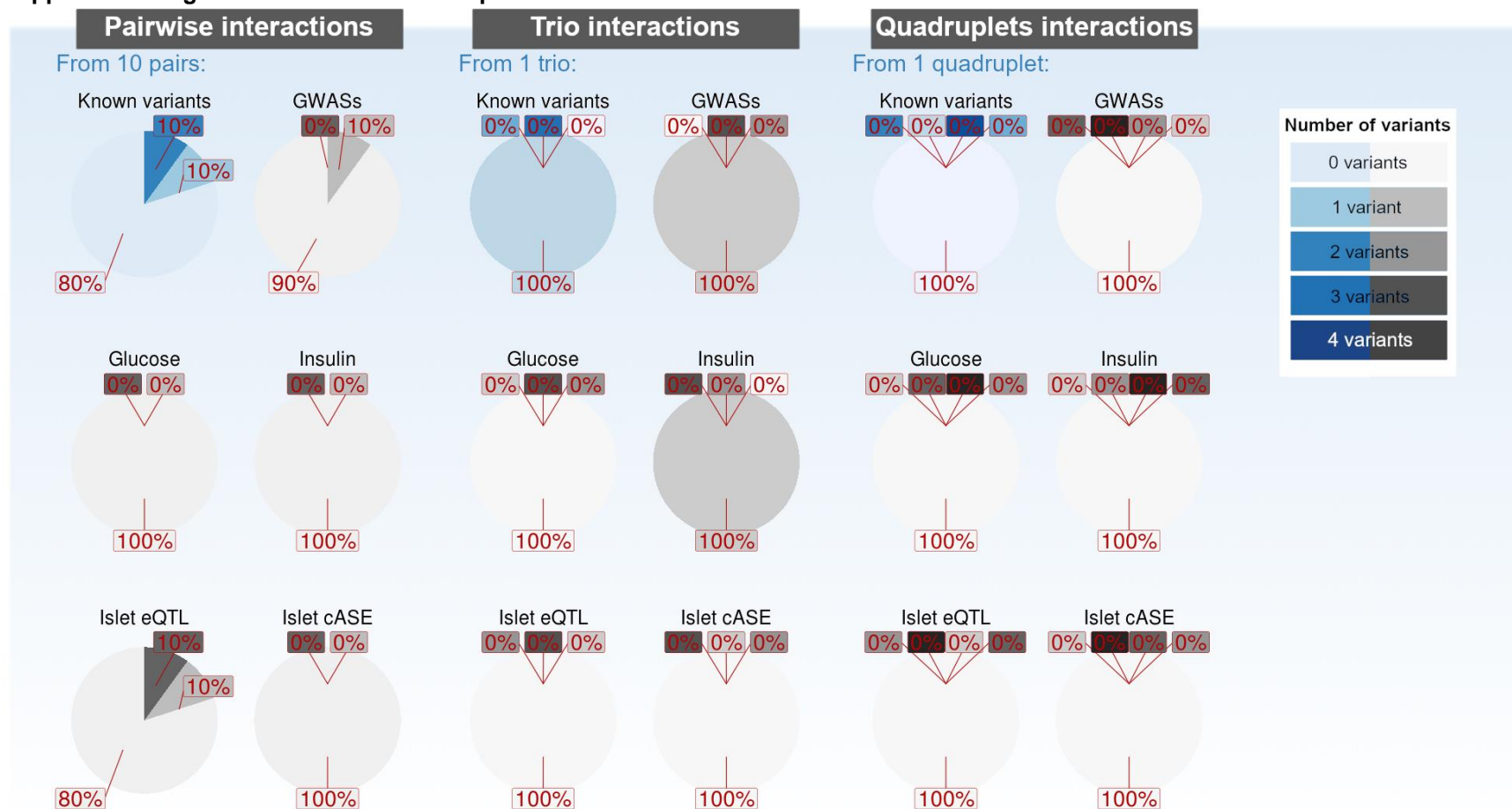
and in the test step (right). The coloured dots represent the results obtained with the best hyperparameters. The numbers inside the parentheses in the 5-fold cross-validation training step correspond to the difference between the precision of the training and the validation for the best hyperparameters. The numbers inside the parentheses in the 5-fold cross-validation validation step correspond to the median number of candidate interacting variants obtained during the training for the best hyperparameters. The numbers inside the parentheses in the test step represent the difference in precision between the validation and test set for the best hyperparameters. The scatterplots show the precision values obtained for each combination of hyperparameters, comparing the results obtained during the 5-fold cross-validation training (x axis) with the validation and test sets (y axis). The different colours of the dots correspond to different depths of the tree (1, 2, 3, 4). The bigger points represent the results obtained with the best hyperparameters in the validation (lighter colours) and the test (darker colours). The red dashed line is defined by the identity ($x=y$).

Supplemental Figure 8. Minor Allele Frequencies from the variants included in the groups obtained from the different scenarios (50, 100, 250, and 500 trees).



As a result from applying the machine learning algorithm, diverse candidate groups of epistatic variants were obtained depending on the number of trees (columns). The variants included in each group (single, pairs, trios, and quadruplets) were analysed to calculate their Minor Allele Frequencies (MAF). The pie charts display the percentage of common ($0.05 \leq \text{MAF}$), low-frequency ($0.01 \leq \text{MAF} < 0.05$), and rare variants observed in each group (rows).

Supplemental Figure 9. Annotations overlap.



The variants present in each group of epistatic variants (pairwise, trio, and quadruplets) were annotated with significant T2D GWAS meta-analysis (Bonàs-Guarch et al., 2018; Mahajan, Taliun, et al., 2018; Scott et al., 2017; The DIAGRAM Consortium et al., 2014), significant European ancestry glyceic traits meta-analysis (J. Chen et al., 2021), and significant human pancreatic islets eQTL and cASE expression results (Alonso, Piron, et al., 2021). The grey pie charts represent separately the proportion of epistatic variants previously associated with T2D or glyceic traits, and the proportion of epistatic variants with an already known effect on pancreatic islet expression. For each pie chart, the number of annotated variants inside a group is represented in a colour scale. The blue pie charts display the proportion of epistatic variants with any previously reported association with T2D, glyceic traits or pancreatic islet expression.

Supplemental Table 1. Evaluation of the performance of different machine learning methods in a subset of the discovery dataset (1,667 GWAS significant features, 11,401 cases, 11,401 controls).

Method	Nearest Neighbours	Linear SVM	RBF SVM	Gaussian Process	Decision Trees	Random Forest	Neural Networks	AdaBoost	Naive Bayes	QDA	XGBoost
Time	10min	25min	28min	>2h	2min	2min	2min	2min	2min	2min	2min
Score	0.52	0.54	0.5		0.54	0.53	0.55	0.56	0.55	0.51	0.56
Other	Nans	Nans	Nans		Nans	Nans	Nans	Nans	Nans	Nans	

Supplemental Table 2. Evaluation of the performance of XGBoost based under case-control imbalance.

Depth	N.trees	50			100			250			500		
		Best 70K test	Best 22K test	Best 22K predict	Best 70K test	Best 22K test	Best 22K predict	Best 70K test	Best 22K test	Best 22K predict	Best 70K test	Best 22K test	Best 22K predict
1	Precision (%)	44.65	55.27	8.32	44.65	56.05	10.64	44.65	56.96	10.89	44.59	56.86	10.79
	Accuracy (%)	79.51	55.32	53.67	79.51	55.75	56.07	79.51	56.66	56.53	79.49	56.53	56.33
	Recall (%)	46.5	55.66	52.97	46.5	53.13	51.47	46.5	54.47	52.3	46.54	54.06	52
	F1-score (%)	45.56	55.46	14.38	45.56	54.56	17.63	45.56	55.69	18.03	45.54	55.42	17.87
	MCC (%)	32.96	10.63	3.5	32.95	11.52	4.64	32.96	13.33	5.38	32.92	13.07	5.09
2	Precision (%)	44.65	56.46	10.7	44.65	56.02	8.65	44.65	56.5	8.65	44.23	56.19	10.88
	Accuracy (%)	79.51	56.25	55.39	79.51	55.95	55.68	79.51	56.41	55.46	79.31	56.15	55.61
	Recall (%)	46.5	54.53	52.85	46.5	55.31	52.68	46.5	55.66	53.02	46.85	55.7	53.64
	F1-score (%)	45.56	55.48	17.8	45.56	55.66	14.86	45.56	56.08	14.87	45.5	55.94	18.09
	MCC (%)	32.96	12.51	4.92	32.96	11.91	4.51	32.96	12.83	4.55	32.77	12.29	5.47
3	Precision (%)	44.65	56.31	8.64	44.65	57.07	10.91	44.42	56.99	10.71	44.4	57.02	10.79
	Accuracy (%)	79.51	56.24	55.27	79.51	56.94	55.83	79.4	56.72	55.48	79.39	56.76	55.98
	Recall (%)	46.5	55.61	53.15	46.5	55.96	53.49	46.77	54.71	52.79	46.77	54.85	52.53
	F1-score (%)	45.56	55.96	14.86	45.56	56.51	18.12	45.56	55.83	17.81	45.55	55.91	17.9
	MCC (%)	32.96	12.48	4.5	32.96	13.87	5.54	32.89	13.44	4.94	32.87	13.53	5.14
4	Precision (%)	44.59	57.03	8.73	43.53	56.38	10.79	44.31	57.13	10.8	43.04	56.38	10.79
	Accuracy (%)	79.49	56.94	55.49	78.96	56.28	55.52	79.35	56.95	55.48	48.69	56.28	55.52
	Recall (%)	46.58	56.18	53.54	47.62	55.44	53.21	46.81	55.67	53.33	48.32	55.44	53.21
	F1-score (%)	45.56	56.6	15.01	45.48	55.91	17.94	45.53	56.39	17.96	45.53	55.91	17.94
	MCC (%)	32.94	13.88	4.82	32.53	12.56	5.19	32.81	13.91	5.23	32.42	12.56	5.19

Supplemental Table 3. Evaluation of the average performance of XGBoost in terms of randomness.

Depth	N.trees	50		100		250		500	
	Dataset	Median Prec. (%)	Best Median N.Feat.	Median Prec. (%)	Best Median N.Feat.	Median Prec. (%)	Best Median N.Feat.	Median Prec. (%)	Best Median N.Feat.
1	Reference	54.84*	24	55.63*	42	56.65*	79	56.89*	72
	Random genotype	49.98	4	50	9	50	20	50	35
	Random phenotype	50.07*	4	50.06*	7	50.03	14	50.02	132
2	Reference	55.49*	79	56.62*	133	56.67*	165	56.55*	357
	Random genotype	50.01	83	49.99	180	50	240	50.01*	404
	Random phenotype	50.05*	18	50.04*	32	50.02	272	50	122
3	Reference	56.04*	182	56.53*	261	56.61*	365	56.5*	536
	Random genotype	50	66	50.01	131	49.99	755	50	937
	Random phenotype	50.03	43	50.01	89	50.02	449	50	522
4	Reference	56.24*	324	56.52*	389	56.5*	572	56.15*	516
	Random genotype	50	423	50	348	50.02*	1,307	50	1,598
	Random phenotype	50.02	100	50.03	345	50.02	546	50	801

* mean precision greater than 50% (5% significance level)

Supplemental Table 4. Evaluation of the performance of XGBoost in terms of variable explanation.

Depth	N.trees	50			100			250			500		
	Dataset	Median Prec. (%)	Best Prec. (%)	Best Median N.Feat.	Median Prec. (%)	Best Prec. (%)	Best Median N.Feat.	Median Prec. (%)	Best Prec. (%)	Best Median N.Feat.	Median Prec. (%)	Best Prec. (%)	Best Median N.Feat.
1	Reference	54.84	55.61	24	55.64	56.68	42	56.65*	57.03	79	56.89	57.01	72
	Random	54.13	55.84	26	54.38	56.44	49	55.01	57.77	123	55.89	58.63	235
2	Reference	55.49	56.75	79	56.62	57	133	56.67*	57	165	56.55*	56.84	357
	Random	54.38	56.45	94	54.87	57.85	177	56.12	59.17	394	57.33	60.62	639
3	Reference	56.04	56.82	182	56.53	57.07	261	56.61*	56.86	365	56.49**	56.84	536
	Random	54.63	57.17	225	55.44	58.28	402	56.92	59.92	739	58.24	60.83	1,069
4	Reference	56.25	56.75	324	56.52	56.8	389	56.5**	57.09	572	56.15**	56.82	516
	Random	54.91	57.57	450	55.88	58.43	707	57.46	60.33	1,076	58.75	61.28	1,297

* mean reference set precision equals to mean random set precision (5% significance level)

** mean reference set precision lower than mean random set precision (5% significance level)

Supplemental Table 5. Evaluation of the performance of XGBoost in terms of variable redundancy.

Depth	N.trees	50			100			250			500		
		Dataset	Median Prec. (%)	Best Prec. (%)	Best Median N.Feat.	Median Prec. (%)	Best Prec. (%)	Best Median N.Feat.	Median Prec. (%)	Best Prec. (%)	Best Median N.Feat.	Median Prec. (%)	Best Prec. (%)
1	Clumped	54.63*	55.32	29	55.13*	56.13	64	56.24*	57.17	173	57.24*	58.07	337
	Discovery	54.67	55.4	32	55.07	56.11	72	56.06	56.84	188	56.79	57.83	363
2	Clumped	54.98*	55.88	109	55.94*	56.47	226	57.16*	58.28	559	58.42*	60.12	1,064
	Discovery	55.34	56.11	123	55.9	56.61	252	57.01	57.83	620	58	59.73	1,166
3	Clumped	55.68*	56.61	283	56.65*	57.21	509	58.36*	59.23	1,308	59.95*	61.19	2,355
	Discovery	55.79	56.69	308	56.62	57.34	617	58.04	58.63	1,359	59.53	60.27	2,745
4	Clumped	56.1*	57	610	57.09*	58.13	1,154	58.93*	60.07	2,504	60.72*	62.2	3,998
	Discovery	56.11	56.85	666	57.06	57.34	1,245	58.71	59.52	2,881	60.46	61.02	5,305

* mean clumped set precision equals to mean discovery set precision (5% significance level)

Supplemental Table 6. Evaluation of the performance of XGBoost in terms of missingness.

Depth	N.trees	50			100			250			500		
		Dataset	Median Prec. (%)	Best Prec. (%)	Best Median N.Feat.	Median Prec. (%)	Best Prec. (%)	Best Median N.Feat.	Median Prec. (%)	Best Prec. (%)	Best Median N.Feat.	Median Prec. (%)	Best Prec. (%)
1	Discovery <10%miss	54.63*	55.47	30	55.1*	56.19	70	56.16*	57.05	188	56.83*	58.09	362
	Discovery	54.67	55.4	32	55.07	56.11	72	56.06	56.84	188	56.79	57.83	363
2	Discovery <10%miss	55.04*	56.37	122	55.88*	56.81	250	57.06*	58.1	612	58.32*	59.75	1,164
	Discovery	55.35	56.11	123	55.9	56.61	252	57.01	57.83	620	58	59.73	1,166
3	Discovery <10%miss	55.88*	56.97	306	56.63*	57.06	564	57.18*	59	1,468	59.58*	61.1	2,751
	Discovery	55.79	56.69	308	56.62	57.34	617	58.04	58.63	1,359	59.53	60.27	2,745
4	Discovery <10%miss	56.14*	56.85	670	57.32*	57.76	1,252	58.96*	59.38	2,880	60.52*	61.48	5,469
	Discovery	56.11	56.85	666	57.06	57.34	1,245	58.71	59.52	2,881	60.46	61.02	5,305

* mean discovery (<10% missing values) set precision equals to mean discovery set precision (5% significance level)

Supplemental Table 7. Evaluation of the performance of XGBoost in terms of data availability.

Depth	N.trees	50			100			250			500		
		Dataset	Median Prec. (%)	Best Prec. (%)	Best Median N.Feat.	Median Prec. (%)	Best Prec. (%)	Best Median N.Feat.	Median Prec. (%)	Best Prec. (%)	Best Median N.Feat.	Median Prec. (%)	Best Prec. (%)
1	PCA10	55.26*	56.53	23	56.25**	57.51	47	57.84**	59.28	113	59.17**	60.71	207
	PCA	55.33*	56.44	24	56.32**	57.19	48	57.73**	59.21	119	59.26**	60.35	227
	Discovery	54.67	55.4	32	55.07	56.11	72	56.06	56.84	188	56.79	57.83	363
2	PCA10	56.55**	57.41	81	57.66**	58.63	145	59.54**	60.56	357	60.79**	61.92	629
	PCA	56.46**	57.35	85	57.4**	58.3	168	59.22**	59.87	405	60.27**	61.17	746
	Discovery	55.35	56.11	123	55.9	56.61	252	57.01	57.83	620	58	59.73	1,166
3	PCA10	57.27**	58.06	191	58.21**	59.3	345	60.12**	60.79	802	61.48**	62.04	1,874
	PCA	57.06**	57.62	205	57.89**	58.67	429	59.27**	59.77	911	60.35**	60.96	1,752
	Discovery	55.79	56.69	308	56.62	57.34	617	58.04	58.63	1,359	59.53	60.27	2,745
4	PCA10	57.21**	57.71	469	58.74**	58.97	744	60.33**	60.92	1,448	61.81**	62.04	1,874
	PCA	56.83**	57.05	534	57.59**	58.01	993	59.17**	59.65	1,969	60.1*	60.82	3,268
	Discovery	56.11	56.85	666	57.06	57.34	1,245	58.71	59.52	2,881	60.46	61.02	5,305

* mean PCA and/or PCA10 set precision equals to mean discovery set precision (5% significance level)

** mean PCA and/or PCA10 set precision greater than mean discovery set precision (5% significance level)

Supplemental Table 8. Evaluation of the performance of XGBoost in terms of overfitting.

Depth	50			100			250			500		
	Median Prec. 5-CV val. (%)	Median Prec. test (%)	Best Prec. test (%)	Median Prec. 5-CV val. (%)	Median Prec. test (%)	Best Prec. test (%)	Median Prec. 5-CV val. (%)	Median Prec. test (%)	Best Prec. test (%)	Median Prec. 5-CV val. (%)	Median Prec. test (%)	Best Prec. test (%)
1	54.67*	54.95	55.19	55.07*	55.64	56.33	56.06*	56.42	56.45	56.79*	56.84	57.92
2	55.35*	55.58	55.83	55.9*	56.25	56.53	57.01*	56.97	57.65	58*	58.35	58.49
3	55.79*	55.95	56.17	56.62*	56.53	56.87	58.04*	57.6	58.38	59.53*	59.56	59.91
4	56.11*	55.84	56.29	57.06*	56.99	56.83	58.71*	58.3	58.21	60.46*	60.09	60.52

* mean 5-CV validation set (5-CV val) precision equals to mean test set (test) precision (5% significance level)

Supplemental Table 9. Evaluation of the relation between candidate epistatic groups of variants by depth and by tree.

type	depth1	trees1	depth2	trees2	N1	N2	LD all (r ² >=80)	LD all (r ² >=20)	LD some (r ² >=80)	LD some (r ² >=20)	LD some (%) (r ² >=20)
by tree	1	50	1	100	32	72	3	3	3	3	9.37
	1	100	1	250	72	182	5	6	5	6	8.33
	1	250	1	500	182	367	8	10	8	10	5.49
	1	50	1	500	32	367	3	3	3	3	9.37
	2	50	2	100	96	195	0	0	25	29	30.20
	2	100	2	250	195	487	0	0	31	44	22.56
	2	250	2	500	487	980	0	0	79	111	22.79
	2	50	2	500	96	980	0	0	25	28	29.16
	3	50	3	100	200	400	0	0	53	62	31
	3	100	3	250	400	971	0	0	120	162	40.5
	3	250	3	500	971	1,952	0	0	283	441	45.42
	3	50	3	500	200	1,952	0	0	92	119	59.5
	4	50	4	100	391	755	0	0	174	215	54.98
	4	100	4	250	755	1,859	0	0	341	461	61.05
	4	250	4	500	1,859	3,607	0	0	907	1,350	72.62
	4	50	4	500	391	3,607	0	0	242	327	83.63
by depth	1	50	2	50	32	96	6	7	6	7	21.87
	2	50	3	50	96	200	0	0	25	27	28.12
	3	50	4	50	200	391	0	0	56	69	34.5
	1	100	2	100	72	195	7	9	7	9	12.5
	2	100	3	100	195	400	1	1	35	49	25.12
	3	100	4	100	400	755	0	0	114	156	39
	1	250	2	250	182	487	12	20	12	20	10.98
	2	250	3	250	487	971	0	0	92	134	27.51
	3	250	4	250	971	1,859	0	0	287	457	47.06
	1	500	2	500	367	980	36	51	36	51	13.89
	2	500	3	500	980	1,952	0	0	200	323	32.95
	3	500	4	500	1,952	3,607	0	0	693	1,117	57.22

Supplemental Table 10. Comparative table to evaluate the differences between the marginal effects in the additive logistic regression model and the model including interactions.

Depth	Number of groups	Variable	Median coeff. additive	Median coeff. additive + interaction	Kolmogorov-Smirnov results*
2	10	var1	-0.1029	-0.1925**	Equals
		var2	0.036	-0.004	Equals
3	1	var1	0.146	0.115	-
		var2	0.067	0.030	-
		var3	0.130	0.129	-
4	1	var1	-0.084	0.088	-
		var2	-0.204	0.457	-
		var3	-0.030	0.098	-
		va4	0.045	0.245	-

* 5% significance level

** mean coefficient from the additive model different to mean coefficient from the full model (5% significance level)

Supplemental Table 11. Logistic regression coefficients of 3 examples of variant interaction with a change in variants effect on T2D.

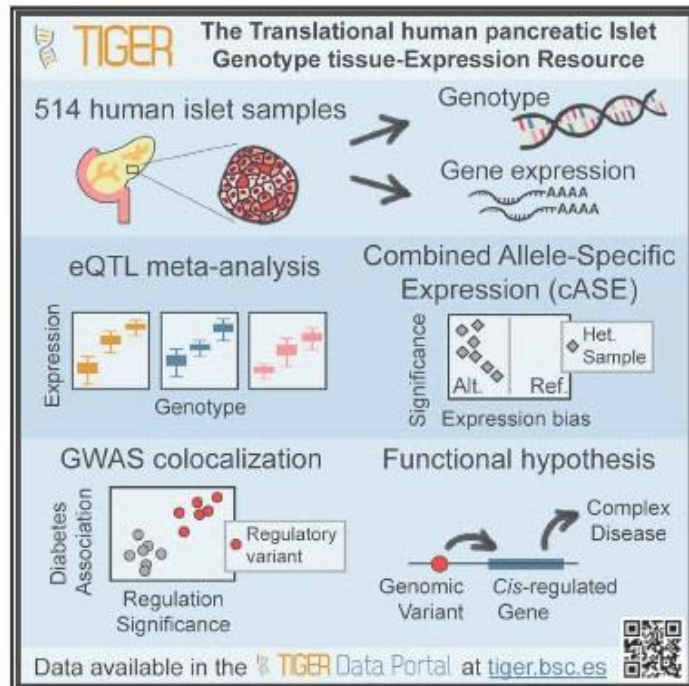
Variants		chr4:96761220 chr1:206513621		chr9:89501123 chr21:25168622		chr1:104373712 chr1:147362531 chr2:147085498 chr11:97009227	
Log.reg. models	Variables	Effect	p-value	Effect	p-value	Effect	p-value
Additive model	var1	0.383729 (OR~1.46)	9.94x10 ⁻⁵	0.125449 (OR~1.13)	2.42x10 ⁻³	-0.013520 (OR~0.98)	7.48x10 ⁻¹
	var2	0.062412 (OR~1.06)	1.05x10 ⁻¹	0.027540 (OR~1.02)	3.78x10 ⁻¹	-0.217645 (OR~0.80)	3.28x10 ⁻⁵
	var3					-0.032436 (OR~0.96)	2.95x10 ⁻¹
	var4					0.042759 (OR~1.04)	1.59x10 ⁻¹
Additive model + interactions (Full model)	var1	-0.698317 (OR~0.49)	3.32x10 ⁻³	-0.124841 (OR~0.88)	7.77x10 ⁻²	0.161312 (OR~1.17)	7.37x10 ⁻²
	var2	0.030166 (OR~1.03)	4.40x10 ⁻¹	-0.030208 (OR~0.97)	3.73x10 ⁻¹	0.440212 (OR~1.55)	3.05x10 ⁻²
	var3					0.105327 (OR~1.11)	1.97x10 ⁻¹
	var4					0.253931 (OR~1.28)	2.41x10 ⁻³
	var1var2	1.342456 (OR~3.82)	3.39x10 ⁻⁷	0.381140 (OR~1.46)	1.28x10 ⁻⁵	-1.209732 (OR~0.29)	1.10x10 ⁻⁵
	var1var3					-0.081683 (OR~0.92)	4.41x10 ⁻¹
	var2var3					-0.671094 (OR~0.51)	1.03x10 ⁻²
	var1var4					-0.239613 (OR~0.78)	2.78x10 ⁻²
	var2var4					-0.723912 (OR~0.48)	6.54x10 ⁻³
	var3var4					-0.172355 (OR~0.84)	9.81x10 ⁻²
	var1var2var3					1.181991 (OR~3.26)	7.28x10 ⁻⁴
	var1var2var4					1.803480 (OR~6.07)	5.00x10 ⁻⁷
	var1var3var4					0.099748 (OR~1.10)	4.61x10 ⁻¹
	var2var3var4					0.818495 (OR~2.26)	1.61x10 ⁻²
var1var2var3var4					-2.102776 (OR~0.12)	3.49x10 ⁻⁶	

TIGER

PUBLICATION

TIGER: The gene expression regulatory variation landscape of human pancreatic islets

Graphical abstract



Authors

Lorena Alonso, Anthony Piron, Ignasi Morán, ..., Josep M. Mercader, Miriam Cnop, David Torrents

Correspondence

mercader@broadinstitute.org (J.M.M.), mcnop@ulb.ac.be (M.C.), david.torrents@bsc.es (D.T.)

In brief

Understanding human islet regulatory genetic variation is essential to better understand the pathophysiology of diabetes and related diseases. Here, Alonso, Piron, Moran et al. present a comprehensive characterization of expression regulatory variation in >500 human islet samples and facilitate its access to the scientific community through the TIGER web portal.

Highlights

- Human pancreatic islets are key drivers of diabetes and related pathophysiology
- TIGER integrates omics and expression regulatory variation in 514 human islet samples
- TIGER expression regulatory variation allows the identification of diabetes effector genes
- The integrated human islet data in TIGER are publicly available through <http://tiger.bsc.es>



Alonso et al., 2021, Cell Reports 37, 109807
 October 12, 2021 © 2021 The Authors.
<https://doi.org/10.1016/j.celrep.2021.109807>



Resource

TIGER: The gene expression regulatory variation landscape of human pancreatic islets

Lorena Alonso,^{1,25} Anthony Piron,^{2,3,25} Ignasi Morán,^{1,25} Marta Guindo-Martínez,¹ Sílvia Bonàs-Guarch,^{4,5} Goutham Atla,^{4,5} Irene Miguel-Escalada,^{4,5} Romina Royo,¹ Montserrat Puiggròs,¹ Xavier Garcia-Hurtado,^{4,5} Mara Suleiman,⁶ Lorella Marselli,⁶ Jonathan L.S. Esguerra,⁷ Jean-Valéry Turatsinze,² Jason M. Torres,^{8,9} Vibe Nylander,¹⁰ Ji Chen,¹¹ Lena Eliasson,⁷ Matthieu Defrance,² Ramon Amela,¹ MAGIC,²⁴ Hindrik Mukder,¹² Anna L. Gloyn,^{9,10,13,14,15} Leif Groop,^{7,12,16} Piero Marchetti,⁶ Decio L. Eizirik,^{2,17} Jorge Ferrer,^{4,5,18} Josep M. Mercader,^{1,19,20,21,26,*} Miriam Cnop,^{2,22,26,27,*} and David Torrents^{1,23,26,*}

¹Life Sciences Department, Barcelona Supercomputing Center (BSC), Barcelona 08034, Spain

²ULB Center for Diabetes Research, Université Libre de Bruxelles, Brussels 1070, Belgium

³Interuniversity Institute of Bioinformatics in Brussels (IB2), Brussels 1050, Belgium

⁴Bioinformatics and Genomics Program, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology (BIST), Barcelona 08003, Spain

⁵Centro de Investigación Biomédica en Red de Diabetes y Enfermedades Metabólicas Asociadas (CIBERDEM) Barcelona 08013, Spain

⁶Department of Clinical and Experimental Medicine and AOUP Cisanello University Hospital, University of Pisa, Pisa 56126, Italy

⁷Unit of Islet Cell Exocytosis, Lund University Diabetes Centre, Malmö 214 28, Sweden

⁸Clinical Trial Service Unit and Epidemiological Studies Unit, Nuffield Department of Population Health, University of Oxford, Oxford OX3 7LF, UK

⁹Wellcome Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford, Oxford OX3 7LF, UK

¹⁰Oxford Centre for Diabetes, Endocrinology, and Metabolism, Radcliffe Department of Medicine, University of Oxford, Oxford OX3 7LE, UK

¹¹Exeter Centre of Excellence for Diabetes Research (EXCEED), University of Exeter Medical School, Exeter EX4 4PY, UK

¹²Unit of Molecular Metabolism, Lund University Diabetes Centre, Malmö 214 28, Sweden

¹³Division of Endocrinology, Department of Pediatrics, Stanford University School of Medicine, Stanford, CA 94304, USA

¹⁴NIHR Oxford Biomedical Research Centre, Churchill Hospital, Oxford OX3 7DQ, UK

¹⁵Stanford Diabetes Research Centre, Stanford University, Stanford, CA 94305, USA

¹⁶Finnish Institute of Molecular Medicine Finland (FIMM), Helsinki University, Helsinki 00014, Finland

¹⁷WELBIO, Université Libre de Bruxelles, Brussels 1050, Belgium

¹⁸Section of Epigenomics and Disease, Department of Medicine, Imperial College London, London SW7 2AZ, UK

¹⁹Programs in Metabolism and Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA

²⁰Diabetes Unit and Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA 02114, USA

²¹Department of Medicine, Harvard Medical School, Boston, MA 02115, USA

²²Division of Endocrinology, Erasmus Hospital, Université Libre de Bruxelles, Brussels 1070, Belgium

²³Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona 08010, Spain

²⁴Members of the MAGIC consortium are provided in Appendix S1

²⁵These authors contributed equally

²⁶Senior author

²⁷Lead contact

*Correspondence: mercader@broadinstitute.org (J.M.M.), mcnop@ulb.ac.be (M.C.), david.torrents@bsc.es (D.T.)

<https://doi.org/10.1016/j.celrep.2021.109807>

SUMMARY

Genome-wide association studies (GWASs) identified hundreds of signals associated with type 2 diabetes (T2D). To gain insight into their underlying molecular mechanisms, we have created the translational human pancreatic islet genotype tissue-expression resource (TIGER), aggregating >500 human islet genomic datasets from five cohorts in the Horizon 2020 consortium T2DSysTems. We impute genotypes using four reference panels and meta-analyze cohorts to improve the coverage of expression quantitative trait loci (eQTL) and develop a method to combine allele-specific expression across samples (cASE). We identify >1 million islet eQTLs, 53 of which colocalize with T2D signals. Among them, a low-frequency allele that reduces T2D risk by half increases *CCND2* expression. We identify eight cASE colocalizations, among which we found a T2D-associated *SLC30A8* variant. We make all data available through the TIGER portal (<http://tiger.bsc.es>), which represents a comprehensive human islet genomic data resource to elucidate how genetic variation affects islet function and translates into therapeutic insight and precision medicine for T2D.





INTRODUCTION

Diabetes is a complex metabolic disease, characterized by elevated blood glucose levels, that affects >463 million people worldwide. Type 2 diabetes (T2D) accounts for >85% of diabetes cases and is strongly related to age, obesity, and sedentary lifestyle. Epidemiologic studies forecast increases in global prevalence up to 25% by 2030 (Khan et al., 2020; Saeedi et al., 2019; Wild et al., 2004). This makes the study and understanding of diabetes a top research and healthcare priority. Progressive pancreatic islet dysfunction is central to the majority of all types of diabetes and thereby key to gain insight into disease pathophysiology.

Great efforts have been dedicated to uncover the link between genetic variation and complex disease susceptibility through large-scale genetic studies. For T2D, >700 genetic loci have been identified to date (Bonàs-Guarch et al., 2018; Mahajan et al., 2018; Spracklen et al., 2020; Vujkovic et al., 2020). The vast majority of variants in these loci do not disrupt protein coding sequences (Miguel-Escalada et al., 2019; Pasquali et al., 2014). Thus, the mechanisms by which these variants influence predisposition to disease remain to be elucidated. As the number of newly identified risk variants keeps increasing, their functional interpretation constitutes the main bottleneck to gain insight into the underlying molecular mechanisms and, thus, to develop more effective and targeted preventive and therapeutic strategies (Claussnitzer et al., 2020).

To provide functional interpretation of non-coding variation, large international efforts have generated and integrated genomic, transcriptomic, and epigenomic data from a large variety of healthy and diseased samples to build comprehensive and genome-wide maps of functional annotations. Among others, the Genotype-Tissue Expression (GTEx) project uses expression quantitative trait loci (eQTL) analysis to link genetic variation with gene expression across 54 different human tissues (Aguet et al., 2020). The Roadmap Epigenomics Mapping project (Bernstein et al., 2010) and the International Human Epigenome project (Bujold et al., 2016) also provide a broad characterization of epigenomic signatures in a variety of tissues and cell types.

The functional interpretation of genetic variants, which are usually associated with moderate or small effect sizes, requires tools and resources that focus on cells and tissues that are affected in the disease of interest. The islets of Langerhans, which are clusters of specialized endocrine cells that are essential to maintain glucose homeostasis, play a central role in the etiology of T2D (Eizirik et al., 2020; Krentz and Gloyn, 2020). Because human islets are difficult to obtain (Barovic et al., 2019; Burgarella et al., 2013; Meier et al., 2015), large multi-tissue resources such as GTEx do not contain islet data and at best use whole pancreas as a proxy, despite the fact that 97% of the pancreatic tissue consists of exocrine cells that mask islet signals. Hence, the development of publicly available resources and tools that include data on islets is essential to translate T2D genetic signals into molecular and physiological mechanisms.

The first studies of eQTL in human islets pinpointed genes that may be influenced by genetic variants and thus possibly mediate T2D risk (van de Bunt et al., 2015; Fadista et al., 2014). Despite the small number of samples, they identified a few loci linked

to differential expression of islet genes, which were enriched in genome-wide association study (GWAS) signals for T2D and related traits. More recently, the InsPIRE Consortium generated a large islet eQTL study with a sample size of 420 islet donors, which identified 46 T2D GWAS signals that colocalize with islet eQTL (Viñuela et al., 2020).

To further expand the understanding of human islet regulatory genomics and its role in T2D, the Horizon 2020 T2DSys consortium gathered an extensive collection of human islet samples with gene expression, epigenomic data, and genotypic and phenotypic information, with a total of 514 samples, 207 of which were analyzed by the InsPIRE Consortium. In this study, we discovered 40 T2D risk signals that colocalize with eQTL or ASE signals by improving genotype imputation methods and analyses and by developing a new method to combine allele-specific expression (cASE) across samples, knowledge previously unknown.

Importantly, the results from this study are made publicly available to the community through the Translational human pancreatic Islet Genotype tissue-Expression Resource (TIGER, <http://bsc.tiger.es>) portal (Figure 1A). This portal integrates the newly generated data with publicly available T2D genomic and genetic resources to facilitate the translation of genetic signals into their functional and molecular mechanisms.

RESULTS

A catalog of genetic variation and gene expression in human pancreatic islets

To study gene expression and the effects of genetic variation in human pancreatic islets, we obtained newly generated and published human islet data from 514 organ donors of European background, distributed across 5 cohorts (Center for Genomic Regulation, Lund University, University of Oxford/University of Alberta, Università di Pisa, and Université Libre de Bruxelles) (Method details). The large majority of these samples came from non-diabetic adult donors, and only 30 were from diabetic organ donors (Table S1).

The DNA of 307 samples was isolated, sequenced, and genotyped (Table S1; Method details) and aggregated to be harmonized with the existing data from 207 samples. After quality control, filtering of RNA sequencing (RNA-seq) and genotyping array data (Method details), we had both high-quality genotypes and RNA-seq data for 404 human islet samples (Figure 1B), including 21 from diabetic donors.

To fully characterize the genetic variation present in the samples, genotype imputation was performed separately for each cohort using 4 different reference panels, as previously described (Bonàs-Guarch et al., 2018; Guindo-Martínez et al., 2021), 1000 Genomes Project (The 1000 Genomes Project Consortium et al., 2015), Genome of the Netherlands (GoNL) (Boomsma et al., 2014), the Haplotype Reference Consortium (McCarthy et al., 2016), and UK10K (Walter et al., 2015). The results were integrated by selecting, for each variant, the imputed genotypes from the reference panel that achieved the best imputation quality (IMPUTE2 info score > 0.7; Method details). We have previously demonstrated that this approach results in increased overall coverage of genetic variation, as well as an

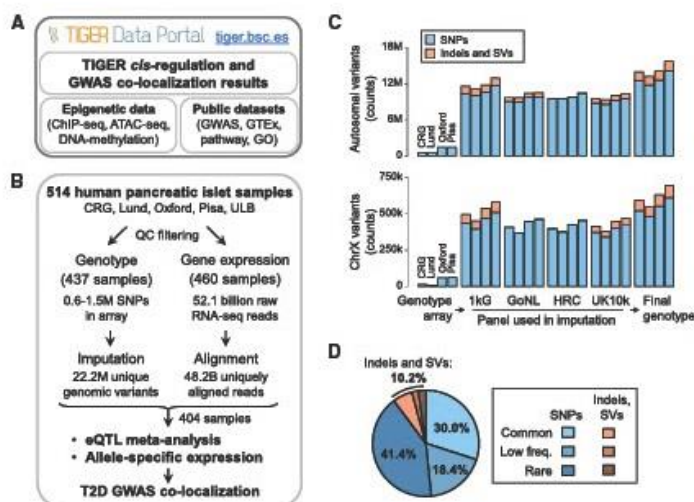


Figure 1. Project overview and genotype imputation

(A) Overview of the TIGER data portal. (B) Datasets of the T2DSystems Consortium and project workflow.

(C and D) Multi-panel genotype imputation identified 13.1–15.7 million autosomal variants (top) and 550,000–700,000 chrX variants (bottom) (C), with (D) a large proportion of low-frequency (minor allele frequency [MAF] 1%–5%) and rare (<1%) variants, with 10.2% of structural variants (SVs), including small indels and large SVs.

increased number of significant associations, including those that are covered by only one of the reference panels (Guindo-Martínez et al., 2021). This allowed imputation of >22 million unique high-quality genetic variants across all of the samples, 10% of which were indels and small structural variants (SVs), and >1.05 million variants in chromosome X (Figures 1C and 1D; Table S2). Notably, this strategy allowed the accurate imputation of 4 million low-frequency (minor allele frequency [MAF] between 0.05 and 0.01) and 10 million rare (0.01 > MAF > 0.001) variants.

In addition, we performed bulk RNA-seq in 514 human islet samples, 460 of which were retained after stringent quality control, including >52 billion raw short reads. We uniquely aligned >48 billion reads (median of 93 million per sample) (Table S3), which allowed us to observe >22,000 genes expressed at >0.5 transcripts per million (TPM) (Method details).

An atlas of eQTLs in human pancreatic islets

To explore the association between genetic variation and gene expression, we performed an eQTL meta-analysis across 4 cohorts. We performed a cis-eQTL analysis in 404 samples, using data from each cohort independently. For each analysis, we corrected for known covariates (age, sex, and body mass index [BMI]), 7 genetic ancestry principal components, and probabilistic estimation of expression residuals (PEER) factors for hidden confounding factors (Stegle et al., 2012). The eQTL results from each of the 4 cohorts were then meta-analyzed (Figure 2A). This resulted in >1.11 million significant eQTLs in >21,115 eGenes (12,802 protein coding genes, 8,313 non-coding) at a 5% false discovery rate (FDR) after Benjamini-Hochberg correction for multiple testing (Benjamini and Hochberg, 1995) (Figure 2B). The quantile-quantile plot showed no baseline inflation in the results. More than 12% of all significant eQTLs were small indels or larger SVs, and this type of variation was the top associated variant for

14% of all genes. This is in line with what has been observed in primary human immune cell types, in which indels comprised 12.5% of the variants in the 95% credible sets for eQTLs (Kundu et al., 2020), and in GTEx, in which SVs were found to have a stronger effect than single nucleotide variants (Chiang et al., 2017).

To assay the potential functional impact of the identified eQTL variants, we tested for their enrichment in human islet regulatory regions, defined by a variety of pancreatic islet chromatin assays (Miguel-Escalada et al., 2019). We observed that eQTL variants overlapped with gene promoters with very strong fold enrichment when compared with a control set of genetic variants (3.1-fold for 1% FDR eQTL variants, $p = 3 \times 10^{-166}$) (Method details), as well as with strong enhancers (Miguel-Escalada et al., 2019) (2-fold, $p = 1.4 \times 10^{-16}$), and open-chromatin regions (1.4-fold, $p = 3.9 \times 10^{-45}$) (Figures 2C and S1). These results are consistent with eQTL studies in other tissues (Aguet et al., 2020).

Next, we contrasted the TIGER human islet results with the latest GTEx eQTL datasets, which comprised 54 human tissues, including whole pancreas, but not islets (Aguet et al., 2020). Of all significant human islet eQTLs, 64.7% were also significant in at least 1 GTEx tissue, whereas 35.3% were exclusive to human islets (Figure 2D, left panel). Only 30.5% of human islet eQTLs were also significant in whole pancreas in GTEx, an overlap that is similar to the rest of the GTEx tissues (26% mean overlap with T2D-related tissues, 29% with other tissues), highlighting that whole pancreas is not a better proxy for pancreatic islets than other tissues. In addition, when considering rare and low-frequency variants, the proportion of TIGER islet exclusive eQTLs increased to 76.5% (Figure 2D, right panel). These observations highlight again the importance of assaying human islets, since a sizeable proportion of the eQTLs cannot be found in other tissues. Interestingly, these observations also held true when we compared TIGER results with recently published InsPIRE eQTLs (Viñuela et al., 2020). Because of its imputation approach, TIGER interrogated a larger number of genomic variants (Figure S2A). Overall, 56.1% of the significant eQTLs were exclusive to our analysis (not assayed or non-significant in InsPIRE; Viñuela et al., 2020) (Figure S2B). Identification of eQTLs driven by low-frequency or rare variants may be more clinically effective, as significant low-frequency variants tend to

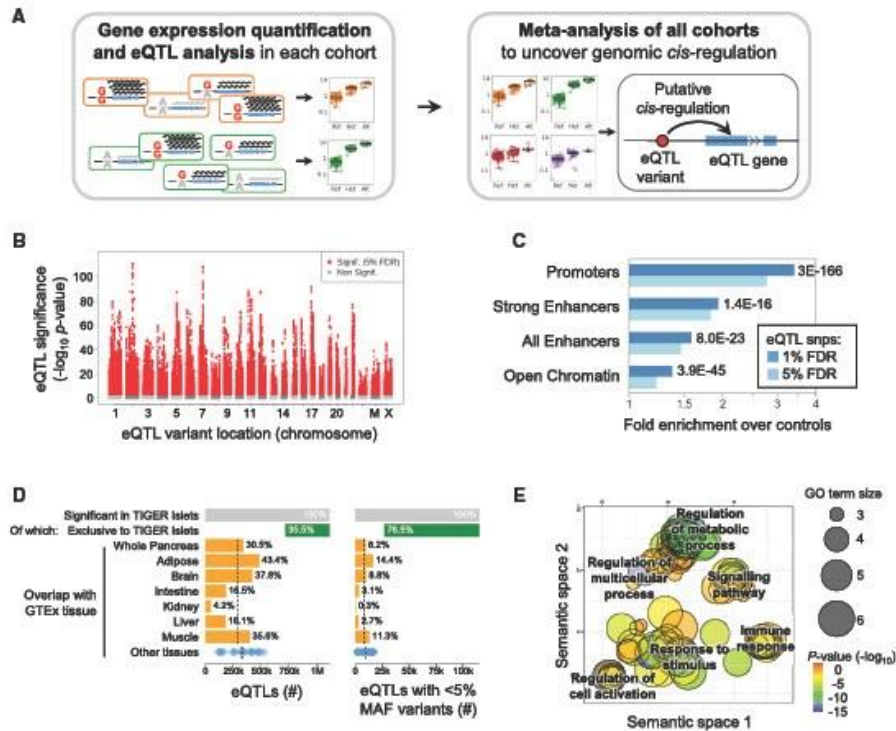


Figure 2. cis-eQTL meta-analysis in human pancreatic islets

(A) Overview of the meta-analysis.

(B) Manhattan plot of all eQTLs, including chrX, analyzed with female-only (F) or male-only (M) samples, and jointly (X).

(C) Fold enrichment over controls of significant eQTL variants, in islet regulatory chromatin regions. p values for 1% FDR eQTL enrichments are shown.

(D) Proportion of exclusive eQTLs in TIGER human islets (green) and previously found in GTEx project: tissues related to T2D etiology (orange), other tissues (blue); means in dashed lines. Right panel restricted to low MAF variants only.

(E) Gene Ontology analysis of the genes of TIGER-specific eQTLs.

have larger effects on disease risk and gene expression (Flanick, 2019). Notably, the proportion of TIGER exclusive eQTLs increased to 74.7% for low-frequency variants (Figure S2C), despite similar sample sizes between the studies. Overall, we identified 125,918 low-frequency eQTLs compared to 113,285 low-frequency eQTLs identified in the InsPIRE study (Figure S2C). This resulted in 20,742 eGenes, including the 69% of the 14,881 eGenes described in InsPIRE (Figure S2D). For eQTLs with variants present in both studies, the statistical strength of the association was correlated, as was the direction of effect for those <5% FDR significant in at least 1 of the 2 studies (Figures S2E and S2F). This indicates that the findings in the 2 studies are consistent, even when considering signals that did not reach significance in 1 of the 2.

Gene Ontology analysis of the significant human islet eQTL genes revealed signaling (including G protein-coupled receptor signaling) and metabolic regulation terms (Figure S3). In contrast, comparing TIGER-specific eQTL genes against those

also present in GTEx tissues revealed strong enrichment for these terms as well as “response to stimulus” or “regulation of cell activation,” and immune system terms (including “lymphocyte/T cell activation” and “regulation of immune system process”) (Figure 2E). This suggests that these eQTLs involve β cell physiology genes, including some related to immune processes with potential relevance for T1D (Ramos-Rodríguez et al., 2019).

Islet eQTLs colocalize with T2D GWAS signals

To assess whether the identified eQTLs can help to identify effector transcripts for T2D risk variants, we investigated the intersection between cis-eQTLs and known T2D associations (Bonás-Guarch et al., 2018; Mahajan et al., 2018; Vujkovic et al., 2020) by performing colocalization analyses using COLOC (Giambartolomei et al., 2014) (Method details).

This analysis uncovered 49 eQTL variants associated with the expression of 53 genes that significantly colocalized with T2D

Table 1. Human pancreatic islet colocalization of eQTL meta-analysis with T2D GWAS

Chr	SNP	Gene	COLOC		T2D GWAS					eQTL	
			PP.H4.abf	SNP.PP.H4	EAF	EA	NEA	OR	p	p	Direction
1	rs1127215	PTGFRN	1.00	0.99	0.42	T	C	0.95	2.3E-13	4.8E-15	--
1	rs1127215	CD101	1.00	0.96	0.42	T	C	0.95	2.3E-13	1.2E-7	--
1	rs1493694	NBPF7	0.81	0.09	0.11	T	C	1.09	2.1E-16	1.0E-5	?+?
1	rs340874	RP11-478J18.2	0.98	1.00	0.56	C	T	1.07	5.6E-26	1.3E-6	+++
1	rs4659836	TBCE	0.82	0.12	0.65	A	G	1.04	4.7E-9	2.9E-7	--
3	rs3887925	ST6GAL1	1.00	1.00	0.55	T	C	1.06	1.4E-17	2.1E-13	+++
3	rs3887925	AC007690.1	1.00	1.00	0.55	T	C	1.06	1.4E-17	5.2E-9	+++
3	rs7640294	SERBP1P3	0.97	0.06	0.56	A	C	1.04	3.0E-8	1.6E-9	+++
4	rs1531583	CPLX1	0.87	0.13	0.046	T	G	1.12	1.2E-12	1.2E-6	+++
4	rs1580278	BDH2	0.81	0.73	0.53	A	C	0.96	2.9E-10	1.1E-9	+++
4	rs58730668	ACSL1	0.89	0.04	0.14	C	T	0.93	1.0E-13	2.5E-5	+++
6	rs6557267	RGS17	0.94	0.08	0.42	T	C	1.04	6.0E-8	8.2E-8	--
8	rs1059592	RP11-582J16.5	0.81	0.12	0.35	A	G	1.03	4.5E-5	4.1E-15	--
8	rs77292833	LRP12	0.84	0.05	0.12	G	C	0.96	1.6E-5	8.1E-8	+++
9	rs10811660	CDKN2B-AS1	0.99	0.48	0.17	A	G	0.85	6.6E-79	1.6E-7	--
9	rs10963924	SAXO1	0.82	0.09	0.43	C	G	1.04	9.2E-10	1.6E-5	--
10	rs827237	PCBD1	0.99	0.19	0.21	T	C	1.04	2.3E-7	2.4E-10	--
11	rs15818	HMBS	0.84	0.06	0.4	G	A	1.03	4.5E-5	2.5E-7	+++
11	rs529623	FXYD2	0.92	0.83	0.52	C	T	0.97	5.8E-6	3.4E-7	+++
11	rs57635800	HSD17B12	0.95	0.24	0.29	A	G	1.05	8.5E-13	1.1E-19	--
12	rs731304	ABCC9	0.80	0.19	0.24	A	G	0.97	1.1E-5	3.0E-11	+++
12	rs76895963	CCND2	0.36	1.00	0.02	G	T	0.62	5.3E-70	1.7E-6	+++?
12	rs77864822	RMST	0.99	0.81	0.07	G	A	0.93	2.2E-8	2.9E-14	+++
12	rs77864822	RP11-528M18.2	0.95	0.17	0.07	G	A	0.93	2.2E-8	3.6E-6	+++
13	rs34584161	CDK8	1.00	0.98	0.24	G	A	0.95	2.9E-10	1.3E-17	--
13	rs488321	KL	0.98	0.27	0.83	C	T	0.95	6.8E-10	4.3E-6	+++
14	rs10151752	ACTR10	0.86	0.26	0.59	G	A	0.97	7.2E-8	4.0E-6	+++
14	rs1803283	RP11-600F2.4.7	0.81	0.02	0.65	T	C	1.04	1.4E-7	2.5E-5	+-
15	rs13737	RP11-817O13.8	0.84	0.10	0.24	T	G	0.96	7.3E-10	2.3E-6	+++
17	rs7218899	USP36	0.96	0.41	0.51	T	C	0.97	1.5E-6	2.4E-10	+++
17	rs8070260	ZNHIT3	0.94	0.13	0.53	G	A	0.97	1.1E-5	4.1E-8	--
18	rs303760	NPC1	0.95	0.08	0.36	T	C	1.03	3.8E-6	2.4E-24	--

Colocalizations not reported in [Viñuela et al. \(2020\)](#). The *R* COLOC package reports the approximate Bayesian factor posterior probability (*PP.H4.abf*) that there is one common causal variant and the posterior probability (*SNP.PP.H4*) that the *SNP* is the associated causal variant. The *GWAS* establishes the link between the *SNP* and T2D; the effect alleles (*EA*) with a frequency (*EAF*) are shown with the associated effect odds ratio (*OR*) and the *p* value. The *GWAS* data are as reported by the *DIAGRAM* Consortium ([Mahajan et al., 2018](#)). The eQTL *p* value is reported with the direction of the effect: up- ("+") or downregulation ("-") direction for the effect allele in the 4 meta-analysis cohorts (order: CRG, Oxford, Lund, and Pisa). "?" means that not enough samples are available in the cohort for the minor allele to compute a *p* value.

GWAS loci (Table S4), 32 of which were not previously reported (Table 1; Figure S4; Data S1). Among the 49 colocalizing signals (Data S1), rs77864822 (MAF = 0.07) minor allele (G) was associated with higher *RMST* (rhabdomyosarcoma 2 associated transcript) expression and decreased T2D risk (odds ratio [OR] = 0.93, $p = 2.2 \times 10^{-8}$) (Figure S4A). By interrogating the latest *GWAS* study on glycemic traits ([Chen et al., 2021](#)), we observed that the protective allele was associated with decreased fasting glucose ($\beta = -0.024$, $p = 4 \times 10^{-11}$), reduced HbA1c ($\beta = -0.087$, $p = 4.6 \times 10^{-4}$), and reduced 2-h glucose in an oral glucose tolerance test ($\beta = -0.064$, $p = 2.4 \times 10^{-4}$) (Table

S4). Interestingly, we identified two low-frequency variants (Figures 3C and 3G), which may have large effect sizes, that colocalized with gene expression, suggesting a target gene and direction of effect (i.e., whether the genetic variant is associated with increased or decreased gene expression). The variant rs1531583 colocalized with *CPLX1* expression (Figures 3A–3C). Interestingly, the same variant was associated with *PCGF3* but not with *CPLX1* gene expression in whole pancreas in GTEx (Figure 3B), demonstrating once again the importance of performing eQTL in the relevant tissue. A detailed analysis of enhancer chromatin marks in human islets showed that

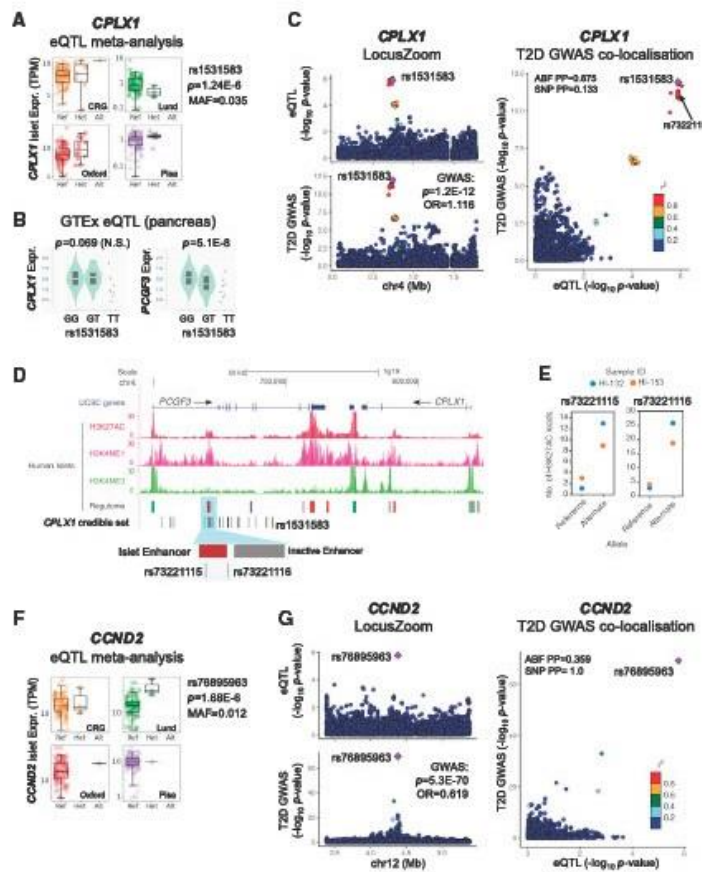


Figure 3. Examples of colocalization of pancreatic islet eQTLs with T2D GWAS

(A) Boxplots representing expression of *CPLX1* across different genotypes of variant rs1531583 in each of the cohorts and final meta-analysis results. (B) rs1531583 was not significant in GTEX whole pancreas for *CPLX1*, but instead it was for *PCGF3* (bottom). (C) LocusZoom plots of islet eQTL (top) and T2D GWAS (bottom) signals for rs1531583-*CPLX1*, and their co-localization (right). ABF, approximate Bayes factor, PP, posterior probability. (D) An islet enhancer overlaps with rs73221115 and rs73221116, part of the *CPLX1* credible set of SNPs. (E) Two human islet samples heterozygous for rs73221115 and rs73221116 showed allelic imbalance in their H3K27ac enhancer chromatin marks. (F) eQTL meta-analysis of *CCND2* and the low-frequency *cis*-regulatory variant rs76895963. (G) Co-localization plots for rs76895963-*CCND2*, as in (B).

to identify ASE in gene expression data in single (Edsgård et al., 2016; Mayba et al., 2014) or multiple samples (Fan et al., 2020; Liang et al., 2021), these methods did not aim to identify candidate *cis*-regulatory variants for the ASE effect.

We implemented a cASE pipeline for the analysis of ASE replicated across multiple samples that differ in age, gender, BMI, and environmental factors, thereby likely to stem from *cis*-regulatory genetic variants (Figure 4A). cASE analysis complements eQTL analysis, and additionally controls for (1) environmental and batch effects, which are important confounding factors in eQTL studies (Akey et al., 2007; Branham et al., 2007; Churchill, 2002; Fare et al., 2003; Irizarry et al., 2005; Yang et al., 2002); (2) sample heterogeneity, which is prevalent in human islets (Leek and Storey, 2007); and (3) *trans* effects, since these would affect the 2 alleles in the same manner and thus cannot result in ASE. cASE combines ASE from each sample in a single Z score statistic that summarizes overall ASE across the cohort of samples (Figure S5; Method details.) (Newhall et al., 1949). Variants that preferentially express the reference allele result in a positive Z score and vice versa (Figure 4A).

Using this strategy, we identified 2,707 genes with 5,271 reporter variants showing cASE in human islets, at 5% FDR (Figure 4B). The similar number of reference and alternate imbalanced variants (2,606 and 2,589, respectively) showed that alignment biases toward the reference allele were successfully controlled (Figures S5B–S5E).

When comparing cASE genes against a set of non-significant genes (matched by gene expression level, Method details), we observed that cASE genes were enriched for islet-specific

rs73221115 ($r^2 = 0.978$ with rs1531583) and rs73221116 ($r^2 = 0.98$ with rs1531583) had allele-specific H3K27ac binding, suggesting that these 2 variants are the most likely causal variants of the *CPLX1* locus (Figures 3D and 3E). We also identified significant colocalization between the low-frequency variant rs76895963, known to be associated with nearly half reduced T2D risk (Steinthorsdottir et al., 2014), and increased *CCND2* expression in islets (Figures 3F and 3G). This variant was also associated with reduced fasting glucose ($\beta = -0.033$, $p = 0.0017$), HbA1c ($\beta = -0.042$, $p = 3.6 \times 10^{-6}$), and 2-h glucose in oral glucose tolerance test ($\beta = -0.095$, $p = 0.01$) (Table S4).

An atlas of cASE in human pancreatic islets

Preferential expression of mRNA copies containing 1 of the 2 alleles of a genetic variant (allele-specific expression [ASE]) can result from *cis*-regulation. However, ASE can occur while the overall amount of expression of a gene remains constant, and therefore this type of regulation cannot be identified by conventional eQTL analysis. While some methods have been developed

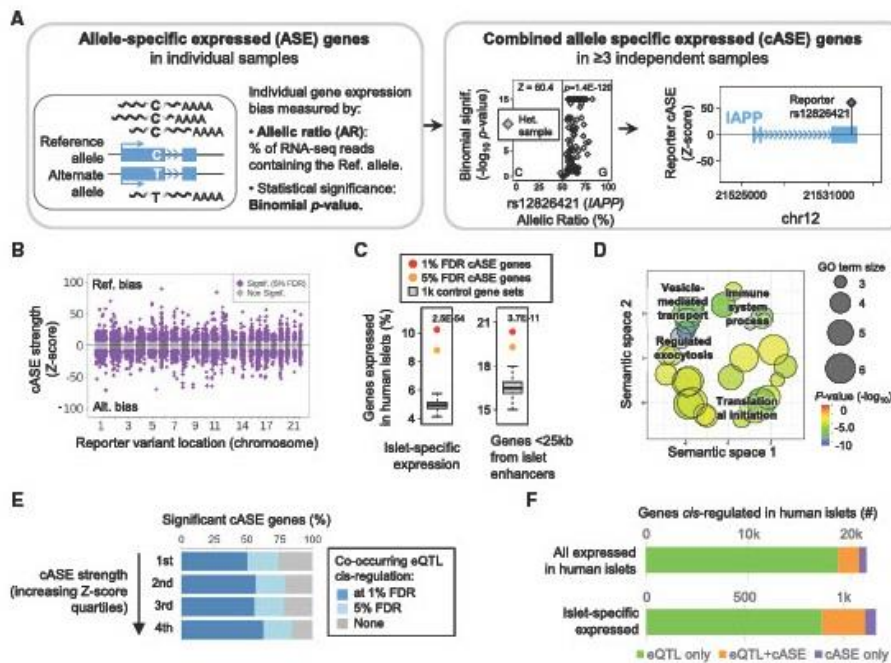


Figure 4. Combined ASE analysis in human islets

(A) Overview of the cASE analysis, with *IAPP* as example of a gene with an imbalanced reporter variant, rs12826421. (B) Manhattan plot of cASE, positive values refer to reference-biased genes, negative to alternate. (C) Significant cASE genes are enriched for islet-specific expression and proximity to islet-regulatory regions. p values for 1% FDR eQTL enrichments are shown. (D) Gene Ontology analysis of cASE significant genes. (E) In genes with significant cASE, the proportion of those also identified as eGenes grew with increasing cASE magnitude. (F) Total number of *cis*-regulated genes (top) and of islet-specific expressed (bottom), identified only by the eQTL analysis (green), cASE (purple), and both (orange).

expression (2.1-fold, $p = 2.5 \times 10^{-54}$ at 1% FDR) and preferentially located near islet regulatory regions (1.23-fold, $p = 3.7 \times 10^{-11}$) (Figure 4C). Gene Ontology analysis (Method details) revealed islet-specific terms such as “vesicle-mediated transport” and “regulated exocytosis” (Figure 4D), related to insulin production and secretion in β cells. As a notable example, the islet amyloid polypeptide gene (*IAPP*) was among the most imbalanced cASE genes. *IAPP* had 7 independent reporter SNPs at 1% FDR (Figure 4A, right panel), all of which had strong imbalance toward the reference allele in the >100 independent samples that were heterozygous for the variants. Notably, there were no significant eQTLs for this gene, highlighting the complementarity between the two methods to identify regulatory variation. These findings highlight the potential of cASE to identify genes involved in regulating pancreatic islet physiology.

Given that eQTL and cASE analyses are complementary methods to detect genes affected by *cis*-regulation, we assessed the concordance between each of them. We interrogated the proportion of genes with significant eQTL of all cASE genes across absolute Z score quartiles (strength of imbalance)

and observed that the proportion of eQTL genes increased with increasing Z scores (Figure 4E), indicating that stronger cASE effects were more likely to be also identified in eQTL analysis, and showing a correlation between the 2 effects.

Of 2,707 cASE significant genes, 2,052 (75.8%) were detected in eQTL analyses, whereas 655 (24.2%) were detected uniquely through cASE (Figure 4F, top panel). The same trend was observed when considering only islet-specific genes. Among 270 islet-specific significant eGenes detected by cASE, 218 were also detected by eQTL analysis, while the remaining 52 were exclusively found by cASE (Figure 4F, bottom panel).

Mapping distal cASE variants allows cASE colocalization analysis and implicates additional T2D effector genes

We next developed an approach to identify distal putative cASE regulatory variants by interrogating all of the variants within the same topologically associated domain as the reporter variant (i.e., the variant located in the transcribed gene region). For each candidate regulatory variant, we stratified samples

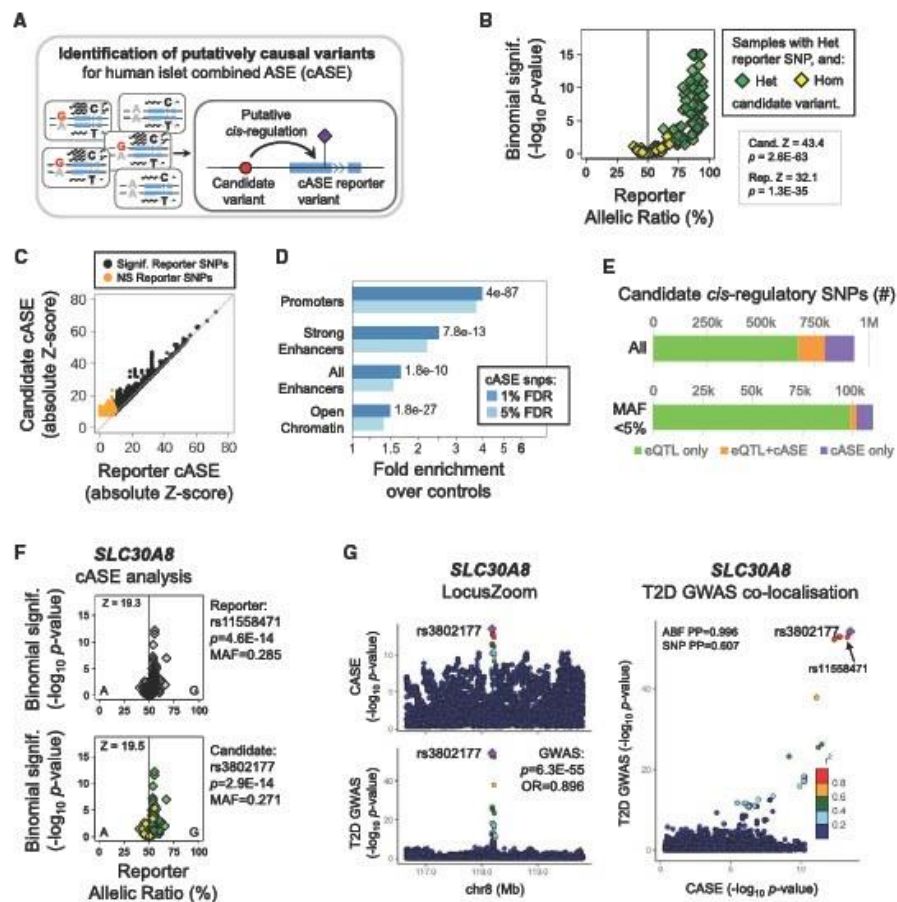


Figure 5. Identification of *cis*-regulatory variants in combined ASE

(A) Overview of the analysis.

(B) An example of *cis*-regulatory variant analysis; the samples Het for the candidate variant (green) have a higher cASE Z score for the reporter SNP, while samples that are Hom for the candidate (yellow) do not show significant imbalance for the reporter SNP.

(C) Candidate variants often have stronger Z scores than the reporters, including some reporter variants that were non-significant by themselves (orange).

(D) Fold enrichment over controls of significant cASE candidate *cis*-regulatory variants, in islet regulatory chromatin regions. p values for 1% FDR cASE enrichments.

(E) Total number of candidate *cis*-regulatory variants (top) and low-frequency variants (bottom) identified by only the eQTL analysis (green), cASE (purple), and both (orange).

(F) cASE analysis for *SLC30A8*, its best reporter SNP (top), and best candidate variant (bottom).

(G) LocusZoom plots of islet cASE (top) and T2D GWAS (bottom) signals for rs3802177-*SLC30A8*, and their colocalization (right).

between the heterozygous and homozygous for the candidate variant. We then recomputed cASE of the reporter variant (i.e., the transcribed variant) for each of the groups (Figure 5A). This approach allowed us to prioritize the candidate variant that had the highest reporter cASE when the candidate regulatory variant was also heterozygous, compared to when the regulatory variant was homozygous (Figure 5B; Method details). This method does not require haplotype phasing since it compares

heterozygous versus homozygous and is agnostic to the direction of the association.

This analysis uncovered 256,981 putative regulatory variants for 3,425 genes, including 570 genes that had no significant reporter variant by themselves, but that did reach significance upon stratifying by the genotype of regulatory variants (Figure 5C, orange points). To assay the potential functional impact of the identified reporter variants, we tested for their enrichment in

Table 2. cASE with T2D GWAS

Chr	SNP	Gene	COLOC		T2D GWAS					cASE				
			PP.H4.abf	SNP.PP.H4	EAF	EA	NEA	OR	p	Reporter variant	Ref	Alt	p	Z score
1	rs1127215	<i>PTGFRN</i>	0.99	0.98	0.42	T	C	0.95	2.3E-13	rs1127656	C	T	8.5E-9	14.6
4	rs10937721	<i>WFS1</i>	0.95	0.26	0.59	C	G	1.09	1.6E-40	rs1046320	G	A	3.2E-16	-20.9
8	rs3802177	<i>SLC30A8</i>	1.00	0.61	0.31	A	G	0.90	6.3E-55	rs11558471	A	G	2.9E-14	19.5
10	rs2280141	<i>PLEKHA1</i>	0.96	0.06	0.48	G	T	0.95	2.0E-13	rs1045216	A	G	1.7E-11	17.2
11	rs35251247	<i>HSD17B12</i>	0.95	0.21	0.29	A	G	1.05	8.5E-13	rs11555762	C	T	5.1E-93	52.9
11	rs35251247	<i>RP11-613D13.5</i>	0.93	0.07	0.29	A	G	1.05	8.5E-13	rs35251247	G	A	6.8E-12	-17.5
11	rs5215	<i>KCNJ11</i>	0.83	0.36	0.63	T	C	0.93	2.0E-26	rs5215	C	T	8.6E-6	-11.1
11	rs529623	<i>FXYD2</i>	0.95	1.00	0.52	C	T	0.97	5.8E-6	rs529623	T	C	3.4E-231	84.1
11	rs529623	<i>RP11-728F11.3</i>	0.91	0.81	0.52	C	T	0.97	5.8E-6	rs869789	G	A	7.2E-16	20.7
12	rs10879261	<i>TSPAN8</i>	0.85	0.08	0.41	G	T	1.05	3.7E-13	rs3763978	C	G	7.2E-11	-16.6
16	rs6600191	<i>ITFG3</i>	0.86	0.24	0.18	C	T	0.94	7.0E-13	rs7193384	C	G	1.1E-7	13.4
18	rs1788762	<i>C18orf8</i>	0.96	0.06	0.64	C	G	0.97	2.3E-6	rs1788820	A	G	3.2E-25	-26.7
18	rs1788762	<i>NPC1</i>	0.96	0.06	0.64	C	G	0.97	2.3E-6	rs1788820	A	G	3.2E-25	-26.7
19	rs3111316	<i>CALR</i>	0.99	0.47	0.59	A	G	1.05	1.6E-12	rs1049481	G	T	1.6E-76	-47.9

The R COLOC package reports the approximate Bayesian factor posterior probability (PP.H4.abf) that there is one common causal variant and the posterior probability (SNP.PP.H4) that the SNP is the associated causal variant. The GWAS establishes the link between the SNP and T2D; the effect alleles (EA) with a frequency (EAF) are shown with the associated effect OR and the p value. The GWAS data are as reported by the DIAGRAM Consortium (Mahajan et al., 2018). The cASE analysis provides the allelic imbalance for the allele represented by the reporter SNP with a reference allele (Ref) and an alternative allele (Alt), a p value (FDR threshold of 0.006), and a Z score. An increased Z score refers to increased expression of the reference allele.

human islet regulatory regions (Miguel-Escalada et al., 2019), observing overlap with gene promoters with very strong fold enrichment when compared with a control set of genetic variants (4-fold for 1% FDR eQTL variants, $p = 4 \times 10^{-89}$) (Method details), as well as with strong enhancers (Miguel-Escalada et al., 2019) (2.5-fold, $p = 7.8 \times 10^{-13}$) and open-chromatin regions (1.5-fold, $p = 1.8 \times 10^{-27}$) (Figure 5D). When comparing these cis-regulatory variants with the 1.11 million eQTLs, we found 123,748 variants were significant by both methods (3,138 with MAF <5%), and a further 133,233 (9,190 with MAF < 5%) were identified only by cASE (Figure 5E), showcasing the relevance of this analysis for enriching genetic cis-regulatory discovery.

Assigning statistical significance to cASE distal regulatory variants allowed us to test for colocalization between cASE regulatory variants and T2D GWAS variants. For each T2D GWAS locus, we assessed all of the regulatory variants for all of the imbalanced genes in the region and identified 14 colocalized locus-gene pairs (Table 2; Figure S6; Data S2). Of these, 6 had also been identified in eQTL/T2D GWAS colocalization analyses, showing consistency between the 2 methods. Interestingly, the 8 colocalizations identified by cASE alone, *WFS1*, *SLC30A8*, *RP11-613D13.5*, *KCNJ11*, *RP11-728F11.3*, *TSPAN8*, *C18orf8*, and *CALR*, suggested that these T2D variants may mediate disease risk by causing an imbalance in allelic expression, rather than altering overall gene expression (Figure S6). A notable example was the highly significant cASE observed in *SLC30A8* (rs11558471; $p = 2.9 \times 10^{-14}$), which showed colocalization with a well-established T2D-associated variant (Figures 5F and 5G; Table S5) for which there was no eQTL colocalization. Thus, cASE analysis uncovered additional disease-relevant genomic regulation and provides a potential biological mechanism underlying the association.

A web portal to explore regulatory variation and genomic pancreatic islet information

Finally, to provide the research community with a user-friendly open access tool to explore these findings and mine the molecular basis of complex diseases influenced by pancreatic islet biology, we created TIGER (<http://tiger.bsc.es>) (Figure S7). This portal integrates the results obtained in this study with other public genomic, transcriptomic, and epigenomic pancreatic islet resources, as well as T2D GWAS meta-analysis summary statistics (Method details).

The TIGER website represents homogeneous gene expression levels from 446 RNA-seq pancreatic islet samples corrected for batch and covariate effects, and enables comparison with GTEx expression data (Aguet et al., 2020) (Method details).

In addition to the eQTL and cASE results and to provide further functional assessment, we gathered islet regulatory information (Akerman et al., 2017; Miguel-Escalada et al., 2019; Pasquali et al., 2014), methylation marks (Hall et al., 2014; Thurner et al., 2018), and chromatin modification datasets (Dunham et al., 2012; Gaulton et al., 2010; Stitzel et al., 2010). Furthermore, to enable the translation of genetic variation to disease risk, we integrated the latest T2D GWAS meta-analysis summary statistics (Bonás-Guarch et al., 2018; Mahajan et al., 2014, 2018; Scott et al., 2017) (Figure 1A).

The TIGER database contains expression and molecular data for 59,625 Gencode genes (version gencode.v23lift37; Frankish et al., 2019) and >26 million variants. The portal allows users to perform both variant and gene-centric queries. The results are displayed in a set of graphical tools and a genomic browser (Down et al., 2011) that help visualize and interpret the molecular context of the query. Each table can be downloaded in csv



format, and the genomic browser integrates tools to search and zoom in on a region, add new tracks, and export the data as publication image. As a result of these efforts, the TIGER resource has already been used in recent studies (Hodson and Rorsman, 2020; Saponaro et al., 2020a, 2020b).

As an example, we present the visualization of *MTNR1B*, a gene associated with T2D and impaired insulin secretion (Lysenko et al., 2009). This gene is lowly expressed in pancreatic islets (median 0.25 TPM), but virtually absent in whole pancreas and other GTEx tissues (median 0 TPM), except for testis (median 0.61 TPM) and brain (median 0.06 TPM), highlighting the utility of this resource for studying human islet-specific expression (Figures S7A and S7B). A T2D risk-associated locus has been described and fine-mapped (Mahajan et al., 2018) to a single variant (rs10830963, $p = 4.8 \times 10^{-43}$, posterior probability [PP] = 0.99; Figures S4B and S7C). Notably, this variant is located within islet H3K27ac peaks, suggesting potential regulatory implications (Figure S7D). The close-up look at this locus illustrates that the TIGER portal can be easily used to interrogate gene expression and the epigenomic and genomic variation regulatory landscape, providing a very valuable resource to the research community to study complex diseases affecting pancreatic islets.

DISCUSSION

By analyzing a large multi-cohort dataset of pancreatic islets with gene expression and dense genotyping data, we have uncovered 1 million significantly associated variant-gene pairs. Of all of the associations we found, 35.3% were islet specific, highlighting the importance of performing tissue-specific eQTL studies (Figure 2D). Remarkably, 17 human islet eQTLs that colocalized with T2D GWAS signals were not associated with gene expression in any GTEx tissue, including whole pancreas, which emphasizes the fact that pancreas cannot be used as a proxy for pancreatic islets and vice versa.

We compared our findings with those obtained in the InsPIRE islet eQTL study that comprised 420 samples (Viñuela et al., 2020), 207 of which were also included in our study. We observed that 18 (34%) of the 53 eQTLs that colocalized with T2D GWAS signals were also identified in InsPIRE (Table S4). The improved power in our study obtained by the use of integrative approaches, such as combined reference panels genotype imputation and meta-analysis allowed us to detect lower MAF eQTL signals (10.4% with <5% MAF), representing a 7-fold increment of low-frequency eQTL variants compared to this previous islet eQTL study. Importantly, the meta-analyses also allow us to compare the heterogeneity of the associations between cohorts and filter out signals that are not consistent across cohorts, thereby avoiding false positives.

We uncovered 32 T2D colocalizations, 2 of which were led by low MAF variants, including variants associated with the expression of *CCND2*, *RMST*, and *CPLX1*. The variant rs76895963 (MAF = 0.02) that upregulates *CCND2* is associated with a nearly 50% reduced risk of T2D (OR = 0.58) (Mahajan et al., 2018; Steinhorsdottir et al., 2014) and is potentially implicated in the perinatal development of human β cells (Osonoi et al., 2020). While the PP of the colocalization was below the threshold of 0.8, the

SNP had a clear eQTL with the gene, and LocusCompare plots showed convincing colocalization (Figure 3G). The variant rs77864822 (MAF = 0.07) upregulates *RMST* expression and decreases T2D risk. *RMST* is a reportedly neuron-specific long non-coding RNA involved in neurogenesis (Ng et al., 2013); it is well expressed in human islet cells (Kaur et al., 2018), but its function in β cells is unknown. The variant rs1531583, with the minor T allele associated with increased T2D risk (Mahajan et al., 2018), upregulates *CPLX1*, encoding complexin-1, again, a reportedly neuron-specific gene. Complexin-1 plays a role in Ca^{2+} -dependent insulin exocytosis in rodent β cells, although it is intriguing that both *CPLX1* silencing and overexpression impaired insulin secretion (Abderrahmani et al., 2004). GWAS often report as a target the gene that is closest to the variant, in this case *PCGF3*. Notably, rs1531583 lies in an intronic region of *PCGF3* and is an eQTL for this gene in several GTEx tissues. In human islets, however, it is specifically associated with *CPLX1* expression and not with *PCGF3*, challenging the hypothesis that the closest gene is often the most likely target gene (Figures 3A–3E).

The imputation with 4 reference panels allowed us to analyze different sources of genetic variation, including indels and SVs. In our study, 12.6% of the eQTL are indels. This stresses the fact that indels are a significant part of the genetic background influencing RNA expression. Unfortunately, the largest available T2D GWAS dataset (Mahajan et al., 2018) did not consider indels, and so we could not include them in our colocalization analyses. In the near future, this approach could be used to fine-map the contribution of indels and SVs to disease risk.

Capitalizing on this valuable pancreatic islet resource, we also analyzed for the first time *cis*-regulation via ASE. We developed a method called cASE, which combines ASE across samples, maximizing the power to detect variants associated with ASE. We identified variants associated with allelic imbalanced expression while not changing overall gene expression, and thus undetectable by eQTL. We extended the cASE results in colocalization analysis and identified 14 T2D colocalizations. Among them, 8 signals non-detected in the eQTL/T2D GWAS colocalization included widely reported T2D-associated signals in *WFS1*, *SLC30A8*, *KCNJ11*, *TSPAN8*, *C18orf8*, and *CALR*. For these, the lead SNP causes allelic imbalance but no overall gene expression change. These findings suggest that a subset of regulatory genetic variants confer disease risk by causing imbalance in the allelic expression of their target genes, a mechanism for which knowledge is lacking. A particular locus of interest was the colocalization for common variant rs3802177 associated with *SLC30A8*. rs3802177 is in strong linkage disequilibrium with rs13266634 T2D-associated variant, widely discussed in the literature (Carvalho et al., 2017; Gupta and Vadde, 2020; Li et al., 2017; Sladek et al., 2007). In our study, both variants had nearly identical p values ($p = 2.9 \times 10^{-14}$ for rs3802177 and $p = 3.3 \times 10^{-14}$ for rs13266634), showing that either or both could induce allelic imbalance. Rare loss-of-function variants in *SLC30A8* strongly reduce T2D risk (Flannick et al., 2019) by enhancing insulin secretion (Dwivedi et al., 2019). However, the direction of effect of the common coding variants is not known. Our cASE results suggest that imbalanced expression toward the rs13266634-T allele is protective for T2D. Since

SLC30A8 loss-of-function decreases risk, these results suggest that the rs13266634-T allele may cause reduced *SLC30A8* function.

This study has a number of limitations. First, there is a substantial overlap of samples between the TIGER and InsPIRE studies. For the variants that were present in both studies, ~70% of TIGER eQTLs were also identified in InsPIRE. The difference in overlapping signals could be due to the lack of power to identify associations or to heterogeneity in the samples or eQTL methodology used. Since TIGER has samples overlapping with InsPIRE, we cannot consider TIGER a replication of InsPIRE results or vice versa. However, results identified in both studies can be considered confirmed. Future efforts should focus on the careful analysis of non-overlapping islet samples from the 2 initiatives. Power will increase further with the integration in TIGER of additional datasets by the human islet community, which we will warmly welcome. A second limitation of this study is that the majority of samples is of European ancestry. Hence, whereas it is a great resource for functional follow up of variants associated with diabetes and related traits, this resource is not useful as a follow-up of variants that are frequent enough only in non-European populations (Mercader and Florez, 2017; Spracklen et al., 2020; Vujkovic et al., 2020). Future human islet omics and genetic studies should focus on collecting data from diverse ancestries. Third, the analysis of pancreatic islet bulk RNA-seq data does not allow the comparison of different cell types that are present in pancreatic islets. Studies using single-cell sequencing will enable the identification of cell-type-specific eQTLs. However, large enough sample sizes of human islet single-cell RNA-seq and paired genotype array datasets are not available yet.

In summary, we generated a large expression regulatory variation resource in human pancreatic islets, a tissue with a central pathogenic role in most, if not all, types of diabetes. The results are available through the TIGER web portal, which constitutes a user-friendly visualization tool that facilitates the exploration of the datasets, democratizing human islet genomic information to all islet researchers and clinicians. We expect that this resource, in combination with the growing number of large-scale genetic and functional studies, will represent a critical step forward toward understanding the molecular underpinnings of complex diseases that affect pancreatic islet biology and provide a path for the identification of novel and personalized drug targets.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
 - Lead contact
 - Materials availability
 - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
 - Islet sample collection and genotyping
 - Centre for Genomic Regulation (CRG)
 - Lund University Diabetes Centre (Lund)

- University of Oxford/University of Alberta (Oxford)
- University of Pisa (Pisa)
- ULB Center for Diabetes Research (ULB)

● METHOD DETAILS

- Genotyping quality control
- Genotype phasing and imputation
- RNA-seq read mapping
- Sample concordance verification between genotype and gene expression
- TIGER web portal development

● QUANTIFICATION AND STATISTICAL ANALYSIS

- eQTL analysis
- Identifying variant regulatory enrichments using GREGOR
- Comparison of TIGER eQTLs with the GTEx and InsPIRE datasets
- Colocalization analysis
- Generation of an unbiased set of ASE reporter variants
- Identification of ASE
- Assessing cASE using Stouffer's Z-score
- Assessing the significance of cASE Z-scores
- Regulatory enrichment of cASE significant genes
- Gene ontology analyses and islet-specific expression
- Identifying candidate SNPs putatively leading to cASE
- Scaling human islet gene expression values to allow comparisons with the GTEx expression datasets in TIGER

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at <https://doi.org/10.1016/j.celrep.2021.109807>.

ACKNOWLEDGMENTS

This work has been supported by the European Union's Horizon 2020 research and innovation program T2DaySystems under grant agreement no. 667191. L.A. was supported by grant BES-2017-081635 of the Severo Ochoa Program, awarded by the Spanish government. I.M. was supported by the FJCI-2017-31878 Juan de la Cierva grant, awarded by the Spanish government. Work in the Onp and Etzirk labs was further supported by the Fonds National de la Recherche Scientifique (FNRS), the Brussels Region Innoviris project DiaType, and the Walloon Region SPW-EER Win2Wal project BetaSource, Belgium. D.L.E. is supported by a grant from the Wellcome-FNRS, Belgium. P.M., L.G., D.L.E., and M.C. are supported by the Innovative Medicines Initiative 2 Joint Undertaking Rhapsody, under grant agreement no. 115881, which is supported by the European Union's Horizon 2020 research and innovation programme, EFPIA and the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 16.0097. J.M.M. is supported by American Diabetes Association Innovative and Clinical Translational Award 1-19-ICTS-068. J.C. is supported by an Expanding Excellence in England Award from Research England. H.M., J.L.S.E., and L.E. are supported by the Swedish Strategic Research Foundation (RC15-0067). A.L.G. is a Wellcome Trust Senior Fellow in Basic Biomedical Science. This work was funded in Oxford and Stanford by the Wellcome Trust (095101, 200837, 106130, and 203141 [all to A.L.G.]) and the NIH (U01-DK105535 and U01-DK085545 [A.L.G.]). The research was funded by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC) (A.L.G.). I.M.-E. was supported by the EFDS/Novo Nordisk Rising Star Programme. Work in the Ferrer lab was supported by the Imperial College London Research Computing Service, the NIHR Imperial BRC, and the Centre for Genomic Regulation (CRG) genomics facility, and grants from Ministerio de Ciencia e

Innovación (BFU2014-54284-R and RTI2018-095666-B-I00), the Medical Research Council (MR/L02036X/1), the Wellcome Trust Senior Investigator Award (WT101033), and the European Research Council Advanced Grant (789055). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR, or the Department of Health. The technical support group from the Barcelona Supercomputing Center is gratefully acknowledged. Finally, we thank the entire Computational Genomics group at the BSC for their helpful discussions and valuable comments on the manuscript. We also acknowledge Cristian Opl and Lala Codó from the Barcelona Supercomputing Center for excellent website design and allocation of technical support and Isabelle Millard and Anylhal Musuaya from the ULB Center for Diabetes Research for excellent technical and experimental support.

AUTHOR CONTRIBUTIONS

LA, A.P., I.M., J.F., J.M.M., M.C., and D.T. conceived and planned the main analyses. J.F. provided unpublished allelic chromatin immunoprecipitation sequencing (ChIP-seq) and RNA-seq datasets and supervised cASE, which was developed and implemented by I.M. while he pursued his PhD in IDIBAPS and Imperial College London. I.M. further applied cASE in the TIGER dataset with the collaboration of L.A., M.G.-M., S.B.-G., M.P., R.A., and J.M.M. A.P. performed the eQTL and colocalization analyses with the collaboration of L.A., M.G.-M., S.B.-G., M.D., R.A., and J.M.M. LA developed the TIGER portal with the collaboration of R.R. and J.M.M. and performed the expression analysis with the collaboration of I.M., A.P., and J.M.M. I.M., A.P., LA, J.M.M., D.T., and M.C. wrote and edited the manuscript. G.A. and I.M.-E. contributed the islet regulatory data and analysis. I.M., S.B.-G., and J.F. contributed the Imperial and CRG data and analysis. J.L.S.E., L.E., H.M., and L.G. contributed the Lund data and analysis. J.-V.T., D.L.E., and M.C. contributed the ULB data and analysis. M.S., L.M., and P.M. contributed the Pisa data and analysis. M.S., L.M., and P.M. contributed the Pisa islet samples. J.L.S.E. contributed the Pisa sample sequencing. V.N. contributed the Pisa sample genotyping. J.M.T., V.N., and A.L.G. contributed the Oxford data and analysis and the genotyping of the Pisa samples. X.G.-H. prepared the chromatin immunoprecipitation, RNA, and DNA samples, and managed the CRG data generation. A.L.G., J.L.S.E., P.M., D.L.E., J.F., J.M.M., M.C., and D.T. provided guidance in the design and during the development of the project. D.L.E., M.C., and D.T. worked on the creation of TIGER. J.C. and MAGIC contributed the MAGIC data and analysis. J.M.M., M.C., and D.T. supervised the study.

DECLARATION OF INTERESTS

A.L.G.'s spouse is an employee of Genentech and holds stock options in Roche.

Received: May 28, 2021

Revised: July 23, 2021

Accepted: September 16, 2021

Published: October 12, 2021

REFERENCES

Abderahmani, A., Niederhauser, G., Plaisance, V., Roehrich, M.E., Lenain, V., Coppola, T., Regazzi, R., and Waerber, G. (2004). Complexin 1 regulates glucose-induced secretion in pancreatic β -cells. *J. Cell Sci.* 117, 2239–2247.

Aguet, F., Barbeira, A.N., Bonazzola, R., Brown, A., Castel, S.E., Jo, B., Kasela, S., Kim-Hellmuth, S., Liang, Y., Oliva, M., et al. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* 369, 1318–1330.

Akerman, I., Tu, Z., Beucher, A., Rolando, D.M.Y., Sauty-Colace, C., Benazra, M., Nakić, N., Yang, J., Wang, H., Pasquall, L., et al. (2017). Human Pancreatic β Cell lncRNAs Control Cell-Specific Regulatory Networks. *Cell Metab.* 25, 400–411.

Akey, J.M., Biswas, S., Leek, J.T., and Storey, J.D. (2007). On the design and analysis of gene expression studies in human populations. *Nat. Genet.* 39, 807–808, author reply 808–809.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al.; The Gene Ontology Consortium (2000). Gene ontology: tool for the unification of biology. *Nat. Genet.* 25, 25–29.

Barovic, M., Distler, M., Schöniger, E., Radisch, N., Aust, D., Weitz, J., Ibersson, M., Schulte, A.M., and Solimena, M. (2019). Metabolically phenotyped pancreatectomized patients as living donors for the study of islets in health and diabetes. *Mol. Metab.* 27S, S1–S6.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B* 57, 289–300.

Bernoulli, J. (1899). *Wahrscheinlichkeitsrechnung (Ars Conjectandi)* (W. Engelmann).

Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M.A., Beaudet, A.L., Ecker, J.R., et al. (2010). The NIH roadmap epigenomics mapping consortium. *Nat. Biotechnol.* 28, 1045–1048.

Bonàs-Guarch, S., Guindo-Martínez, M., Miguel-Escalada, I., Granup, N., Sebastian, D., Rodríguez-Fos, E., Sánchez, F., Planas-Félix, M., Cortes-Sánchez, P., González, S., et al. (2018). Re-analysis of public genetic data reveals a rare X-chromosomal variant associated with type 2 diabetes. *Nat. Commun.* 9, 321.

Boomsma, D.I., Wijmenga, C., Slagboom, E.P., Swertz, M.A., Karssen, L.C., Abdellou, A., Ye, K., Guryev, V., Vermaat, M., van Dijk, F., et al. (2014). The Genome of the Netherlands: design, and project goals. *Eur. J. Hum. Genet.* 22, 221–227.

Branham, W.S., Melvin, C.D., Han, T., Desai, V.G., Moland, C.L., Scully, A.T., and Fuscoe, J.C. (2007). Elimination of laboratory ozone leads to a dramatic improvement in the reproducibility of microarray gene expression measurements. *BMC Biotechnol.* 7, 8.

Bujold, D., Morais, D.A.L., Gauthier, C., Côté, C., Caron, M., Kwan, T., Chen, K.C., Laperle, J., Markovits, A.N., Pastinen, T., et al. (2016). The International Human Epigenome Consortium Data Portal. *Cell Syst.* 3, 496–499.e2.

Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* 47 (D1), D1005–D1012.

Burgarella, S., Merlo, S., Figliuzzi, M., and Remuzzi, A. (2013). Isolation of Langerhans islets by dielectrophoresis. *Electrophoresis* 34, 1068–1075.

Carvalho, S., Molina-López, J., Parsons, D., Corpe, C., Maret, W., and Hogstrand, C. (2017). Differential cytolocalization and functional assays of the two major human SLC30A8 (ZnT8) isoforms. *J. Trace Elem. Med. Biol.* 44, 116–124.

Chen, J., Spracklen, C.N., Marenne, G., Varshney, A., Corbin, L.J., Luan, J., Willems, S.M., Wu, Y., Zhang, X., Horikoshi, M., et al.; Lifelines Cohort Study; Meta-Analysis of Glucose and Insulin-related Traits Consortium (MAGIC) (2021). The trans-ancestral genomic architecture of glycemic traits. *Nat. Genet.* 53, 840–860.

Chiang, C., Scott, A.J., Davis, J.R., Tsang, E.K., Li, X., Kim, Y., Hadzic, T., Damani, F.N., Ganel, L., Montgomery, S.B., et al.; GTEx Consortium (2017). The impact of structural variation on human gene expression. *Nat. Genet.* 49, 692–699.

Churchill, G.A. (2002). Fundamentals of experimental design for cDNA microarrays. *Nat. Genet.* 32 (Suppl), 490–495.

Claussnitzer, M., Cho, J.H., Collins, R., Cox, N.J., Dermitzakis, E.T., Hurles, M.E., Kathiresan, S., Kenny, E.E., Lindgren, C.M., MacArthur, D.G., et al. (2020). A brief history of human disease genetics. *Nature* 577, 179–189.

Cnop, M., Abdulkarim, B., Bottu, G., Cunha, D.A., Igoillo-Esteve, M., Masini, M., Turatsinze, J.V., Griebel, T., Villate, O., Santin, I., et al. (2014). RNA

sequencing identifies dysregulation of the human pancreatic islet transcriptome by the saturated fatty acid palmitate. *Diabetes* 63, 1978–1993.

Delaneau, O., Marchini, J., and Zagury, J.F. (2011). A linear complexity phasing method for thousands of genomes. *Nat. Methods* 9, 179–181.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485, 376–380.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21.

Down, T.A., Pilipari, M., and Hubbard, T.J.P. (2011). Dailiance: interactive genome viewing on the web. *Bioinformatics* 27, 889–890.

Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., Epstein, C.B., Fritzer, S., Harrow, J., Kaul, R., et al.; ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74.

Dwivedi, O.P., Lehtvirta, M., Hastoy, B., Chandra, V., Krentz, N.A.J., Kleiner, S., Jain, D., Richard, A.M., Abaitua, F., Beer, N.L., et al. (2019). Loss of ZnT8 function protects against diabetes by enhanced insulin secretion. *Nat. Genet.* 51, 1596–1606.

Edsgård, D., Iglesias, M.J., Reilly, S.-J., Hamsten, A., Tomvall, P., Odeberg, J., and Emanuelsson, O. (2016). GeneASE: detection of condition-dependent and static allele-specific expression from RNA-seq data without haplotype information. *Sci. Rep.* 6, 21134.

Ezrin, D.L., Pasquall, L., and Cnop, M. (2020). Pancreatic β -cells in type 1 and type 2 diabetes mellitus: different pathways to failure. *Nat. Rev. Endocrinol.* 16, 349–362.

Fadista, J., Vilkin, P., Laakso, E.O., Mollet, I.G., Esguerra, J.L., Taneera, J., Storm, P., Osmark, P., Ladenvall, C., Prasad, R.B., et al. (2014). Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism. *Proc. Natl. Acad. Sci. USA* 111, 13924–13929.

Fan, J., Hu, J., Xue, C., Zhang, H., Suszkat, K., Reilly, M.P., Xiao, R., and Li, M. (2020). ASEP: gene-based detection of allele-specific expression across individuals in a population by RNA sequencing. *PLoS Genet.* 16, e1008786.

Fare, T.L., Coffey, E.M., Dai, H., He, Y.D., Kessler, D.A., Killian, K.A., Koch, J.E., LeProust, E., Marton, M.J., Meyer, M.R., et al. (2003). Effects of atmospheric ozone on microarray data quality. *Anal. Chem.* 75, 4672–4675.

Flannick, J. (2019). The Contribution of Low-Frequency and Rare Coding Variation to Susceptibility to Type 2 Diabetes. *Curr. Diab. Rep.* 19, 25.

Flannick, J., Mercader, J.M., Fuchsberger, C., Udler, M.S., Mahajan, A., Weese, J., Teslovich, T.M., Caulkins, L., Koesterer, R., Barajas-Olmos, F., et al.; Broad Genomics Platform; DiscovEHR Collaboration; CHARGE; LuCamp; ProDIGY; GoT2D; ESP; SIGMA-T2D; T2D-GENES; AMP-T2D-GENES (2019). Exome sequencing of 20,791 cases of type 2 diabetes and 24,440 controls. *Nature* 570, 71–76.

Frankish, A., Diekhans, M., Ferreira, A.M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J.M., Siu, C., Wright, J., Armstrong, J., et al. (2019). GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* 47 (D1), D766–D773.

Gaulton, K.J., Namm, T., Pasquall, L., Simon, J.M., Giresi, P.G., Fogarty, M.P., Panhuls, T.M., Mieczkowski, P., Secchi, A., Bosco, D., et al. (2010). A map of open chromatin in human pancreatic islets. *Nat. Genet.* 42, 255–259.

Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace, C., and Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* 10, e1004383.

Grafelman, J. (2015). Exploring diallelic genetic markers: the HardyWeinberg package. *J. Stat. Softw.* 64, 1–23.

Grafelman, J., and Camarena, J.M. (2008). Graphical tests for Hardy-Weinberg equilibrium based on the ternary plot. *Hum. Hered.* 65, 77–84.

Guindo-Martínez, M., Amela, R., Bonàs-Guarch, S., Puiggròs, M., Salvoro, C., Miguel-Escalada, I., Carey, C.E., Cole, J.B., Rieger, S., Atkinson, E., et al.;

FinnGen Consortium (2021). The impact of non-additive genetic associations on age-related complex diseases. *Nat. Commun.* 12, 2436.

Gupta, M.K., and Vadde, R. (2020). Insights into the structure-function relationship of both wild and mutant zinc transporter ZnT8 in human: a computational structural biology approach. *J. Biomol. Struct. Dyn.* 38, 137–151.

Haeussler, M., Zweig, A.S., Tyner, C., Speir, M.L., Rosenbloom, K.R., Raney, B.J., Lee, C.M., Lee, B.T., Hinrichs, A.S., Gonzalez, J.N., et al. (2019). The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res.* 47 (D1), D853–D858.

Hall, E., Volkov, P., Dayeh, T., Esguerra, J.L.S., Saló, S., Eliasson, L., Rönn, T., Bacos, K., and Ling, C. (2014). Sex differences in the genome-wide DNA methylation pattern and impact on gene expression, microRNA levels and insulin secretion in human pancreatic islets. *Genome Biol.* 15, 522.

Hodson, D.J., and Rorsman, P. (2020). A variation on the theme: SGLT2 inhibition and glucagon secretion in human islets. *Diabetes* 69, 864–866.

Irizarry, R.A., Warren, D., Spencer, F., Kim, I.F., Biswal, S., Frank, B.C., Gabrielson, E., Garcia, J.G.N., Geoghegan, J., Germino, G., et al. (2005). Multi-laboratory comparison of microarray platforms. *Nat. Methods* 2, 345–350.

Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., Sidropoulos, K., Cook, J., Gillespie, M., Haw, R., et al. (2020). The reactome pathway knowledgebase. *Nucleic Acids Res.* 48 (D1), D498–D503.

Jun, G., Flickinger, M., Hetrick, K.N., Romm, J.M., Doherty, K.F., Abecasis, G.R., Boehnke, M., and Kang, H.M. (2012). Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am. J. Hum. Genet.* 91, 839–848.

Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Aifoldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birbaumer, D.P., et al.; Genome Aggregation Database Consortium (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581, 434–443.

Kaur, S., Mirza, A.H., and Pociot, F. (2018). Cell type-selective expression of circular RNAs in human pancreatic islets. *Noncoding RNA* 4, 38.

Khan, M.A.B., Hashim, M.J., King, J.K., Govender, R.D., Mustafa, H., and Al Kaabi, J. (2020). Epidemiology of type 2 diabetes - Global burden of disease and forecasted trends. *J. Epidemiol. Glob. Health* 10, 107–111.

Krentz, N.A.J., and Gloyn, A.L. (2020). Insights into pancreatic islet cell dysfunction from type 2 diabetes mellitus genetics. *Nat. Rev. Endocrinol.* 16, 202–212.

Kundu, K., Mann, A.L., Tardagulla, M., Watt, S., Ponstingl, H., Vasquez, L., Morrell, N.W., Stagle, O., Pastinen, T., Sawcer, S.J., et al. (2020). Genetic associations at regulatory phenotypes improve fine-mapping of causal variants for twelve immune-mediated diseases. *bioRxiv*. <https://doi.org/10.1101/2020.01.15.907436>.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10, R25.

Leek, J.T., and Storey, J.D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* 3, 1724–1735.

Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12, 323.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079.

Li, L., Bai, S., and Sheline, C.T. (2017). HZnT8 (Slc30a8) transgenic mice that overexpress the R325W polymorph have reduced islet Zn²⁺ and proinsulin levels, increased glucose tolerance after a high-fat diet, and altered levels of pancreatic zinc binding proteins. *Diabetes* 66, 551–559.

Liang, Y., Aguet, F., Barbeira, A.N., Ardlie, K., and Im, H.K. (2021). A scalable unified framework of total and allele-specific counts for cis-QTL, fine-mapping, and prediction. *Nat. Commun.* 12, 1424.

- Liu, B., Gloudemans, M.J., Rao, A.S., Ingelsson, E., and Montgomery, S.B. (2019). Abundant associations with gene expression complicate GWAS follow-up. *Nat. Genet.* **51**, 768–769.
- Loh, P.-R., Danecek, P., Palamara, P.F., Fuchsberger, C., Reshet, Y.A., Finucane, H.K., Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G.R., et al. (2016a). Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443–1448.
- Loh, P.R., Palamara, P.F., and Price, A.L. (2016b). Fast and accurate long-range phasing in a UK Biobank cohort. *Nat. Genet.* **48**, 811–816.
- Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al.; GTEx Consortium (2013). The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585.
- Lysenko, V., Nagorny, C.L.F., Erdos, M.R., Wierup, N., Jonsson, A., Spégel, P., Buglioni, M., Saxena, R., Fax, M., Pulizzi, N., et al. (2009). Common variant in *MTNR1B* associated with increased risk of type 2 diabetes and impaired early insulin secretion. *Nat. Genet.* **41**, 82–88.
- MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., et al. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research* **45**, 896–901.
- Mahajan, A., Go, M.J., Zhang, W., Below, J.E., Gaulton, K.J., Ferreira, T., Horikoshi, M., Johnson, A.D., Ng, M.C.Y., Prokopenko, I., et al.; Diabetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium; Asian Genetic Epidemiology Network Type 2 Diabetes (AGEN-T2D) Consortium; South Asian Type 2 Diabetes (SAT2D) Consortium; Mexican American Type 2 Diabetes (MAT2D) Consortium; Type 2 Diabetes Genetic Exploration by Next-generation sequencing in multi-Ethnic Samples (T2D-GENES) Consortium (2014). Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat. Genet.* **46**, 234–244.
- Mahajan, A., Taliun, D., Thumer, M., Robertson, N.R., Torres, J.M., Rayner, N.W., Payne, A.J., Steinthorsdottir, V., Scott, R.A., Granup, N., et al. (2018). Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat. Genet.* **50**, 1505–1513.
- Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39**, 906–913.
- Marselli, L., Piron, A., Suleiman, M., Coll, M.L., Yi, X., Khamis, A., Carrat, G.R., Rutter, G.A., Buglioni, M., Giusti, L., et al. (2020). Persistent or Transient Human β Cell Dysfunction Induced by Metabolic Stress: Specific Signatures and Shared Gene Expression with Type 2 Diabetes. *Cell Rep.* **33**, 108466.
- Mayba, O., Gilbert, H.N., Liu, J., Haverty, P.M., Jhunjhunwala, S., Jiang, Z., Watanabe, C., and Zhang, Z. (2014). MBASED: allele-specific expression detection in cancer tissues and cell lines. *Genome Biol.* **15**, 405.
- McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M., Fuchsberger, C., Danecek, P., Sharp, K., et al.; Haplotype Reference Consortium (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283.
- McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Filcek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122.
- Meier, D.T., Entrup, L., Templin, A.T., Hogan, M.F., Samarasekera, T., Zraika, S., Boyko, E.J., and Kahn, S.E. (2015). Determination of Optimal Sample Size for Quantification of β -Cell Area, Amyloid Area and β -Cell Apoptosis in Isolated Islets. *J. Histochem. Cytochem.* **63**, 663–673.
- Mercader, J.M., and Florez, J.C. (2017). The Genetic Basis of Type 2 Diabetes in Hispanics and Latin Americans: Challenges and Opportunities. *Front. Public Health* **5**, 329.
- Miguel-Escalada, I., Bonàs-Guarch, S., Cebola, I., Ponsa-Cobas, J., Mendieta-Esteban, J., Atia, G., Javierre, B.M., Rolando, D.M.Y., Farabella, I., Morgan, C.C., et al. (2019). Human pancreatic islet three-dimensional chromatin architecture provides insights into the genetics of type 2 diabetes. *Nat. Genet.* **51**, 1137–1148.
- Newhall, R.A., Stouffer, S.A., Schuman, E.A., DeVinney, L.C., Star, S.A., and Williams, R.M. (1949). The American Soldier: Adjustment During Army Life. Volume I. Mississippi Val. Hist. Rev. **36**, 339.
- Ng, S.Y., Bogu, G.K., Soh, B.S., and Stanton, L.W. (2013). The long noncoding RNA *RMST* interacts with *SOX2* to regulate neurogenesis. *Mol. Cell* **51**, 349–359.
- O’Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44** (D1), D733–D745.
- Ongen, H., Bull, A., Brown, A.A., Dermizakis, E.T., and Delaneau, O. (2016). Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* **32**, 1479–1485.
- Osonoi, S., Ichinohe, H., Kudo, K., Yagihashi, S., and Mizukami, H. (2020). 2047-P: Possible Implication of Cyclin D2 in Beta-Cell Proliferation of Human Perinatal Islet. *Diabetes* **69** (Suppl 1).
- Pagès, H. (2015). *SNPlocs.Hsapiens.dbSNP142.GRCh37: SNP locations for Homo sapiens (dbSNP Build 142). R package version 0.99.5.* <https://bioconductor.org/packages/release/data/annotation/html/SNPlocs.Hsapiens.dbSNP142.GRCh37.html>.
- Pasquall, L., Gaulton, K.J., Rodríguez-Seguí, S.A., Mularoni, L., Miguel-Escalada, I., Akerman, I., Tena, J.J., Morán, I., Gómez-Marín, C., van de Bunt, M., et al. (2014). Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nat. Genet.* **46**, 136–143.
- Piñero, J., Bravo, A., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., García-García, J., Sanz, F., and Furlong, L.I. (2016). DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Research* **45**, 833–839.
- Piñero, J., Bravo, A., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., García-García, J., Sanz, F., and Furlong, L.I. (2017). DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.* **45** (D1), D833–D839.
- Plotly Technologies (2015). *Collaborative data science.* <https://plotly>.
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575.
- Ramos-Rodríguez, M., Raurell-Vila, H., Coll, M.L., Avelos, M.J., Subirana-Granés, M., Juan-Mateu, J., Norris, R., Turatainze, J.V., Nakayasu, E.S., Webb-Robertson, B.M., et al. (2019). The impact of proinflammatory cytokines on the β -cell regulatory landscape provides insights into the genetics of type 1 diabetes. *Nat. Genet.* **51**, 1588–1595.
- Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47.
- Rozowsky, J., Abyzov, A., Wang, J., Alves, P., Raha, D., Harmanci, A., Leng, J., Bjornson, R., Kong, Y., Kitabayashi, N., et al. (2011). AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol. Syst. Biol.* **7**, 522.
- Saeedi, P., Petersohn, I., Salpea, P., Malanda, B., Karuranga, S., Uhruw, N., Colaguri, S., Guariguata, L., Motala, A.A., Ogurtsova, K., et al. (2019). Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th edition. *Diabetes Res. Clin. Pract.* **157**, 107843.
- Saponaro, C., Acosta-Montalvo, A., Anguelova, L., Thévenet, J., Chiral, M., Pasquetti, G., Piron, A., Onop, M., Gmyr, V., Prehn, J., et al. (2020a). 1900-P: HNF1A Deficiency Leads to Perturbed Glucagon Secretion in Humans. *Diabetes* **69** (Suppl 1).
- Saponaro, C., Mühlemann, M., Acosta-Montalvo, A., Piron, A., Gmyr, V., Delalleau, N., Moerman, E., Thévenet, J., Pasquetti, G., Coddeville, A., et al. (2020b). Interindividual heterogeneity of SGLT2 expression and function in human pancreatic islets. *Diabetes* **69**, 902–914.

- Satya, R.V., Zavaljevski, N., and Reifman, J. (2012). A new strategy to reduce allelic bias in RNA-seq readmapping. *Nucleic Acids Res.* *40*, e127.
- Schmidt, E.M., Zhang, J., Zhou, W., Chen, J., Mohlke, K.L., Chen, Y.E., and Willer, C.J. (2015). GREGOR: evaluating global enrichment of trait-associated variants in epigenomic features using a systematic, data-driven approach. *Bioinformatics* *31*, 2601–2606.
- Scott, R.A., Scott, L.J., Mägi, R., Marullo, L., Gaulton, K.J., Kaakinen, M., Pervjakova, N., Pers, T.H., Johnson, A.D., Eicher, J.D., et al.; Diabetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium (2017). An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans. *Diabetes* *66*, 2888–2902.
- Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., Boutin, P., Vincent, D., Belisle, A., Hadjad, S., et al. (2007). A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* *445*, 881–885.
- Solimena, M., Schulte, A.M., Marselli, L., Ehehalt, F., Richter, D., Kleeberg, M., Mziat, H., Knoch, K.-P., Parnis, J., Bugliani, M., et al. (2018). Systems biology of the IMIDIA biobank from organ donors and pancreatectomized patients defines a novel transcriptomic signature of islets from individuals with type 2 diabetes. *Diabetologia* *61*, 641–657.
- Spracklen, C.N., Horikoshi, M., Kim, Y.J., Lin, K., Bragg, F., Moon, S., Suzuki, K., Tam, C.H.T., Tabara, Y., Kwak, S.H., et al. (2020). Identification of type 2 diabetes loci in 433,540 East Asian individuals. *Nature* *582*, 240–245.
- Stegle, O., Parts, L., Durbin, R., and Winn, J. (2010). A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comput. Biol.* *6*, e1000770.
- Stegle, O., Parts, L., Pilipari, M., Winn, J., and Durbin, R. (2012). Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* *7*, 500–507.
- Steinthorsdottir, V., Thorleifsson, G., Sulem, P., Helgason, H., Grarup, N., Sigurdsson, A., Helgadóttir, H.T., Johannsdóttir, H., Magnusson, O.T., Gudjonsson, S.A., et al. (2014). Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. *Nat. Genet.* *46*, 294–298.
- Stitzel, M.L., Sethupathy, P., Pearson, D.S., Chines, P.S., Song, L., Erdos, M.R., Welch, R., Parker, S.C.J., Boyle, A.P., Scott, L.J., et al.; NISC Comparative Sequencing Program (2010). Global epigenomic analysis of primary human pancreatic islets provides insights into type 2 diabetes susceptibility loci. *Cell Metab.* *12*, 443–455.
- Supek, F., Bošnjak, M., Škunca, N., and Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS ONE* *6*, e21800.
- The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature* *526*, 68–74.
- The Gene Ontology Consortium. (2017). Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res.*
- Thomas, P.D., Campbell, M.J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A., and Narechania, A. (2003). PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res.* *13*, 2129–2141.
- Thomas, P.D., Kejariwal, A., Guo, N., Mi, H., Campbell, M.J., Muruganujan, A., and Lazareva-Ulitsky, B. (2006). Applications for protein sequence-function evolution data: mRNA/protein expression analysis and coding SNP scoring tools. *Nucleic Acids Res.* *34*, W645–W650.
- Thumer, M., van de Bunt, M., Torres, J.M., Mahajan, A., Nylander, V., Bennett, A.J., Gaulton, K.J., Barrett, A., Burrows, C., Bell, C.G., et al. (2018). Integration of human pancreatic islet genomic data refines regulatory mechanisms at type 2 diabetes susceptibility loci. *eLife* *7*, e31977.
- van de Bunt, M., Manning Fox, J.E., Dai, X., Barrett, A., Grey, C., Li, L., Bennett, A.J., Johnson, P.R., Rajotte, R.V., Gaulton, K.J., et al. (2015). Transcript Expression Data from Human Islets Links Regulatory Signals from Genome-Wide Association Studies for Type 2 Diabetes and Glycemic Traits to Their Downstream Effectors. *PLoS Genet.* *11*, e1005694.
- Vifuela, A., Varshney, A., van de Bunt, M., Prasad, R.B., Asplund, O., Bennett, A., Boehnke, M., Brown, A.A., Erdos, M.R., Fadista, J., et al. (2020). Genetic variant effects on gene expression in human pancreatic islets and their implications for T2D. *Nat. Commun.* *11*, 4912.
- Vujkovic, M., Keaton, J.M., Lynch, J.A., Miller, D.R., Zhou, J., Tchandjieu, C., Huffman, J.E., Assimes, T.L., Lorenz, K., Zhu, X., et al.; HPA Consortium; Regeneron Genetics Center; VA Million Veteran Program (2020). Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis. *Nat. Genet.* *52*, 680–691.
- Wakefield, J. (2009). Bayes factors for genome-wide association studies: comparison with P-values. *Genet. Epidemiol.* *33*, 79–86.
- Walter, K., Min, J.L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., Perry, J.R., Xu, C., Futema, M., Lawson, D., et al.; UK10K Consortium (2015). The UK10K project identifies rare variants in health and disease. *Nature* *526*, 82–90.
- Wild, S., Roglic, G., Green, A., Sicree, R., and King, H. (2004). Global prevalence of diabetes: estimates for the year 2000 and projections for 2030. *Diabetes Care* *27*, 1047–1053.
- Willer, C.J., Li, Y., and Abecasis, G.R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* *26*, 2190–2191.
- Wu, D., Gu, J., and Zhang, M.Q. (2013). FastDMA: An Infinium HumanMethylation450 Beadchip Analyzer. *PLoS ONE* *8*, e74275.
- Yang, Y.H., Dudoit, S., Luu, P., Lin, D.M., Peng, V., Ngai, J., and Speed, T.P. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* *30*, e15.
- Yates, A.D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M.R., Armean, I.M., Azov, A.G., Bennett, R., et al. (2020). Ensembl 2020. *Nucleic Acids Res.* *48* (D1), D682–D688.



STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
RNA-seq and genotyping array data (in this paper)	Marselli et al., 2020	EGA: EGAS00001005535
RNA-seq and genotyping array data	Fadista et al., 2014	GEO:GSE50244
RNA-seq and genotyping array data	van de Bunt et al., 2015	EGA:EGAD00001001601
RNA-seq data	Cnop et al., 2014	GEO:GSE53949
RNA-seq and genotyping array data	Akerman et al., 2017	EGA:EGAS00001002865
RNA-seq and genotyping array data	Miguel-Escalada et al., 2019; data not shown	EGA pending accession number
Expression array	Solimena et al., 2018	GEO:GSE76896
DNA-methylation	Hall et al., 2014	EGA:EGAD00001003946
Bisulphite sequencing	Thurner et al., 2018	EGA:EGAD00001003947
Cohesin	Miguel-Escalada et al., 2019	EGA:EGAD00001005203
Mediator	Miguel-Escalada et al., 2019	EGA:EGAD00001005203
H3K27ac	Miguel-Escalada et al., 2019	EGA:EGAD00001005203
ATAC-seq	Miguel-Escalada et al., 2019	EGA:EGAD00001005203
Islet regulome annotations, ChIP-seq and ATAC-seq processed files	Miguel-Escalada et al., 2019	EGA:EGAD00001005203
Pancreatic islet enhancer clusters	Pasquali et al., 2014	
H3K4me1	Pasquali et al., 2014	
Long non-coding RNAs (lncRNAs) annotation	Akerman et al., 2017	
Pancreatic islet open chromatin DNase	Stitzel et al., 2010	ENCODE (2012-2016) Open Chromatine Dnase
Pancreatic islet open chromatin DNase	Gaulton et al., 2010	ENCODE (2012-2016) Open Chromatine Dnase
Glycemic traits data	MAGIC investigators (http://magicinvestigators.org/); members of MAGIC are provided in Appendix S1	
70KforT2D GWAS meta-analysis summary statistics	Bonàs-Guarch et al., 2018	http://cg.bsc.es/70kfort2d/
DIAGRAM 1000G GWAS meta-analysis Stage 1 Summary statistics	Scott et al., 2017	https://diagram-consortium.org/downloads.html
DIAGRAM Trans-ethnic T2D GWAS meta-analysis	Mahajan et al., 2014	https://diagram-consortium.org/downloads.html
DIAMANTE T2D GWAS meta-analysis	Mahajan et al., 2018	https://diagram-consortium.org/downloads.html
GTEx Analysis V7 - Transcript TPMs	GTEx Portal	https://www.gtexportal.org/home/
FastDMA probe full annotation	Wu et al., 2013	http://bioinfo.au.tsinghua.edu.cn/member/jgu/fastdma/
Gene Ontology	The Gene Ontology Consortium, 2017	http://geneontology.org/
Reactome	Reactome Pathway database	https://reactome.org/download-data/
DisGeNET, May 2017	Piñero et al., 2016	https://www.disgenet.org/
GWAS Catalog version 1.0 release 2021-06-08	MacArthur et al., 2017	https://www.ebi.ac.uk/gwas/downloads
Ensembl Variant Effect Predictor version 87.27	McLaren et al., 2016	https://m.ensembl.org/info/data/ftp/index.html
RefSeq BUILD.37.3	O'Leary et al., 2016	ftp://ftp.ncbi.nlm.nih.gov/genomes/Homo_sapiens/ARCHIVE/BUILD.37.3

(Continued on next page)



Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Gencode v23 lift 37 annotation	Frankish et al., 2019	ftp://ftp.ebi.ac.uk/pub/databases/genocode/Genocode_human/release_23/GRCh37_mapping/genocode.v23lift37_annotation.gtf.gz
gnomAD version 2.0.2	gnomAD database	https://gnomad.broadinstitute.org/downloads

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Miriam Cnop (mcnop@ulb.ac.be)

Materials availability

This study did not generate new unique reagents.

Data and code availability

RNA-seq and genotyping array data from PISA cohort Sequence data have been deposited at the European Genome-phenome Archive (EGA), which is hosted by the EBI and the CRG, under accession number EGAS00001005535.

Further information about EGA can be found on <https://ega-archive.org> “The European Genome-phenome Archive of human data consented for biomedical research” (<https://www.nature.com/ng/journal/v47/n7/full/ng.3312.html>).

RNA-seq and genotyping array data from CRG cohort should be requested through Miguel-Escalada et al. (2019) and coauthor Goutham Atla.

The eQTL and cASE results are available for browsing at TIGER (<http://tiger.bsc.es>), and the full summary statistics are available for download.

Source data and publicly available resources used for this study supporting all findings are detailed in the [key resources table](#).

The cASE code is available through https://github.com/moran-BSC/TIGER_cASE.

Any additional information required to reanalyze the data reported in this work paper is available from the Lead Contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Islet sample collection and genotyping

TIGER data consist of 514 RNA-seq and 485 genotyped array data of deidentified cadaveric human pancreatic islet samples from five research centers: 1) Centre for Genomic Regulation, 2) Lund University Diabetes Centre, 3) University of Oxford/University of Alberta, 4) Department of Endocrinology and Metabolism, University of Pisa and 5) ULB Center for Diabetes Research, Université Libre de Bruxelles (Table S1). For the latter two centers, islets are prepared from the body and tail of the pancreas.

Centre for Genomic Regulation (CRG)

The DNA of 127 CRG samples was isolated, sequenced, and genotyped using Illumina’s Human OmniExpress 12 v1 and 2.5-8 v1.1 chips, as described in Miguel-Escalada et al. (2019). Genotype array was done in 125 samples with Illumina’s Genome Studio software providing information on a total of 624k SNPs.

Lund University Diabetes Centre (Lund)

The DNA of 89 Lund samples from cadaver donors of European ancestry provided by the Nordic Islet Transplantation Programme was isolated as described in Fadista et al. (2014). The samples were genotyped using Illumina’s HumanOmniExpress 12v1 C chips passing standard quality control metrics providing information on a total of 609k SNPs.

University of Oxford/University of Alberta (Oxford)

The DNA of 118 Oxford samples was isolated from either spleen or the exocrine fraction of the islet isolation using the Tissue DNA Purification Kit. When no other tissue was available, DNA was extracted from human islets using the Trizol fraction remaining after extraction of RNA as described in van de Bunt et al. (2015). The samples were genotyped using Illumina’s Human Omni 2.5 exome array following the Illumina Infinium protocol providing information on a total of 2.5M SNPs.



University of Pisa (Pisa)

The DNA of 154 Pisa samples was isolated according to previously described in [Marselli et al. \(2020\)](#) and sequenced. Genotype calling was done in 153 samples with Illumina's Human Omni 2.5 exome array providing information on a total of 2.6M SNPs.

ULB Center for Diabetes Research (ULB)

The 43 ULB samples were isolated in Pisa using collagenase digestion and density gradient purification from beating-heart organ donors with no medical history of diabetes or metabolic disorders. Following islet shipment to Brussels, mRNA was extracted and processed following the RNeasy QIAGEN protocol as described in [Cnop et al. \(2014\)](#).

METHOD DETAILS

Genotyping quality control

PLINK v1.9 ([Purcell et al., 2007](#)) was used to do standard quality control of the genotype data, at the variant and sample level ([Bonàs-Guarch et al., 2018](#)). At the variant level, we discarded rare variants (Minor Allele Frequency MAF < 0.01) and applied Hardy-Weinberg equilibrium test filtering ($p \leq 1 \times 10^{-6}$) ([Graffelman, 2015](#); [Graffelman and Camarena, 2008](#)). Further, we filtered the variants below a missingness threshold of 0.05. At the sample level, we discarded samples presenting a gender discordance between the reported gender in the metadata and the genetic sex, as well as the subjects with at least a 3rd degree of relatedness, those below a missingness threshold of 0.02 and, finally, individuals not clustering within the 4 standard deviations of the first four principal components from the multidimensional scale analysis. The ancestry of the individuals was assessed by principal components analysis comparisons with phase3 1000 Genomes Project populations ([The 1000 Genomes Project Consortium, 2015](#)).

After QC this resulted in a total of: 1) 103 individuals, 559,083 SNPs in the CRG cohort, 2) 88 individuals, 596,273 SNPs in the Lund cohort, 3) 102 individuals, 1,487,651 SNPs in the Oxford cohort and 4) 144 individuals, 1,542,765 SNPs in the Pisa cohort.

Genotype phasing and imputation

The autosomal genotypes were phased with Eagle3 ([Loh et al., 2016a, 2016b](#)) using the Human Reference Consortium Project reference panel ([McCarthy et al., 2016](#)). The X chromosome was phased without reference panel with SHAPET ([Delaneau et al., 2011](#)). Then, GUIDANCE ([Guindo-Martinez et al., 2021](#)) integrating IMPUTE2 ([Marchini et al., 2007](#)) was used for imputation, using 4 reference panels: the 1000 Genomes Project phase 3 ([The 1000 Genomes Project Consortium, 2015](#)), the Genome of the Netherlands Project ([Boomsma et al., 2014](#)), the Haplotype Reference Consortium Project ([McCarthy et al., 2016](#)) and the UK10K Project ([Walter et al., 2015](#)), with an IMPUTE2 info score threshold of ≥ 0.7 . This resulted in a total of 13.7-16.3M SNPs for each cohort separately, that were merged considering the best info score obtained across all panels, resulting in 22,983,795 genotyped and imputed genetic variants with MAF > 0.001.

RNA-seq read mapping

RNA from 514 human donor islet samples was isolated and purified, and was used to construct RNA-seq libraries. These bulk RNA-seq assays generated a total of > 72 billion pair-ended fragments of 75, 76, 100, 101, 125 bp read lengths.

To perform eQTL analysis, we aligned all samples against the transcriptome reference genome v2.3lif37 ([Frankish et al., 2019](#)) with STAR v2.4.0 ([Dobin et al., 2013](#)), using

- `--paired-end -p 8`

An alternative mapping strategy was used for RNA-seq read mapping to be used for cASE. Given that the standard reference genome contains only one allele in polymorphic sites, standard RNA-seq read mapping can produce reference-biased alignments, leading to false positives in the study of ASE. To align RNA-seq datasets in an allele unbiased manner, two modified reference genomes were built, defined as a 'masked' and an 'enhanced' genome. The 'masked' reference genome was built by substituting with an 'N' the nucleotide position of each common SNP in dbSNP142 ([Pagès, 2015](#)) (MAF > 1%), using the `vcf2diploid.jar` ([Rozowsky et al., 2011](#)) tool. To construct the 'enhanced' reference genome, we modified the scripts developed by [Satya et al. \(2012\)](#) to accommodate RNA-seq reads, which added artificial contigs to the reference genome containing all possible SNP allele combinations. For this step, we used the subset of 4M common SNPs located within gene coordinates in the Ensembl ([Yates et al., 2020](#)), RefSeq ([O'Leary et al., 2016](#)) and UCSC ([Haeussler et al., 2019](#)) annotations, or within previously identified human islet lncRNAs ([Akerman et al., 2017](#)) ([Figure S5](#)).

STAR v2.2.0 ([Dobin et al., 2013](#)) was used to align the RNA-seq datasets against the masked genome, using

- `--outFilterMultimapNmax 1--outFilterMismatchNmax 10`
- `--outSAMstrandField intronMotif--outSAMattributes All`

in order to allow up to 10% of nucleotide mismatches, suppress multimapped reads, and make the output compatible with downstream software. Bowtie v2.0.5 ([Langmead et al., 2009](#)) was used to align the RNA-seq data against the enhanced genome, using

- `--n-cell L,0,0.03--score-min C,-14,0 -N 1 -X 50000`

to allow up to 3 nucleotide mismatches evenly distributed within the read, and long range read pairs. Bowtie2 (Langmead and Salzberg, 2012) was chosen because it does not map the RNA-seq spliced reads, (only the reference allele-containing spliced sequences were present in the enhanced genome) which prevents the generation of allelic alignment bias.

After mapping the RNA-seq datasets to the two modified reference genomes, the outputs of both alignments were combined into one non-redundant set of reads, using the read merging C++ scripts available in our github repository (https://github.com/Imoran-BSC/TIGER_cASE, scripts 02 and 03). Reads that aligned to the same genomic positions by both methods were kept, as well as reads mapped only by one of the two methods. In addition, all reads that mapped partially to intronic regions were discarded. The resulting set of reads was named 'unbiased alignment' (Figure S5A). This method successfully eliminated alignment bias in heterozygous positions (Figure S5B), and mapped 86.2% of all RNA-seq reads. When comparing this alignment with one using the standard reference genome and STAR v2.2.0 using a subset of the samples, we recovered an extra 8.5% more reads using the unbiased alignment method (Figure S5C).

Sample concordance verification between genotype and gene expression

To avoid mislabeled samples leading to mismatching errors between genotype-phenotype samples, and to discard samples with poor quality or possible contamination, we used verifyBamID v1.1.3 (Jun et al., 2012) with "--best," applied to the RNA-seq alignments sorted and indexed with samtools v1.1 (Li et al., 2009), and comparing with their genotypes. After these steps, 404 samples with good quality genotype and RNA-seq data and concordance remained for further analysis.

TIGER web portal development

The TIGER web portal (<http://tiger.bsc.es>) is the comprehensive integration in an ElasticSearch v1.4.4 database of a) T2D GWAS variants identified in 70KforT2D (Bonàs-Guarch et al., 2018), diagram DIAMANTE (Mahajan et al., 2018), diagram Trans-ethnic (Mahajan et al., 2014), diagram 1000G (Scott et al., 2017) T2D meta-analyses or included in the GWAS Catalog v1 release 2021-06-08 (Buniello et al., 2019), b) variant annotation and characterization through Variant Effect Predictor v87.27 (McLaren et al., 2016) and Gnomad v2.0.2 (Karczewski et al., 2020), c) epigenomic marks from islet DNA-methylation sites (Hall et al., 2014; Turner et al., 2018), chromatin accessibility (Dunham et al., 2012; Gaulton et al., 2010; Stitzel et al., 2010) and CHIP-seq profiles (Miguel-Escalada et al., 2019), d) annotation from Gene Ontology (Ashburner et al., 2000; The Gene Ontology Consortium, 2017), lncRNAs (Akerman et al., 2017) and islet regulome (Miguel-Escalada et al., 2019; Pasquali et al., 2014) in a publicly available platform. Genes are referenced to Gencode annotation v23 lift 37 (Frankish et al., 2019) and RefSeq BUILD.37.3 (O'Leary et al., 2016) and enriched with DisGeNET (Piñero et al., 2017) (May 2017) and Reactome Pathway (Jassal et al., 2020) database information. It contains results on gene expression integrating the results of a) gene expression from normalized islet RNA-seq counts, microarrays (Solimena et al., 2018), and the Genotype-Tissue Expression database (GTEx) (Lonsdale et al., 2013), and b) computed eQTL and cASE.

The portal was built upon [ICGC software codebase], the front-end coded in angular v1.5.7 with embedded biodalliance v1.4.4 genomic browser (Down et al., 2011), plotly v1.54.1 (Plotly Technologies, 2015) and highcharts libraries and the back-end coded in Java.

QUANTIFICATION AND STATISTICAL ANALYSIS

eQTL analysis

The *cis*-eQTL analysis of 404 human pancreatic islets for which both RNA-seq and genotyping data remained after QC was performed by cohort with fastQTL v2.0 tool (Ongen et al., 2016). The analysis was run for regions one million base pairs up- or downstream of the transcription start site of each gene using *gencode.v23lift37* (Frankish et al., 2019) version. For each cohort, we corrected for known covariates (age, sex and BMI), 7 genomic ancestry principal components, and 15 PEER v1.3 (Stegle et al., 2010) factors in order to account for hidden confounding factors. For the X chromosome, we used 5 PEER factors and 4 genomic ancestry principal components and the *cis*-eQTL analysis was performed stratified by sex and combined. The full command for fastQTL is

```
fastQTL-log 'chr1.log'-vcf 'chr1.bcf'-bed 'rsem.bed' -C 'covariates.tsv'-threshold '0.01'-out 'chr1.fastQTL.gz'
```

Age and BMI missing metadata were imputed using the cohort mean.

The by-cohort fastQTL (Ongen et al., 2016) results were then meta-analyzed with METAL (Willer et al., 2010) using the sample size strategy and computing heterogeneity. For the X chromosome, the meta-analysis was run over the 4 cohorts for both sexes together and over the 8 eQTL analysis (4 cohorts, 2 sexes). The full configuration files for METAL are given by:

```
SEPARATOR WHITESPACE
MARKER ensg.snp
ALLELE a0 a1
EFFECT slope
PVALUE pval
```



```
WEIGHT N
PROCESS cohort_CRG
PROCESS cohort_OXFORD
PROCESS cohort_LUND
PROCESS cohort_PISA
OUTFILE metal .tsv
ANALYZE HETEROGENEITY
QUIT
```

Identifying variant regulatory enrichments using GREGOR

To test the eQTL and cASE variants for enrichment in islet regulatory overlaps, we used the Genomic Regulatory Elements and Gwas Overlap algoRithm (GREGOR) (Schmidt et al., 2015), designed to calculate such enrichment while controlling for linkage-disequilibrium between variants, MAF and distance to nearest gene. We used the 1% and 5% FDR set of significant eQTL variants, after selecting them by linkage disequilibrium < 0.2 using PLINKv1.9 (Purcell et al., 2007) with “-indep-pairwise 100k 5 0.2”. We tested enrichment against a set of human islet regulatory regions, including gene promoters, enhancers, and open-chromatin derived from ChIP-seq experiments in human islets (Figures 2C and S1) (Miguel-Escalada et al., 2019). Specifically, we used an R^2 threshold of 0.99, a window size of 1,000,000, a `min_neighbor_num` of 500, and European (EUR) as the population.

Comparison of TIGER eQTLs with the GTEx and InsPIRE datasets

To assess the degree of concordance between the TIGER significant eQTLs and those reported in the GTEx v8 dataset (Aguet et al., 2020), we searched for exact variant-target gene matches among the dataset of significant eQTLs in all 54 GTEx tissues. To analyze the overlap of eQTLs with low-frequency variants, we repeated the analysis, but first filtered the TIGER and GTEx eQTLs to include only those with variants with a MAF < 0.05 in the EUR population of the 1000 genomes phase-3 dataset (The 1000 Genomes Project Consortium, 2015).

To obtain a relevant comparison with the InsPIRE (Viñuela et al., 2020) dataset, we first applied the same multiple-testing correction method used in this study to the full nominal p values of the InsPIRE dataset. The Benjamini-Hochberg corrections for 1 and 5% FDR resulted in the nominal p -value thresholds of $p = 8.55 \times 10^{-6}$ and $p = 6.2 \times 10^{-4}$, corresponding to 974,435 and 1,408,891 significant eQTLs. Two eQTLs were considered significant by both methods if they were detected at $< 5\%$ FDR in both studies, and had an exact match in both variant and target gene. The low-frequency variant eQTLs were determined as described above.

Colocalization analysis

COLOC 4.0 (Giambartolomei et al., 2014) R package was used for the colocalization analysis of *cis*-eQTL and T2D GWAS. We used the `coloc.abf` method which implements a variation of the Approximate Bayes Factor computations (Wakefield, 2009). The `coloc.abf` function was called with two R lists, one for the eQTL and one for the GWAS:

```
list(pvalues = ..., N = ..., MAF = ..., snp = ..., type = "quant")
```

with a vector of p -values, N the sample size, MAF the minor allele frequency and `snp` the rsid of the variant.

In order to select regions for colocalization analyses, we selected genes associated with at least one significant eQTL SNP which had been previously reported as a GWAS lead variant (Bonàs-Guarch et al., 2018; Mahajan et al., 2018; Vujkovic et al., 2020). The significant eQTL SNPs were determined based on a 0.05 threshold Benjamini-Hochberg FDR (Benjamini and Hochberg, 1995). Similarly, we used the p -values of the cASE analysis to perform colocalization, considering loci with an at least 5% FDR significant signal. The colocalization was run over regions ranging from one million base pairs downstream to one million upstream of the *cis*-regulatory target gene transcription start site.

The colocalization plots were generated by the `locuscompare` R package v1.0.0 (Liu et al., 2019) (Data S1 and S2).

Generation of an unbiased set of ASE reporter variants

To identify loci under mappability related allelic biases, a C++ script available in the github repository (https://github.com/Imoran-BSC/TIGER_cASE, script 01) was used to generate all possible reads containing both alleles of all possible reporter SNPs. A splice junction database was created using the Ensembl (Yates et al., 2020), RefSeq (O'Leary et al., 2016), UCSC (Haeussler et al., 2019) and human islet lncRNA (Akerman et al., 2017) gene annotations, to take splice junctions into account.

The resulting dataset, consisting of 240M artificial reads, was aligned using the unbiased mapping strategy described above, and the allelic ratios (i.e., the percentage of reference-allele carrying reads) were quantified. Since the same number of reads were purposely generated carrying both alleles, any observed allelic imbalance would derive exclusively from mapping biases. SNPs whose allelic ratio was not between 49%–51% were blacklisted. Additionally, all SNPs located within 100 bps of a common or low-frequency indel present in dbSNP142 (Pagès, 2015) were also blacklisted.

The remaining curated set of 3.97M SNPs were used as bona-fide SNPs for reporting ASE.

Identification of ASE

The number of reads containing the reference and alternate alleles RNA-seq reads overlapping each reporter SNP were quantified using the `mpileup` command of `samtools` v1.1 (Li et al., 2009), with the flags “-A -B -d 20000”, and the `ComputePileupFreqs.pl` script (Satya et al., 2012). Sample-specific ASE was assessed calculating the allelic ratio, i.e., the fraction of reads containing the reference allele over the total number of reads. We selected the set of SNPs with at least 3 heterozygous samples with ≥ 15 RNA-seq reads (of which ≥ 10 non-clonal), resulting in a set of $> 170k$ informative reporter SNPs.

A binomial test (Bernoulli, 1899) was used to assess the significance of ASE for all reporter SNPs, using the number of reads carrying the reference and alternate alleles. To account for any possible remaining alignment bias in the datasets, the median allelic ratio for each possible bi-allelic SNP (AC, AG, AT, CG, CT, GT) across the genome was calculated and used as null, instead of the theoretical 50%. Similarly, the allelic ratios were proportionally adjusted using the sample and nucleotide-pair specific median value.

The resulting p -values were used to calculate a sample-specific 1% and 5% FDR Benjamini-Hochberg (Benjamini and Hochberg, 1995) thresholds, to correct for multiple testing.

Assessing cASE using Stouffer's Z-score

To assess cASE in a given heterozygous variant in many independent samples, the Stouffer's Z-score (Newhall et al., 1949) method was used. This method combines independently obtained p -values into a Z statistic, which increases in absolute value with significance. The method allows for weighting of independent p -values and, additionally, it accounts for a positive or negative direction in the magnitude associated with the p -values. Thus, this method allows to differentiate between significant reference and alternate reporter variants, as well as providing a way to account for the variance inherent to differing numbers of informative RNA-seq reads in each reporter.

For each reporter, a Z-score was calculated as follows:

$$Z = \frac{\sum w_i Z_i}{\sqrt{\sum w_i^2}}$$

where w_i was the total read coverage of sample i , and Z_i was the transformed binomial p -value p_i :

$$Z_i = \pm \theta^{-1} \left(1 - \frac{p_i}{2} \right)$$

where the sign was positive if the value of the allelic ratio was $> 50\%$, zero if exactly 50%, and negative otherwise, and θ^{-1} was the inverse of the standard normal cumulative distribution function, calculated using the `qnorm` function in R. A threshold of 10^{-15} was imposed as the minimum possible binomial p -value, in order to prevent single events with very significant p -values from dominating the Z-score value, while still maintaining their relevance. Therefore, Stouffer's Z-score (Newhall et al., 1949) method accounted for consistency in the overall reference or alternate direction of the allelic bias across samples, and considered all p -values into account, regardless of their sample-specific significance.

Z-scores were only calculated if the reporter SNP was heterozygous in 3 or more samples, and only samples with a read coverage of ≥ 15 RNA-seq reads, of which ≥ 10 non-clonal, were used in the calculation.

Assessing the significance of cASE Z-scores

To assess the significance of the obtained Z-scores, we performed 1,000 permutations of the reference/alternate read counts between heterozygous SNPs, and calculated their binomial p -values and resulting control Z-scores (https://github.com/imoran-BSC/TIGER_cASE, script 04). To account for the differences in gene expression, all reporter SNPs were distributed in 5 bins: one containing all SNPs with a median coverage of 0 reads, and 4 more bins containing the remaining SNPs according to their read coverage quartile, and the read counts of heterozygous SNPs were only shuffled within their bins. By permuting only the values of the heterozygous SNPs while keeping the reference and alternate homozygous values invariant, the distribution of the number of samples in heterozygosity for each SNP was kept constant.

The resulting null distribution of Z-scores was therefore attributable only to stochasticity, and so for each empiric Z-score, a p -value was calculated from this null distribution. The Benjamini-Hochberg method (Benjamini and Hochberg, 1995) was then used to obtain q -values from these p -values and thus correct for multiple testing.

Regulatory enrichment of cASE significant genes

To calculate these regulatory enrichments, we first generated a null distribution of control genes that were non-significant for cASE but had similar expression levels. First, we separated the cASE significant genes in 4 bins of expression, and randomly selected the same number of non-significant genes of the same expression quartile, 1,000 times. We then calculated, in the 1% and 5% FDR cASE genes and in each of the 1,000 control sets, the proportion of genes that were in the islet-specifically expressed genes list (Miguel-Escalada et al., 2019) (Figure 4C, left). The same procedure was performed to calculate the enrichment for proximity to islet enhancers, by calculating the proportion of genes located at less than 25kb from islet enhancers (Miguel-Escalada et al., 2019). The p -values were obtained by approximating these permuted control distributions as Gaussian distributions and deriving a p -value using the `pnorm` R function.



Gene ontology analyses and islet-specific expression

Gene ontology terms in the analyses of eQTL and cASE genes were obtained using the PANTHER (Protein ANALYSIS THrough Evolutionary Relationships) (Thomas et al., 2003, 2006) classification system.

For eQTL, we analyzed all 5% FDR significant genes versus a background list of all genes expressed in islets (Figure S3), and the list of TIGER exclusive eQTL genes versus a background of all eQTL genes shared with GTEx (Figure 2E).

For cASE, we studied 5% FDR cASE genes versus a background dataset of all genes for which the calculated cASE was non-significant (Figure 4D). The visualization of the syntactic terms was obtained using the REVIGO web tool (Supek et al., 2011).

Identifying candidate SNPs putatively leading to cASE

We aimed to characterize the set of SNPs putatively causal of cASE (referred to as ‘candidate SNPs’). To that end, we first identified all variant pairs consisting of a cASE-significant reporter and a candidate variant, as long as both were located within the same topologically associating domain (TAD) (Dixon et al., 2012), plus a boundary leeway of ± 200 kbs. Then, we separated the samples using the candidate variant genotype in two groups: those heterozygous (Het), and those homozygous (Hom). Finally, we calculated the reporter Z-score of both sample groups, and selected the candidate variants with significant Z-scores for the Het individuals, which were also non-significant for the Homs (https://github.com/moran-BSC/TIGER_cASE, script). The underlying hypothesis was that if the candidate variant was homozygous, it was unlikely to be causal.

Putative causal variants were also interrogated for the set of non-cASE significant reporter variants, following the same procedure described above. This produced an additional 1,247 genes that reached cASE significance only after being considered with these putative causal variants.

Scaling human islet gene expression values to allow comparisons with the GTEx expression datasets in TIGER

The RNA-seq expression of human islet samples was measured with RSEM v1.3.0 (Li and Dewey, 2011) in 60,261 transcripts from Gencode database (v23lift37 annotation) (Frankish et al., 2019) using STAR v2.5.3.a (Dobin et al., 2013) and BOWTIE v2.3.2 (Langmead and Salzberg, 2012) hg19 aligned-reads as follows:

```
STAR-runMode genomeGenerate-genomeFastaFiles GRCh37.primary_assembly.genome.fa-sjdbGTFfile gencode.v23lift37.
annotation.gtf
rsem-prepare-reference-gtf gencode.v23lift37.annotation.gtf-bowtie2 GRCh37.primary_assembly.genome.fa
rsem-calculate-expression-paired-end-star-paired-end -p 8
```

We obtained measures of raw counts, counts normalized by transcript length (TPM - transcripts per million) and fragment length (FPKM - fragments per kilobase). The batch effects and covariate differences between samples captured in the TPM measures were removed with limma removeBatchEffect function (Ritchie et al., 2015), using the log10 normalized expression of the genes that were expressed in at least 80% of human islet samples. The results of this normalization were evaluated with Spearman correlation, ensuring that there was a correlation above 0.8 between all the samples independently of the cohort after correction.

TPM expression datasets from the 54 tissues available in GTEx (Lonsdale et al., 2013) (20 samples per tissue) were collected, and a decile distribution analysis was performed excluding genes from GTEx samples that miss expression in at least 50% of the samples. Then, TIGER islet expression was scaled to fit these measures according to the following criteria:

- (1) Each GTEx decile bin $[D_{G_i}, D_{G_{j+1}}]$ has TPM values in $[T_{G_i}, T_{G_{j+1}}]$, thus the corresponding decile straight will be: $y_G = (T_{G_{j+1}} - T_{G_i})x + T_{G_i}$.
- (2) Each pancreatic islet decile bin $[D_{PI_i}, D_{PI_{i+1}}]$ has TPM values in $[T_{PI_i}, T_{PI_{i+1}}]$, thus the corresponding decile straight will be: $y_{PI} = (T_{PI_{i+1}} - T_{PI_i})x + T_{PI_i}$.

From Equation (2) one can derive: $x = \frac{y_{PI} - T_{PI_i}}{T_{PI_{i+1}} - T_{PI_i}}$ (3) thus, allowing the relation between the TPM pancreatic islet values y_{PI} and the TPM GTEx values y_G by replacing (3) in (1): $y_G = \left(\frac{T_{G_{j+1}} - T_{G_i}}{T_{PI_{i+1}} - T_{PI_i}} \right) y_{PI} - T_{PI_i} \left(\frac{T_{G_{j+1}} - T_{G_i}}{T_{PI_{i+1}} - T_{PI_i}} \right) + T_{G_i}$ the scaling factor.

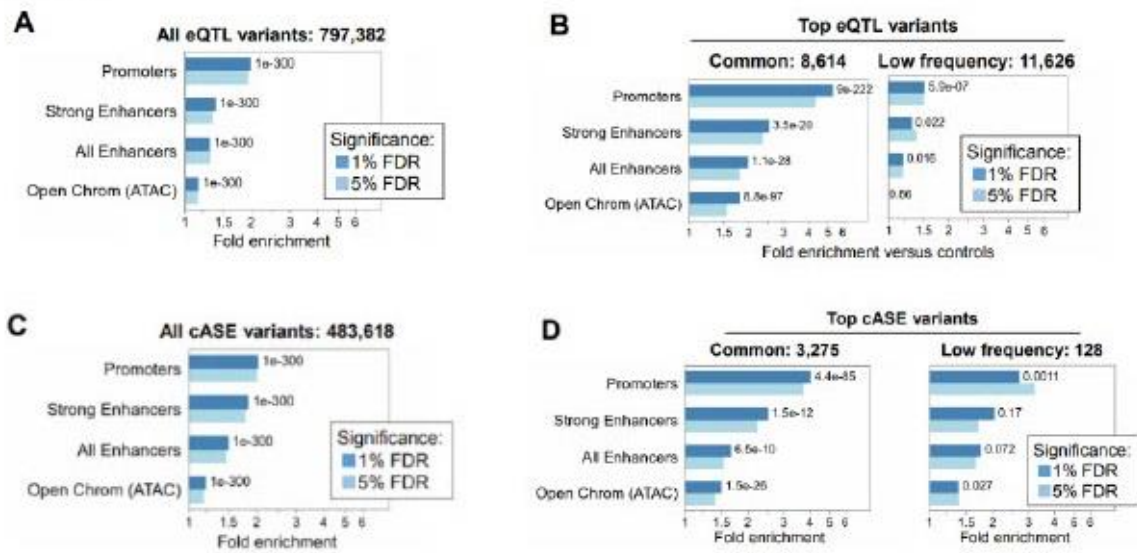
Supplemental information

TIGER: The gene expression regulatory variation landscape of human pancreatic islets

Lorena Alonso, Anthony Piron, Ignasi Morán, Marta Guindo-Martínez, Sílvia Bonàs-Guarch, Goutham Átla, Irene Miguel-Escalada, Romina Royo, Montserrat Puiggròs, Xavier Garcia-Hurtado, Mara Suleiman, Lorella Marselli, Jonathan L.S. Esguerra, Jean-Valéry Turatsinze, Jason M. Torres, Vibe Nylander, Ji Chen, Lena Eliasson, Matthieu Defrance, Ramon Amela, MAGIC, Hindrik Mulder, Anna L. Gloyn, Leif Groop, Piero Marchetti, Decio L. Eizirik, Jorge Ferrer, Josep M. Mercader, Miriam Cnop, and David Torrents

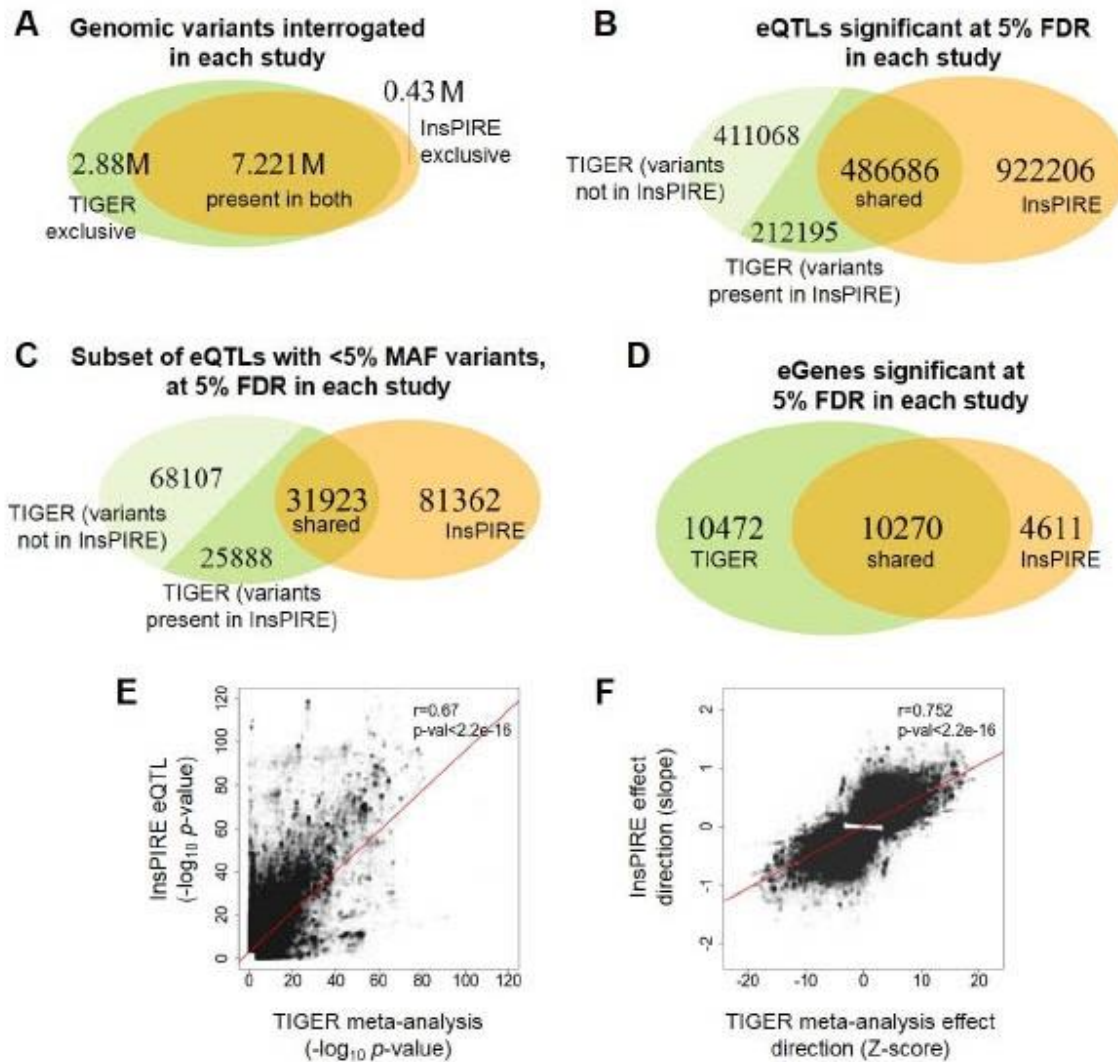
Supplemental Figures

Figure S1.



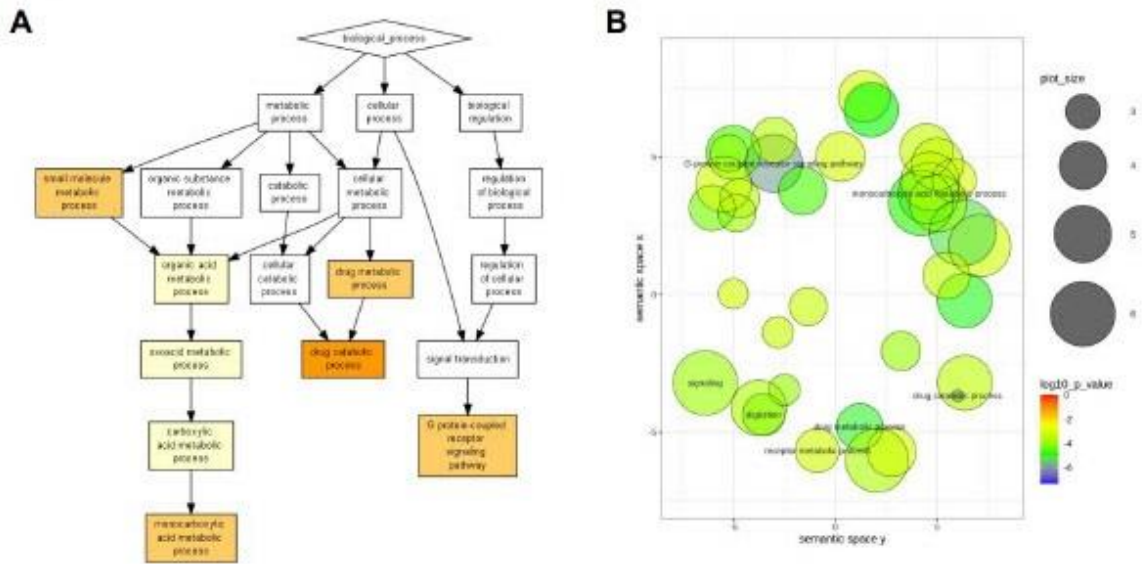
Fold enrichment over controls of significant eQTL and cASE variants, in islet regulatory chromatin regions, related to Figures 2C and 5D. *p*-values for 1% FDR eQTL enrichments. A) all eQTL variants, B) Top eQTL variants: common ($\geq 5\%$ MAF) and low-frequency variants ($1\% < \text{MAF} < 5\%$), C) all cASE variants, D) Top cASE variants: common ($\geq 5\%$ MAF) and low-frequency variants ($1\% < \text{MAF} < 5\%$).

Figure S2.



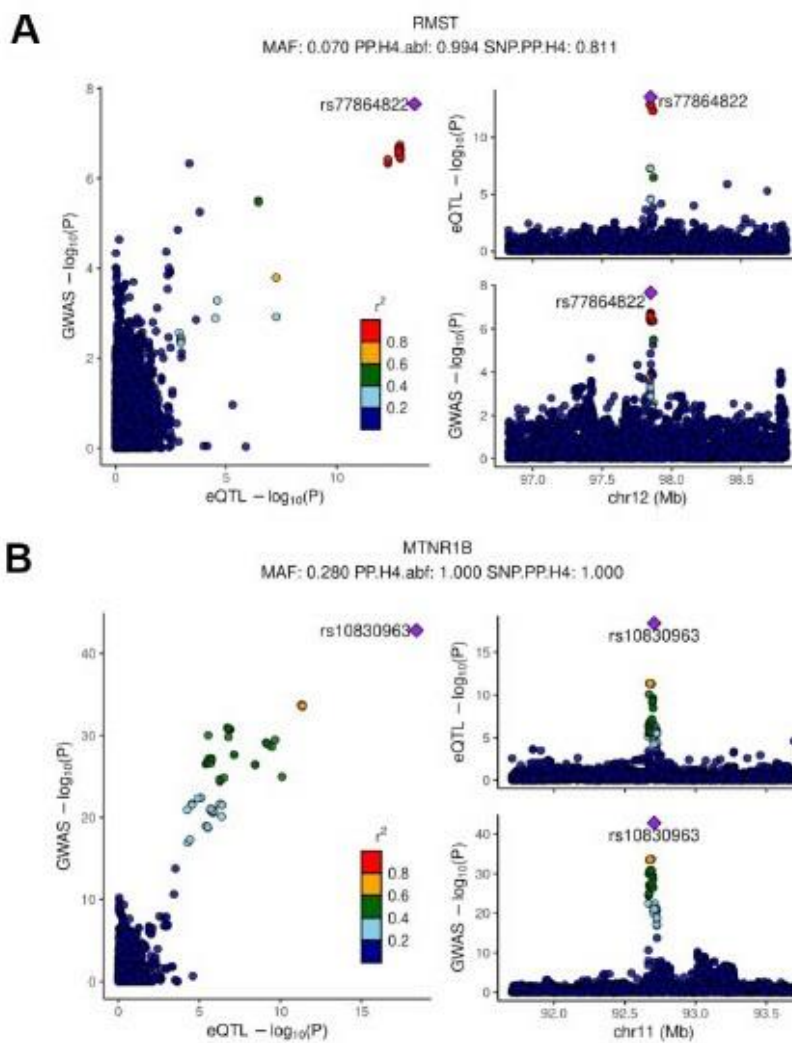
Comparison between TIGER and InsPIRE eQTL results, related to Figure 2D. A) Overlap between the number of genomic variants interrogated in TIGER and InsPIRE. 7.22M were in both studies (brown), while 2.88M (28.5%) (green) were exclusive to TIGER, and 430k (orange) to InsPIRE. B) Overlap between the significant eQTL results in both studies. 37% of TIGER results were eQTLs with TIGER-exclusive variants (light green), 19% with variants also present in InsPIRE (green), and 44% were shared with InsPIRE (brown). C) As in B) but restricted to eQTLs with variants of minor allele frequency <5%. D) Overlap between the eGenes detected by both studies. E) The p-values obtained in InsPIRE correlated (Pearson's $r=0.67$, red line) with the meta-analysis p-values of TIGER, for all eQTLs with variants present in both studies at <5% FDR significant in either study. F) For these eQTLs, the direction of effect was also consistent between studies (Pearson's $r=0.752$, red line).

Figure S3.



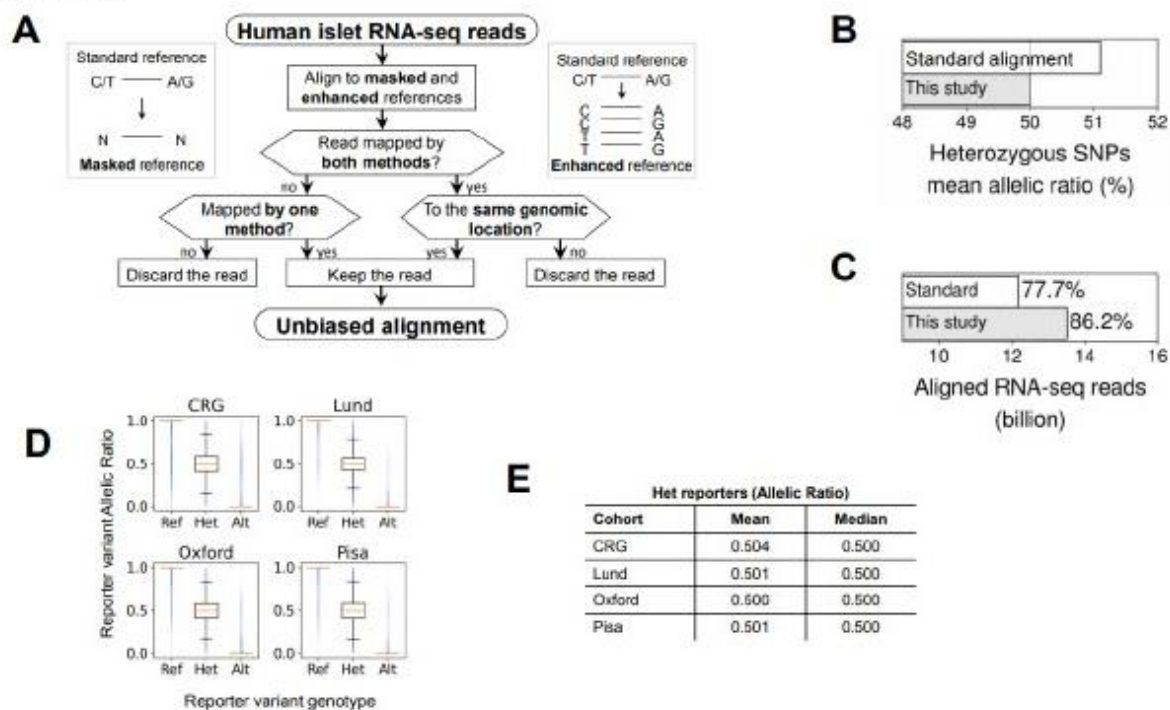
Gene ontology of the eQTL genes, related to Figure 2E. A) Gene ontology analysis of the genes with significant eQTLs compared with a background of all genes expressed in human islets, visualized with the Gene Ontology enRichment anaLysis and visualizAtion (GORilla) web tool. B) Same information visualized using Revigo.

Figure S4.



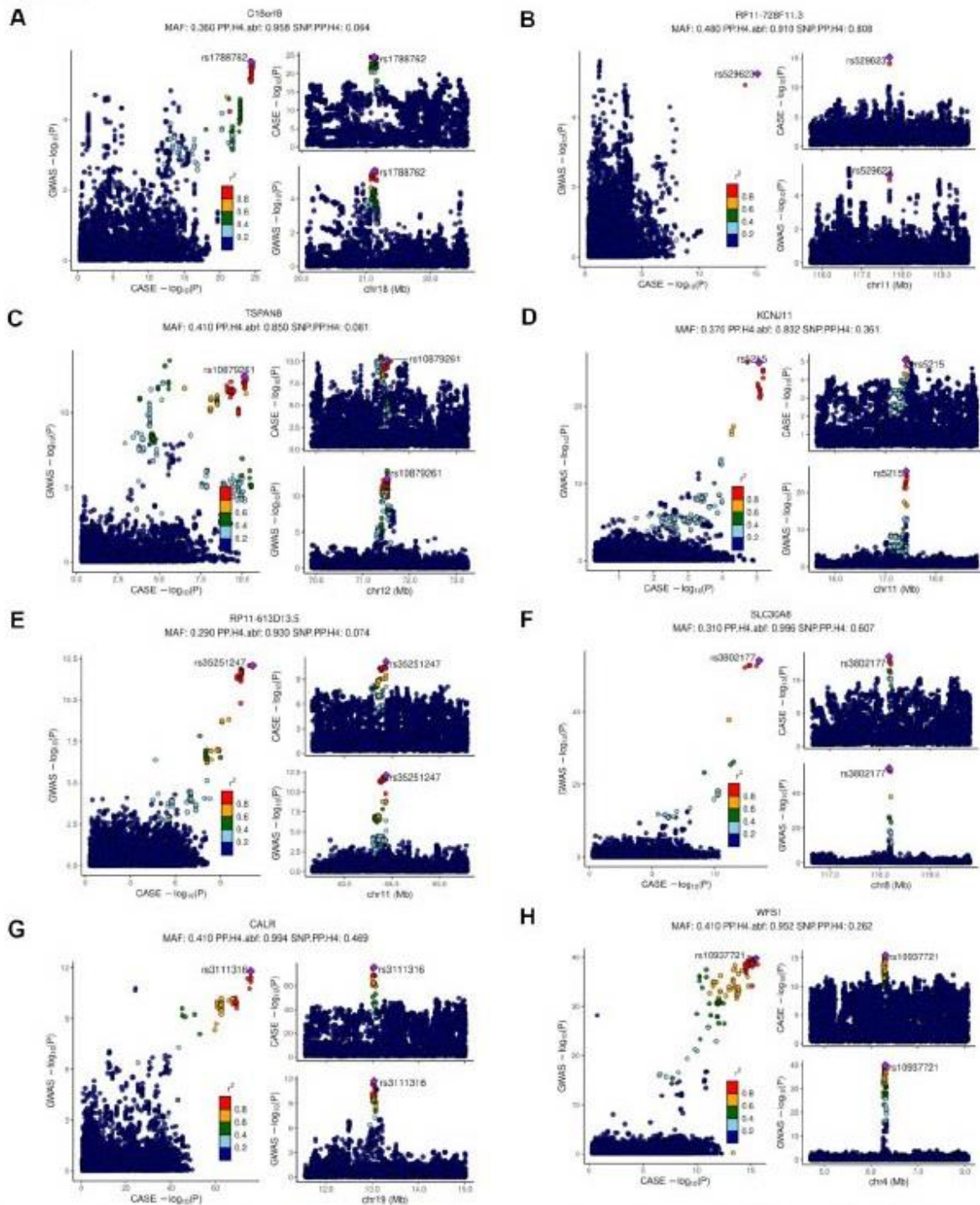
Colocalization plots of eQTL signals, related to Tables 1 and S4 . LocusCompare plots depicting A) *RMST* and B) *MTNR1B*. Significant colocalizations between eQTL and T2D GWAS analyses. The lead variant is represented by a purple diamond. The linkage disequilibrium between the lead variant and the other variants is given as the square of the correlation coefficient r^2 and is indicated in a color scale. The $-\log_{10}(p\text{-values})$ for each variant — which are located in a region of one mega-base pair up- and downstream from the gene transcription start site — are depicted in three panels: (left) $p\text{-values}$ of eQTL as x-axis and GWAS as y-axis, (bottom right) $p\text{-values}$ of GWAS in the gene region and (top right) $p\text{-values}$ of eQTL in the gene region. The title shows the gene name; MAF: minor allele frequency; PP.H4.abf: Posterior probability of colocalization; SNP.PP.H4: posterior probability of lead variant being the associated causal variant.

Figure S5.



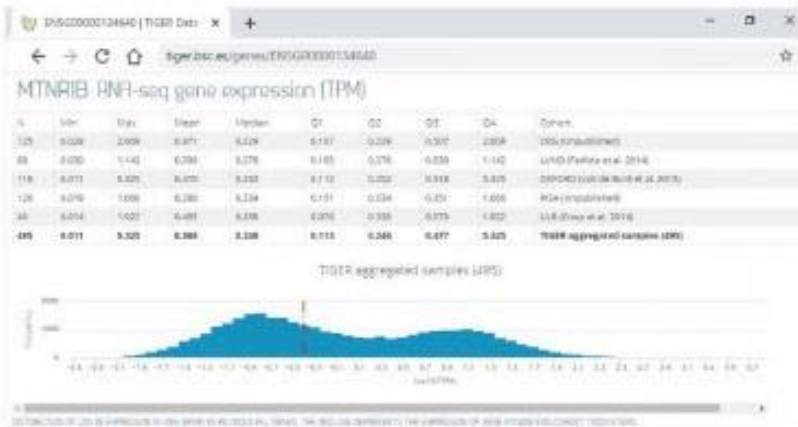
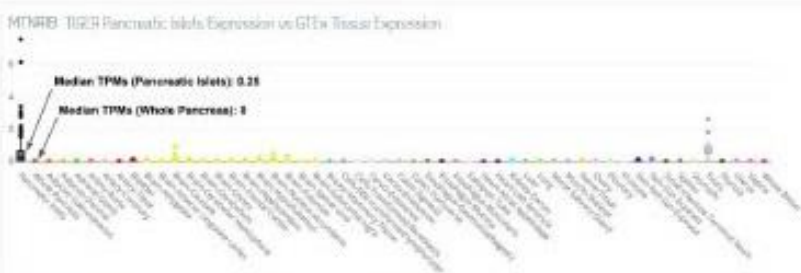
Overview of the RNA-seq alignment used in cASE analysis, related to Figure 4A and STAR Methods. A) Decision tree of the RNA-seq reads to keep or discard, after alignment using both a masked and an enhanced reference genome. **B)** Mean allelic ratio resulting from the ‘unbiased alignment’ method compared versus an alignment using a standard reference genome. **C)** As in B but showing the percentage of raw reads aligned. **D)** Allelic ratio of reporter variants separated by their genotypes, across the four cohorts. **E)** Mean and median allelic ratio values for each of the cohorts.

Figure S6.



Colocalization plots of cASE signals without an observed eQTL/T2D GWAS colocalization, related to Tables 2 and S5. LocusCompare plots depicting significant colocalizations between cASE and T2D GWAS analyses. A) *C18orf18* B) *RP11-728F11.3* C) *TSPAN8* D) *KCNJ11* E) *RP11-613D13.5* F) *SLC30A8* G) *CALR* H) *WFS1*. The lead variant is represented by a purple diamond. The linkage disequilibrium between the lead variant and the other

variants is given as the square of the correlation coefficient r^2 and is indicated in a color scale. The $-\log_{10}(p\text{-values})$ for each variant — which are located in a region of one mega-base pair up- and downstream from the gene transcription start site — are depicted in three panels: (left) p -values of cASE as x-axis and GWAS as y-axis, (bottom right) p -values of GWAS in the gene region and (top right) p -values of cASE in the gene region. The title shows the gene name; MAF: the minor allele frequency; PP.H4.abf: Posterior probability of colocalization; SNP.PP.H4: posterior probability of lead variant being the associated causal variant.

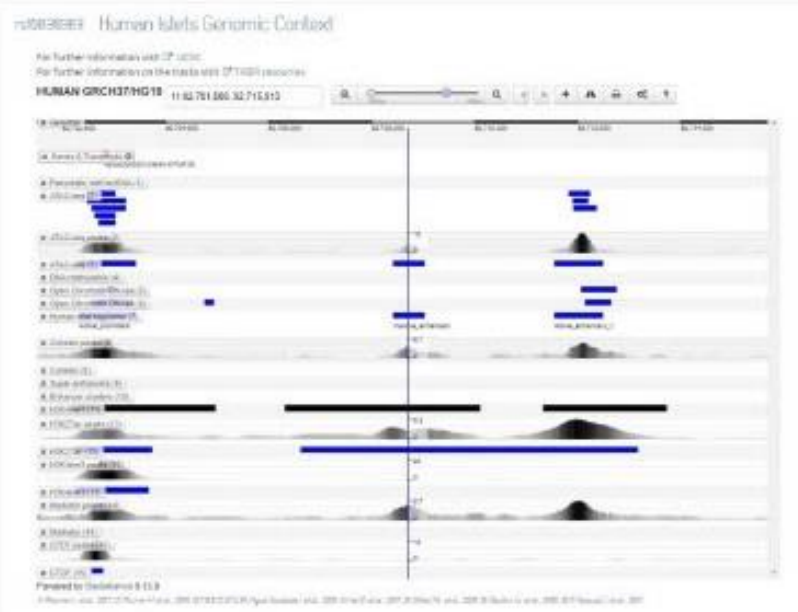
Figure S7**A****B****C**

R1 variant at ± 100bp of MTNR1B and their associations

Showing 1 - 10 of 2212 variant associations [10000]

Show data: INFO P R2 R1 DIAGRAM DIAGRAMS DIAGRAM (1000) DIAGRAM (non-ref)

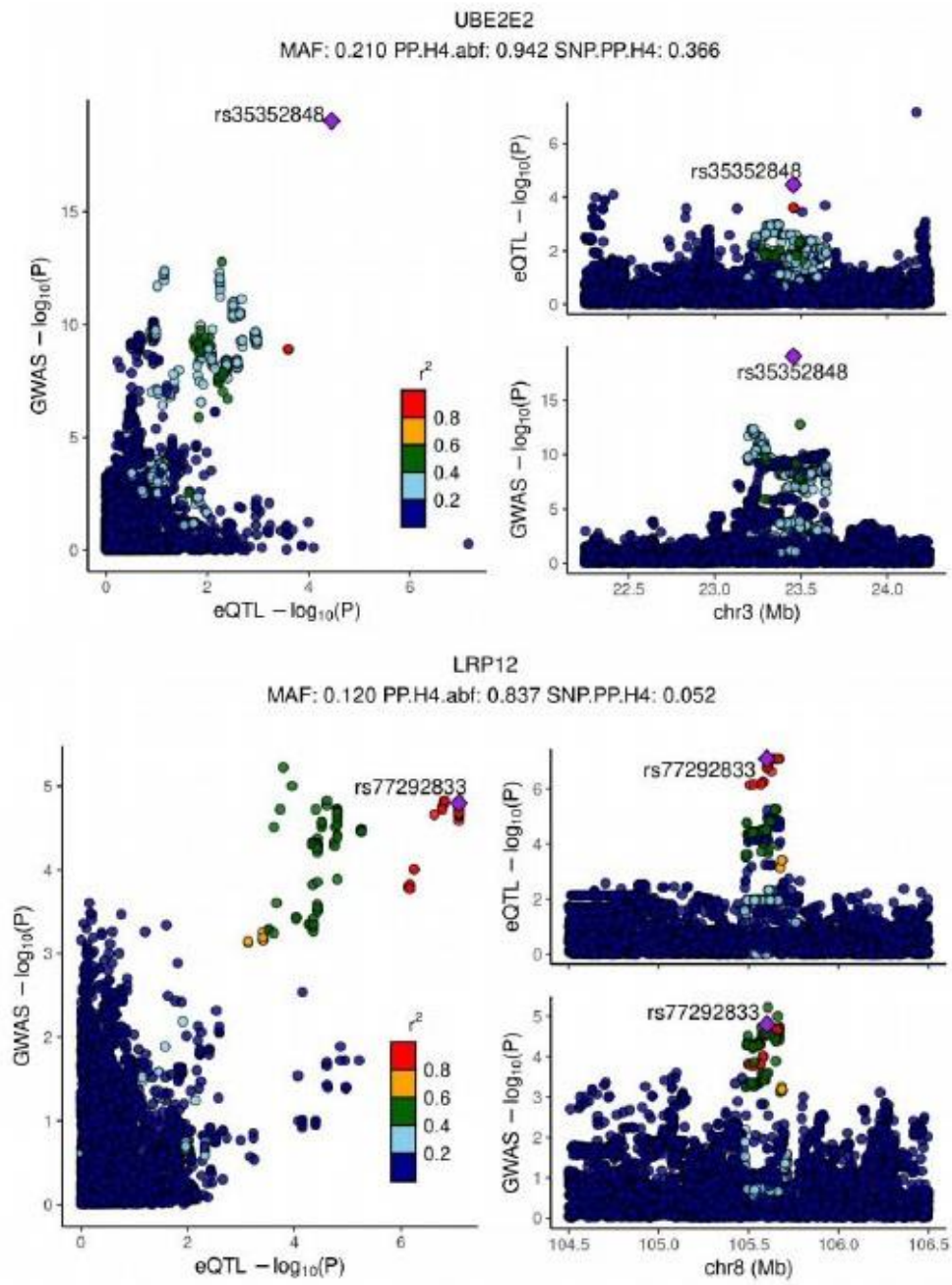
rsID	chr	pos	ref	alt	freq	info	r2	r1	di	diags	diag1000	diagnonref
rs1000000	1	11,222,222	A	G	0.5	0.9	0.9	0.9	0.9			
rs1122222	1	11,222,222	A	G	0.5	0.9	0.9	0.9	0.9			
rs1333333	1	11,222,222	A	G	0.5	0.9	0.9	0.9	0.9			
rs1444444	1	11,222,222	A	G	0.5	0.9	0.9	0.9	0.9			
rs1555555	1	11,222,222	A	G	0.5	0.9	0.9	0.9	0.9			
rs1666666	1	11,222,222	A	G	0.5	0.9	0.9	0.9	0.9			
rs1777777	1	11,222,222	A	G	0.5	0.9	0.9	0.9	0.9			
rs1888888	1	11,222,222	A	G	0.5	0.9	0.9	0.9	0.9			
rs1999999	1	11,222,222	A	G	0.5	0.9	0.9	0.9	0.9			
rs2111111	1	11,222,222	A	G	0.5	0.9	0.9	0.9	0.9			

D

TIGER platform example, related to Figures 1A, S4B and STAR Methods. A) *MTNR1B* normalized $\log_{10}(\text{TPM})$ expression in islets; table (top) displays *MTNR1B* normalized TPM expression in each cohort and across the cohorts (bold); histogram (bottom) shows $\log_{10}(\text{TPM})$ gene expression distribution in 495 human islets samples, the red dashed line corresponds to *MTNR1B* $\log_{10}(\text{TPM})$ expression. B) *MTNR1B* normalized TPM expression in islets vs other GTEx tissues where each boxplot represents one tissue; *MTNR1B* has higher expression in pancreatic islets (black) compared to the whole pancreas (brown), which has almost no expression. C) Table showing the list of variants in a 100Kb window around *MTNR1B* and displaying results from either eQTL or DIAMANTE GWAS data sorted by ascending eQTL p-value; the eQTL variant rs10830963 ($p=4.04\times 10^{-19}$) colocalizes with DIAMANTE ($p=1.50\times 10^{-43}$). D) 15Kb human islet genomic context of variant rs10830963 (chr11:92708710); islet significant regions (black/blue boxes) and peaks are represented in each track, the blue line corresponds to rs10830963 position.

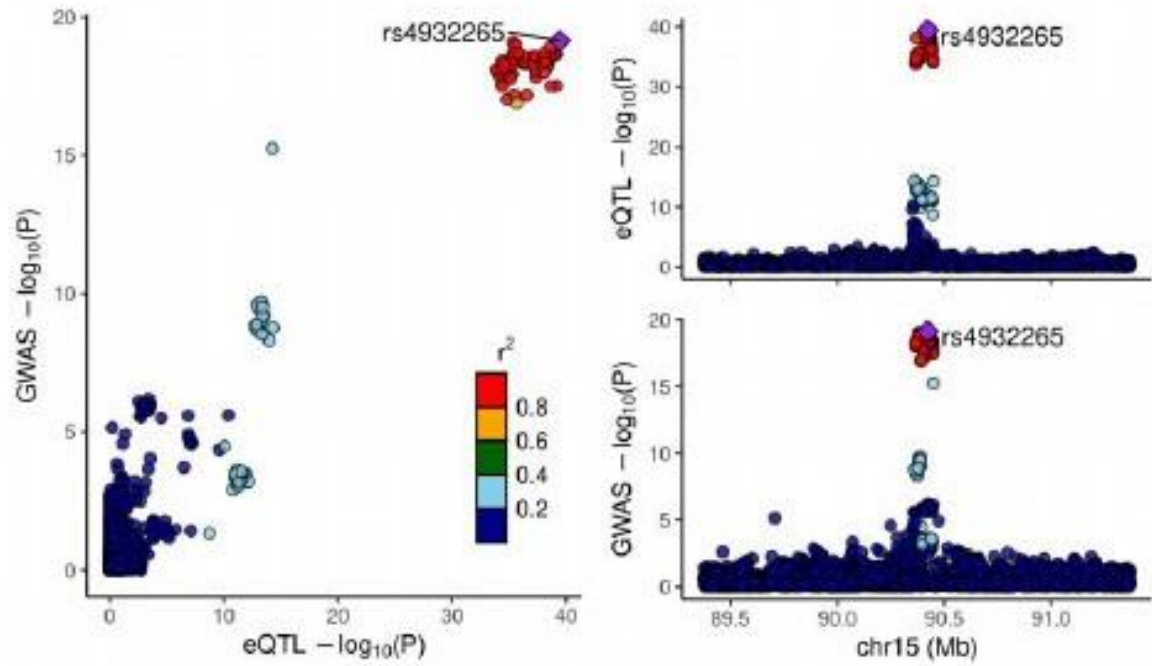
Supplemental Data

Data S1



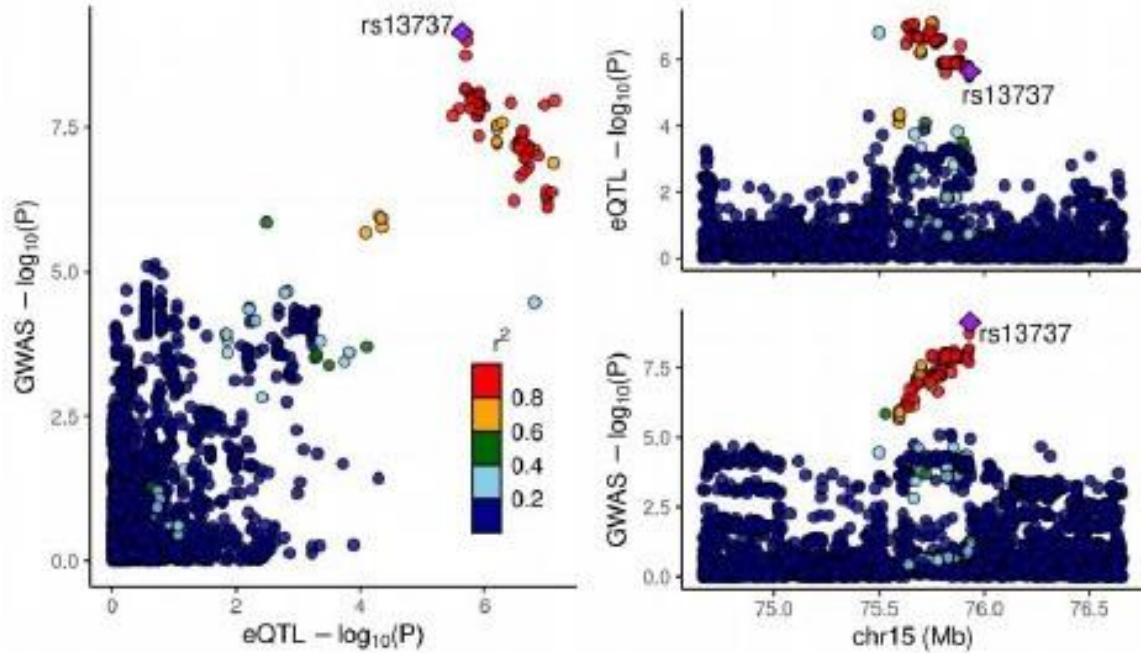
AP3S2

MAF: 0.270 PP.H4.abf: 0.974 SNP.PP.H4: 0.505



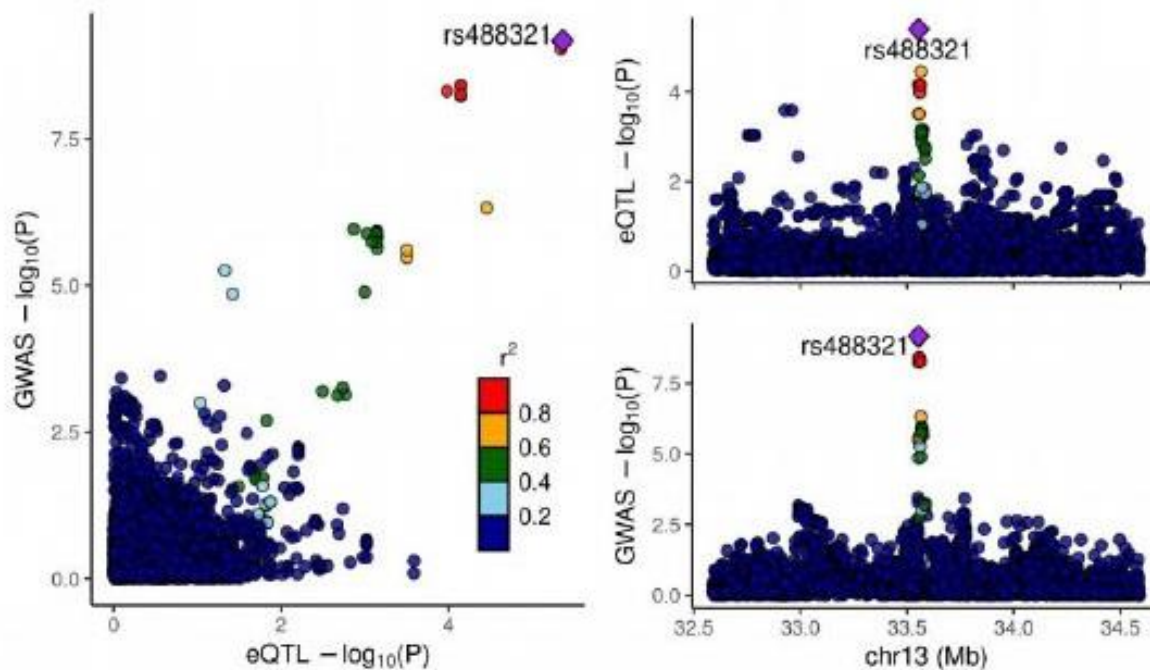
RP11-817O13.8

MAF: 0.240 PP.H4.abf: 0.839 SNP.PP.H4: 0.097



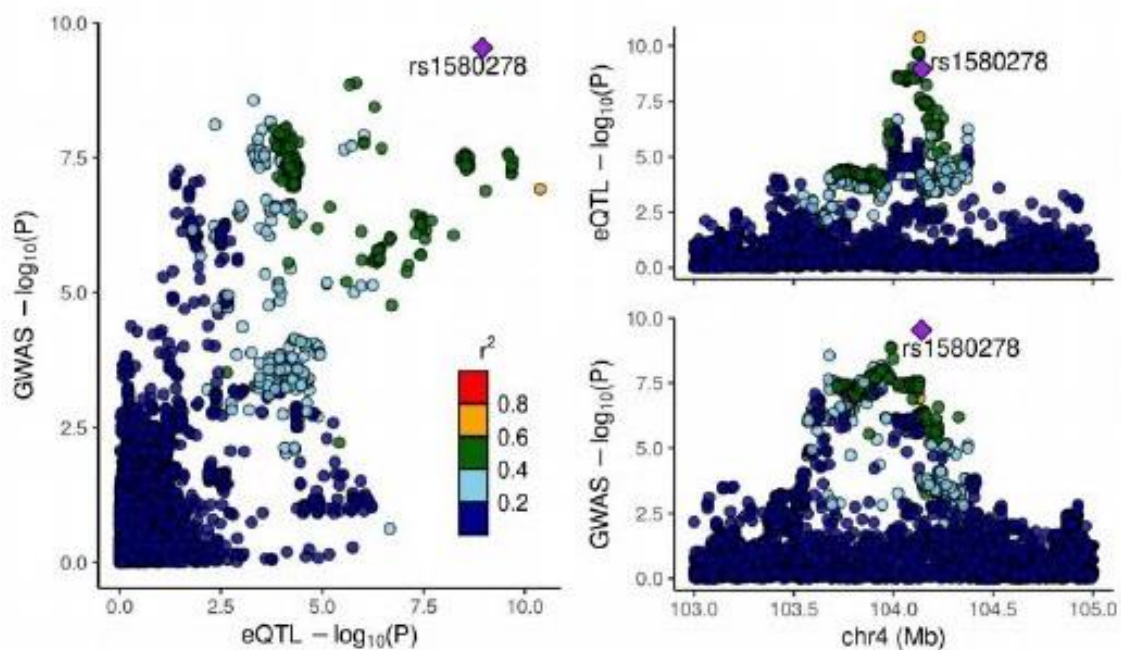
KL

MAF: 0.170 PP.H4.abf: 0.978 SNP.PP.H4: 0.266



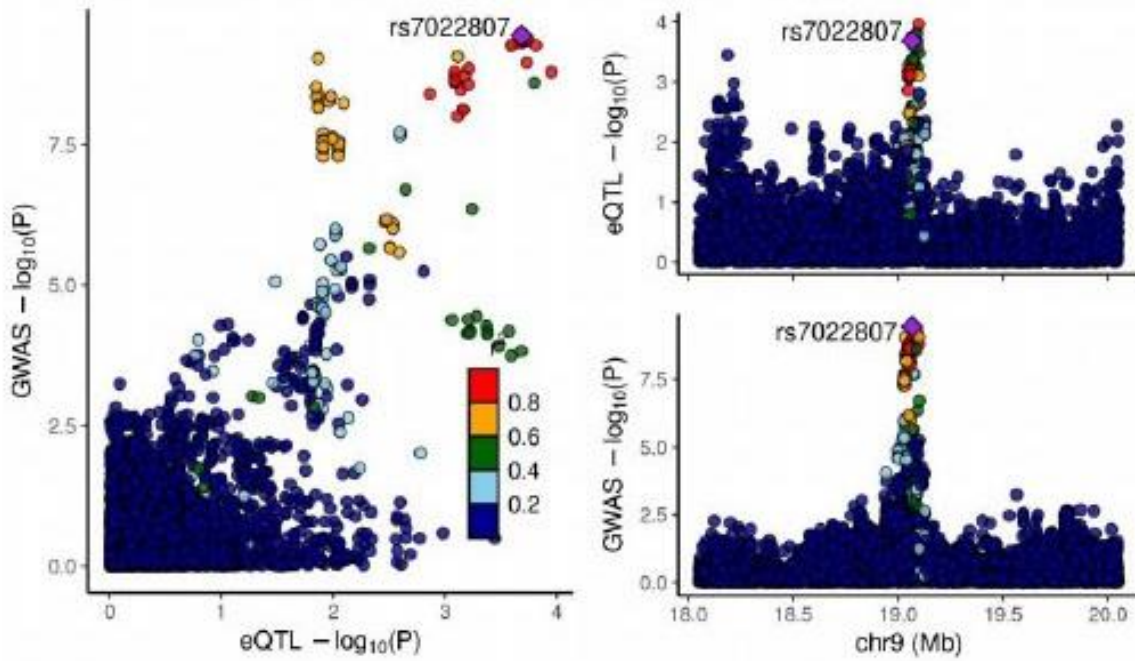
BDH2

MAF: 0.470 PP.H4.abf: 0.812 SNP.PP.H4: 0.727



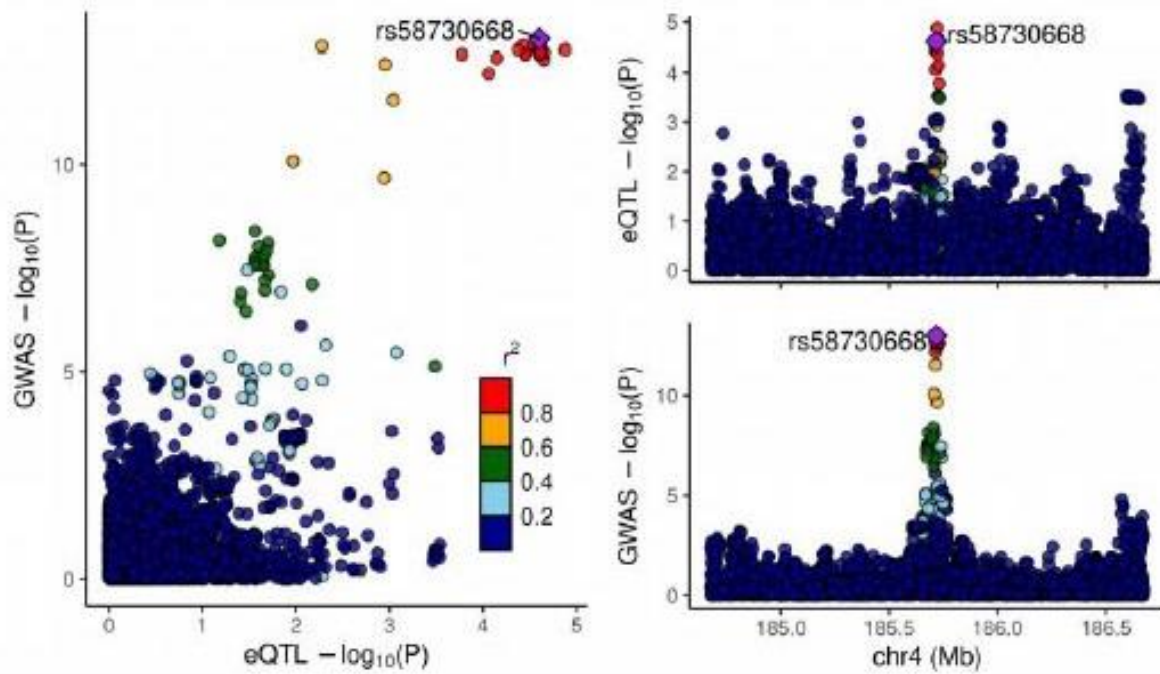
HAUS6

MAF: 0.400 PP.H4.abf: 0.820 SNP.PP.H4: 0.098



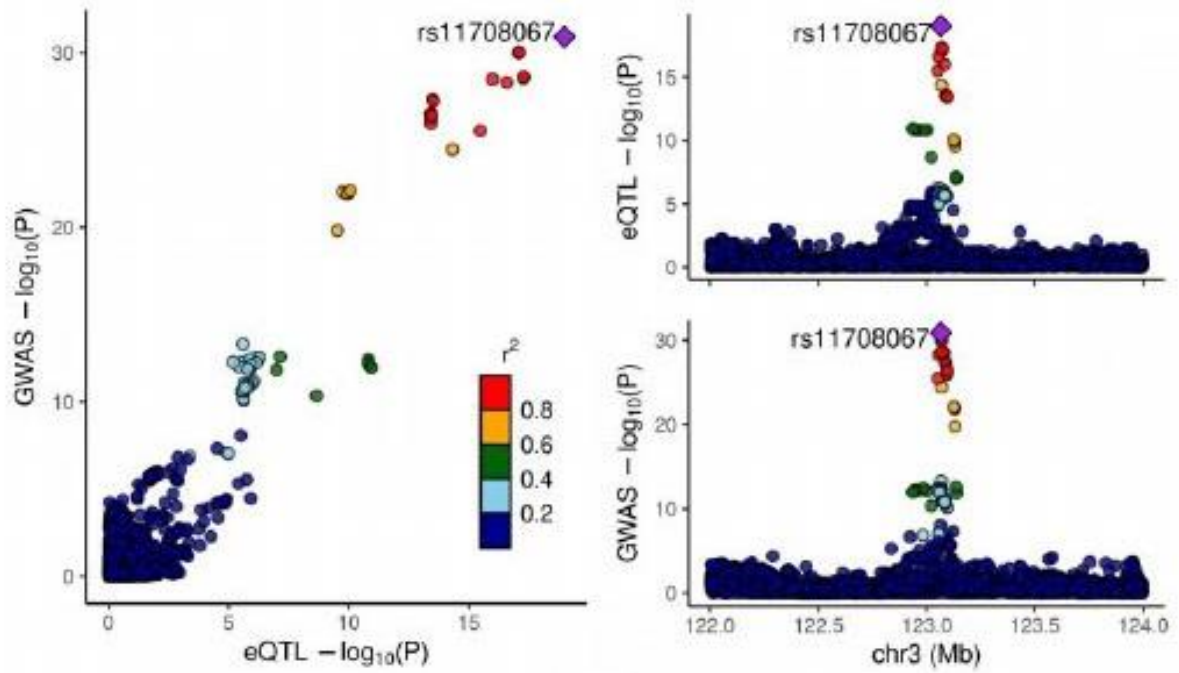
ACSL1

MAF: 0.140 PP.H4.abf: 0.893 SNP.PP.H4: 0.039



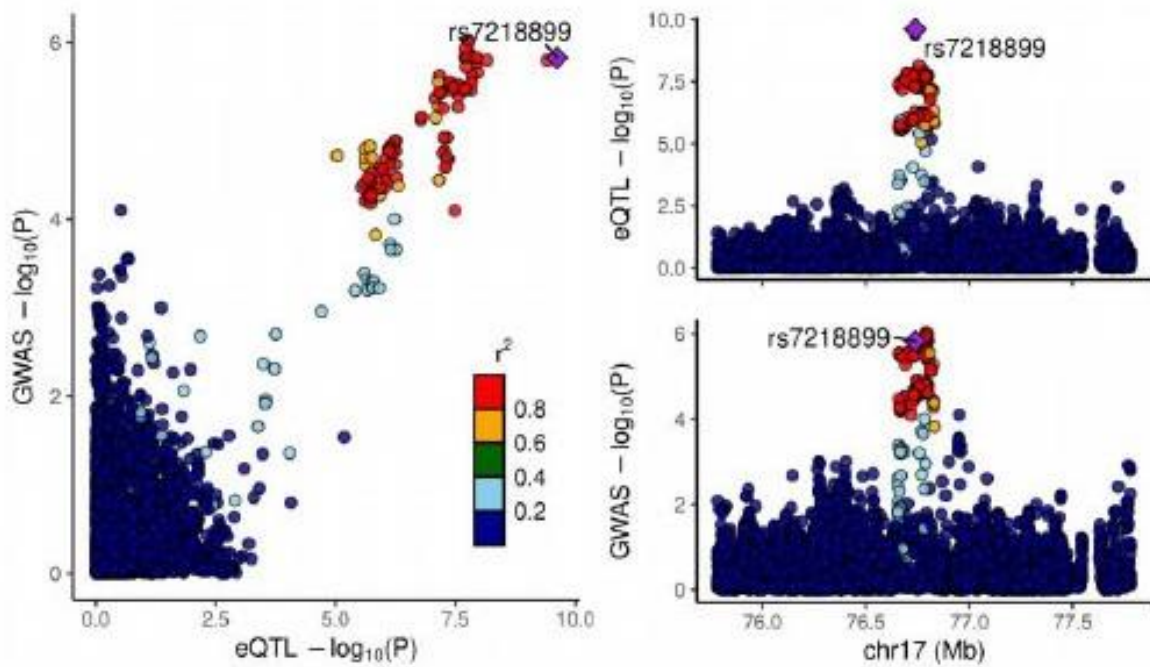
ADCY5

MAF: 0.230 PP.H4.abf: 0.999 SNP.PP.H4: 0.990



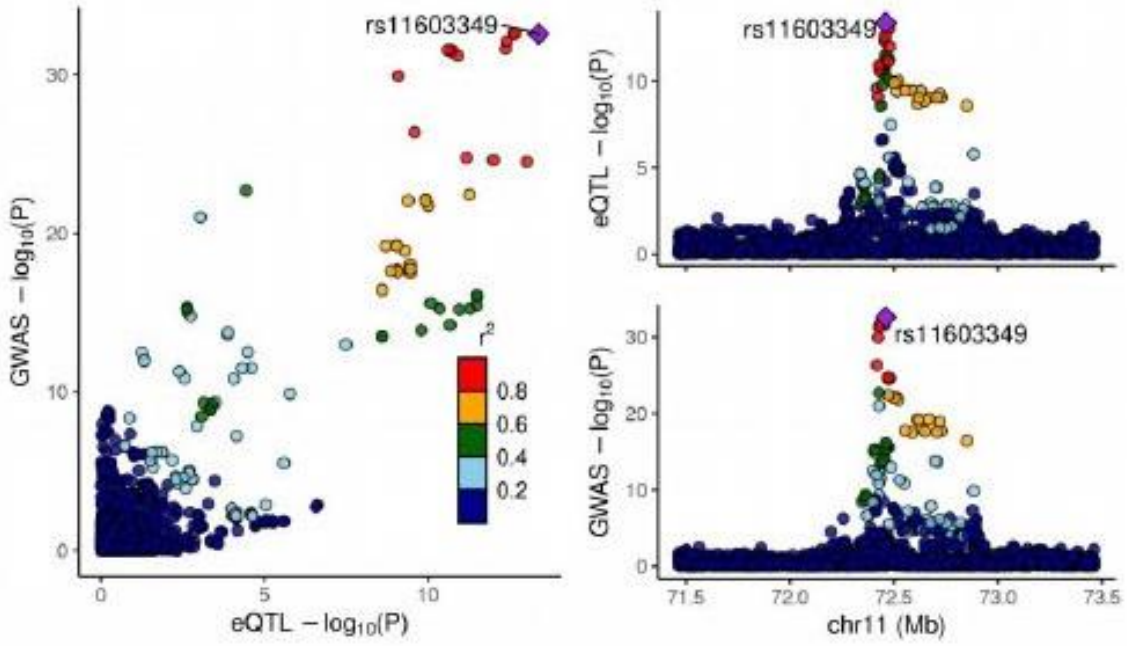
USP36

MAF: 0.490 PP.H4.abf: 0.959 SNP.PP.H4: 0.415



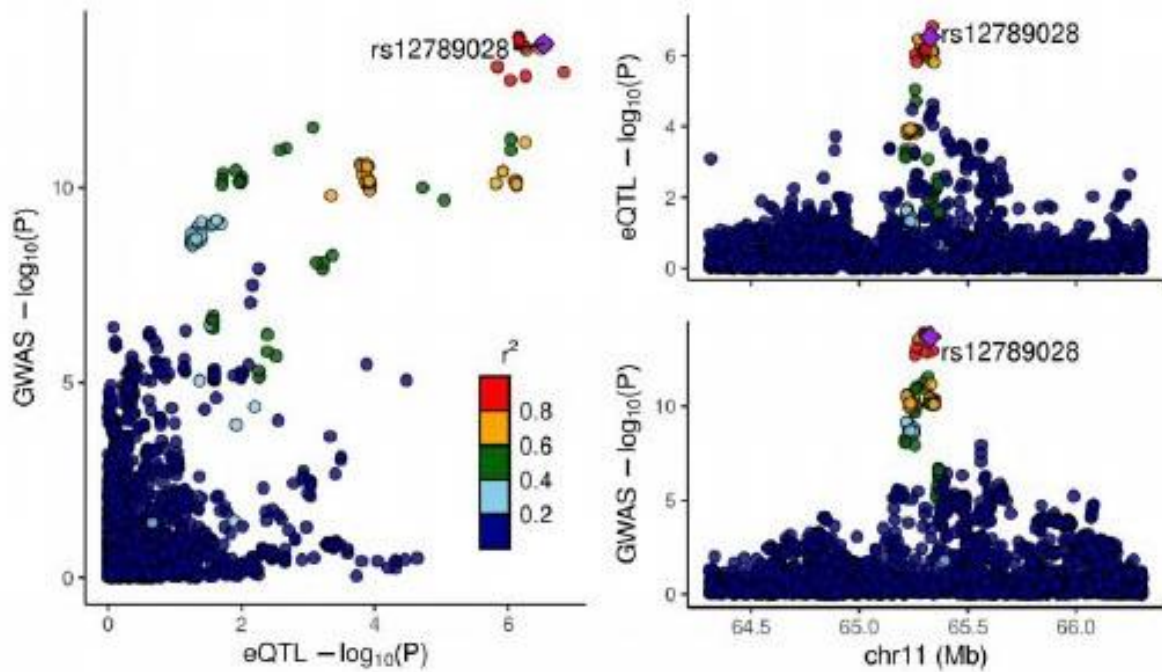
STARD10

MAF: 0.160 PP.H4.abf: 0.991 SNP.PP.H4: 0.416



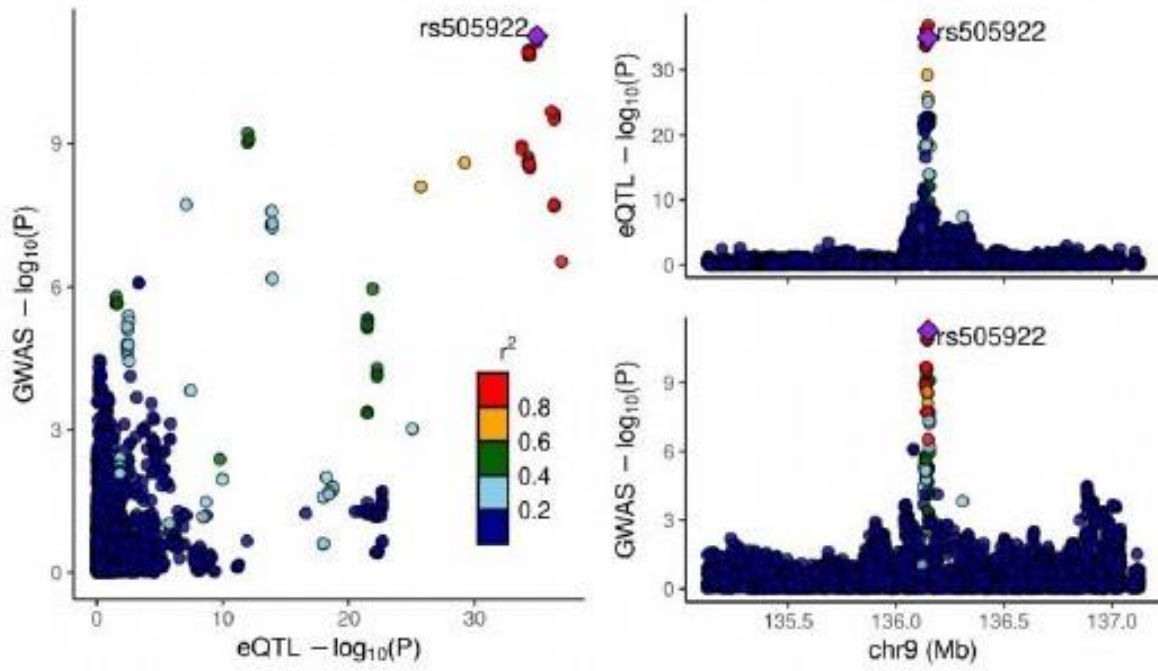
LTBP3

MAF: 0.190 PP.H4.abf: 0.973 SNP.PP.H4: 0.156



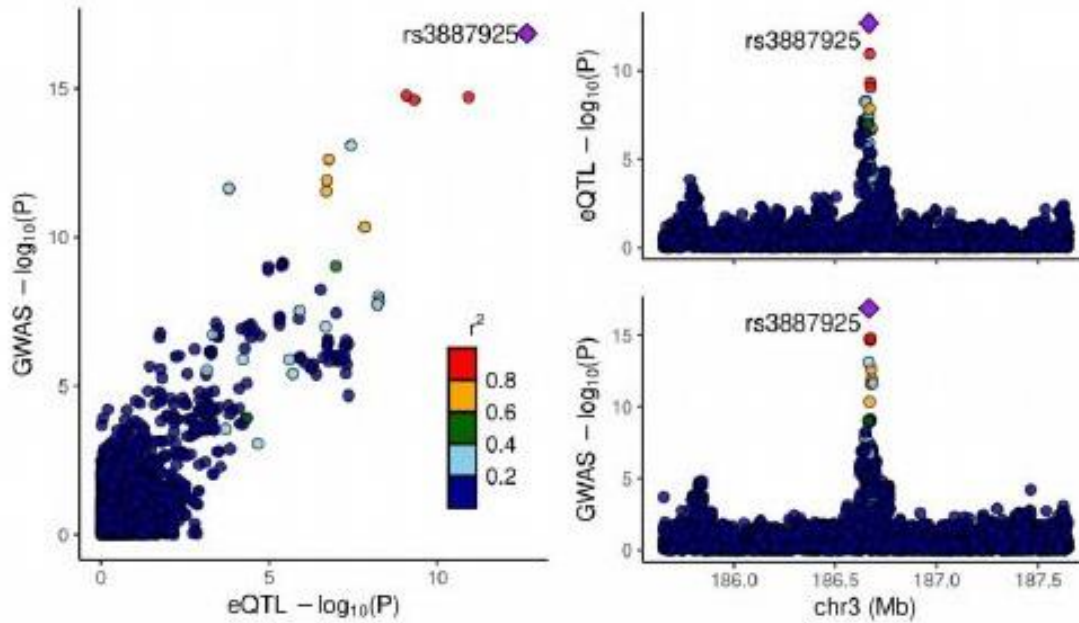
ABO

MAF: 0.330 PP.H4.abf: 0.840 SNP.PP.H4: 0.220



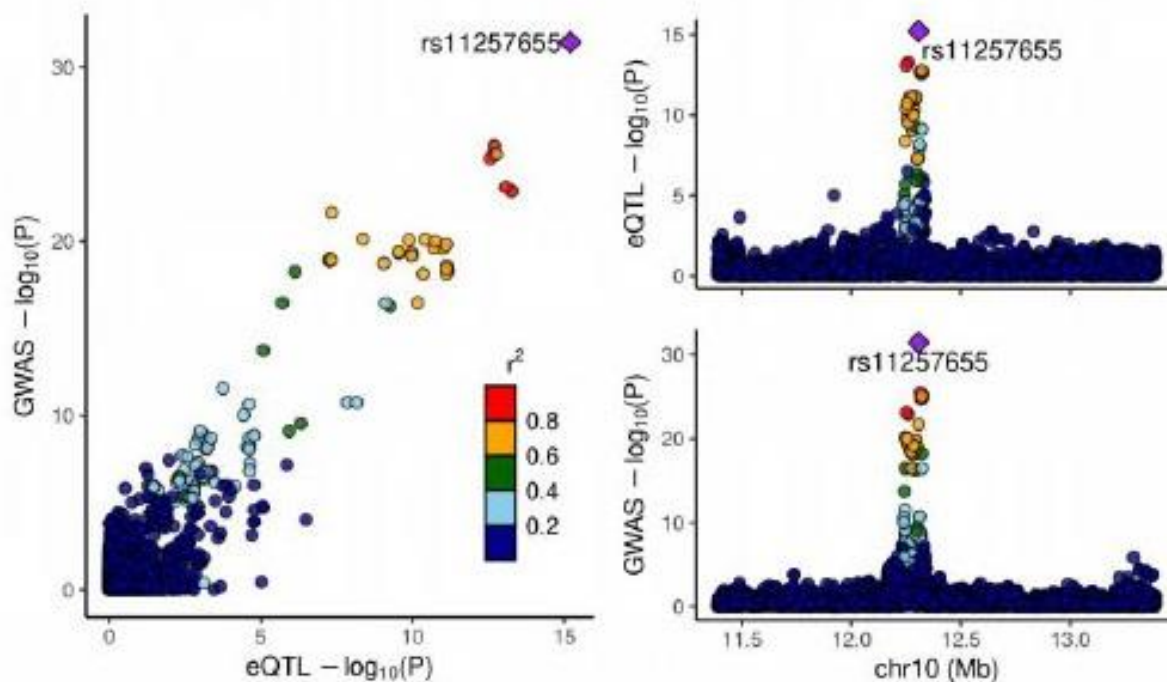
ST6GAL1

MAF: 0.450 PP.H4.abf: 1.000 SNP.PP.H4: 1.000



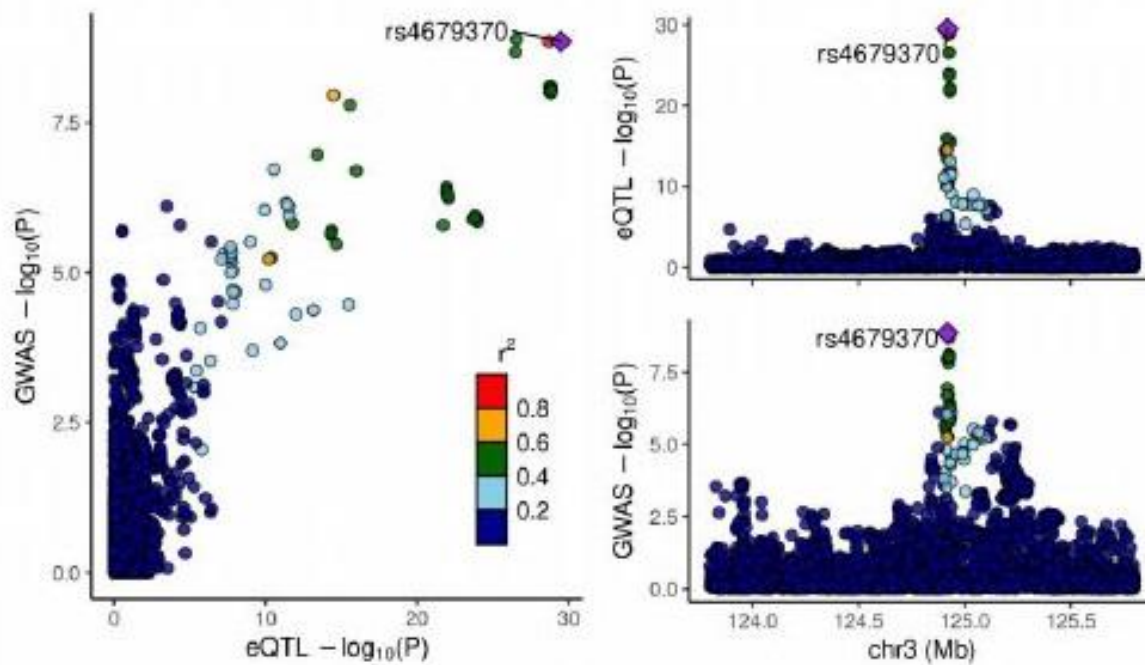
CAMK1D

MAF: 0.220 PP.H4.abf: 1.000 SNP.PP.H4: 1.000



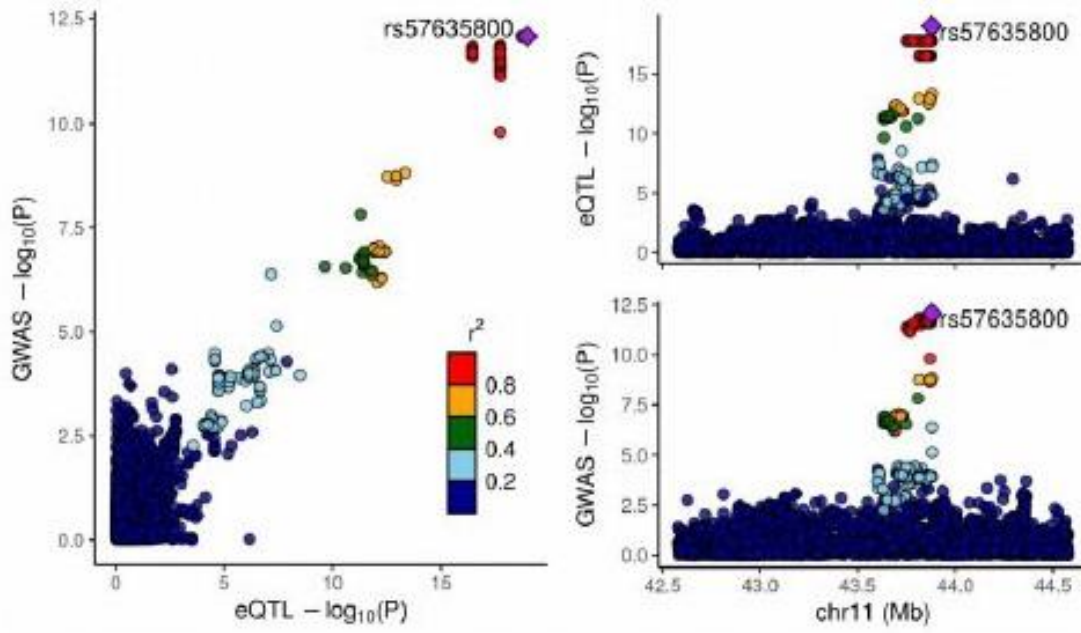
SLC12A8

MAF: 0.460 PP.H4.abf: 0.990 SNP.PP.H4: 0.633



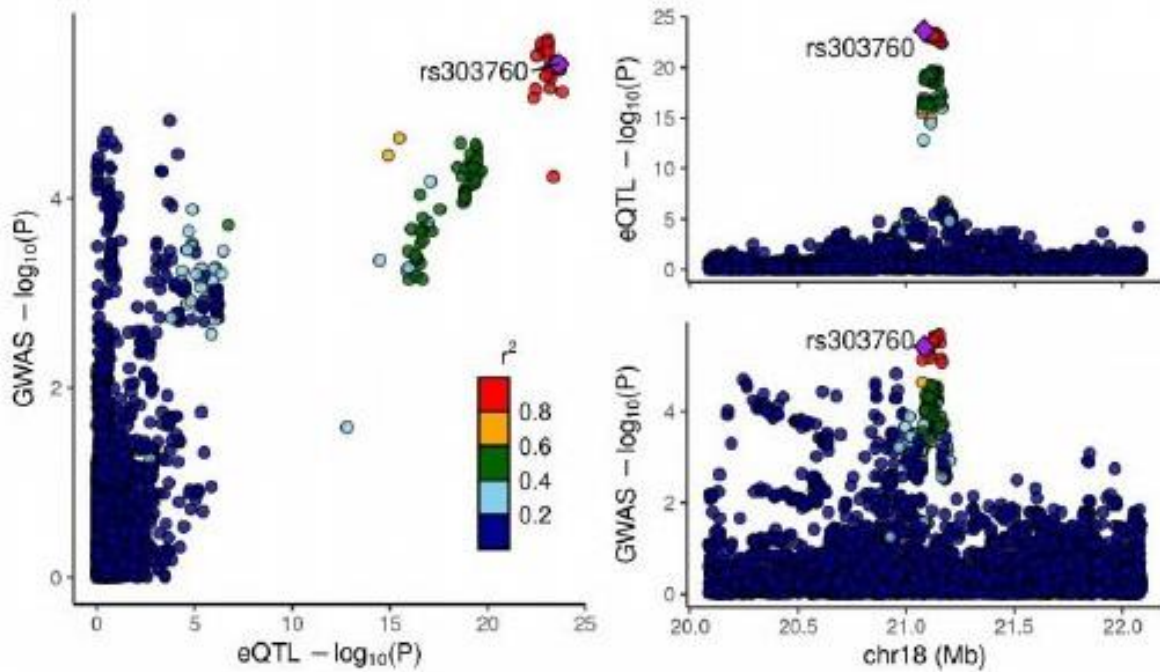
HSD17B12

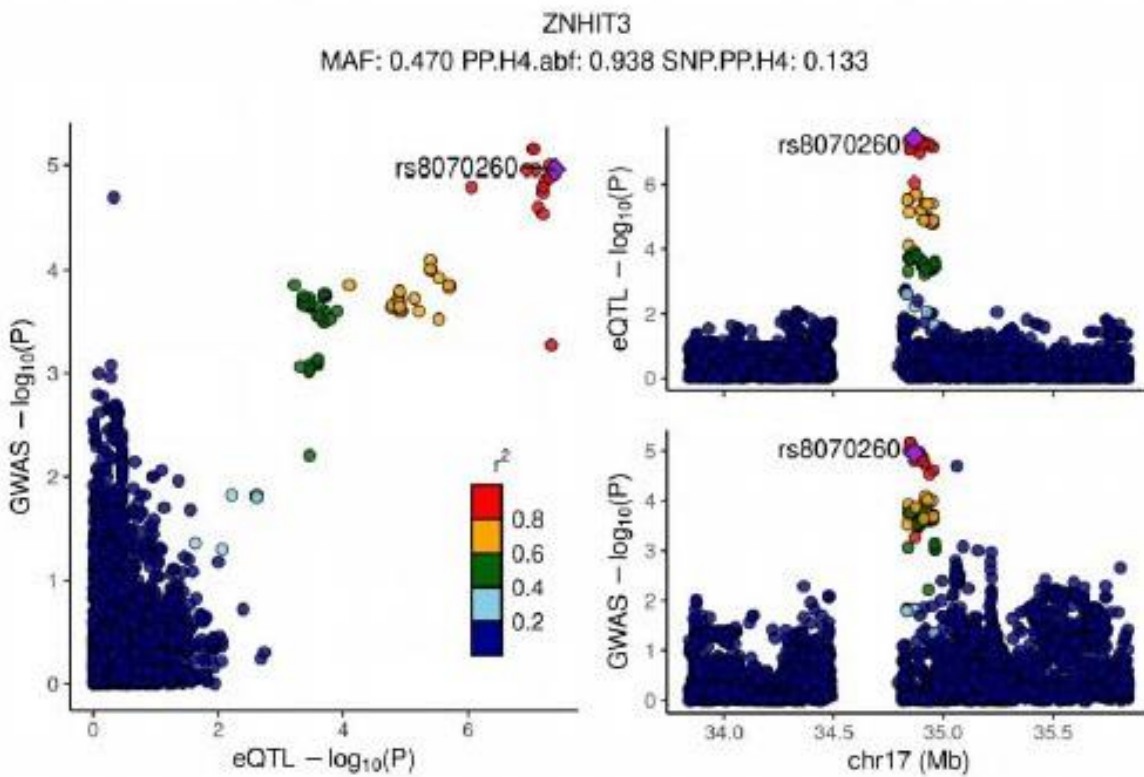
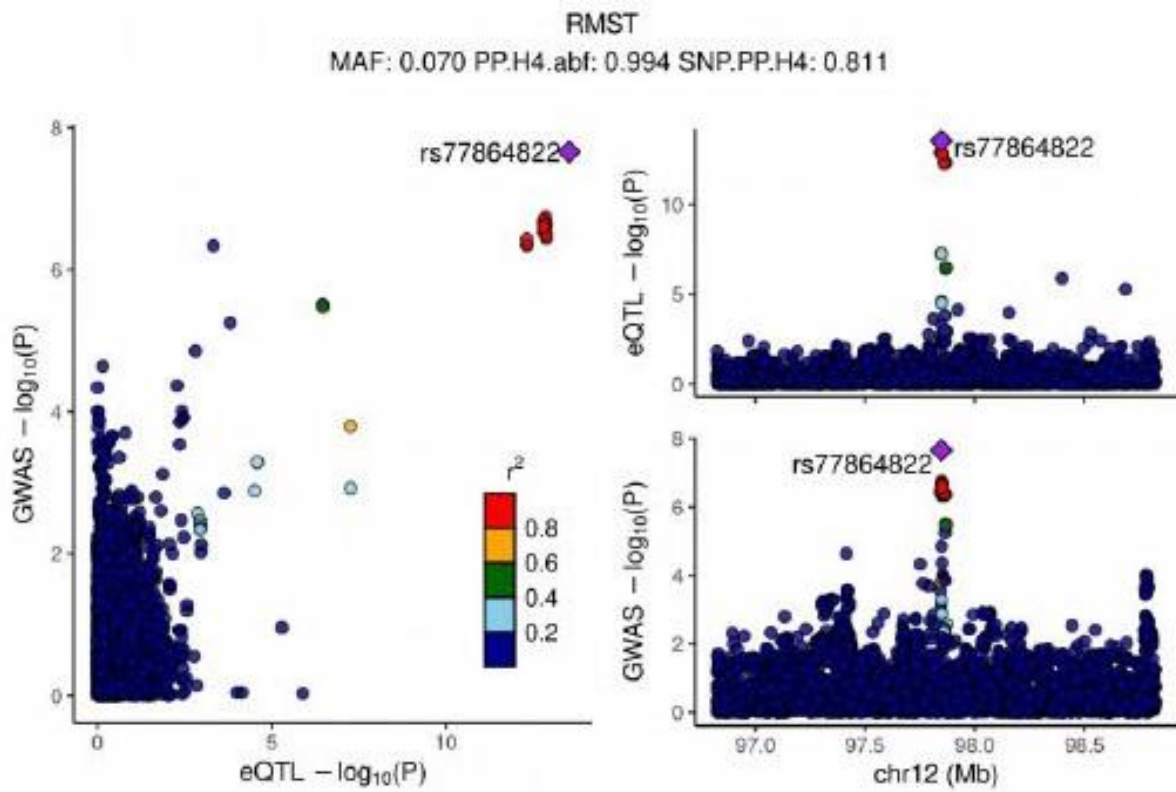
MAF: 0.290 PP.H4.abf: 0.953 SNP.PP.H4: 0.235



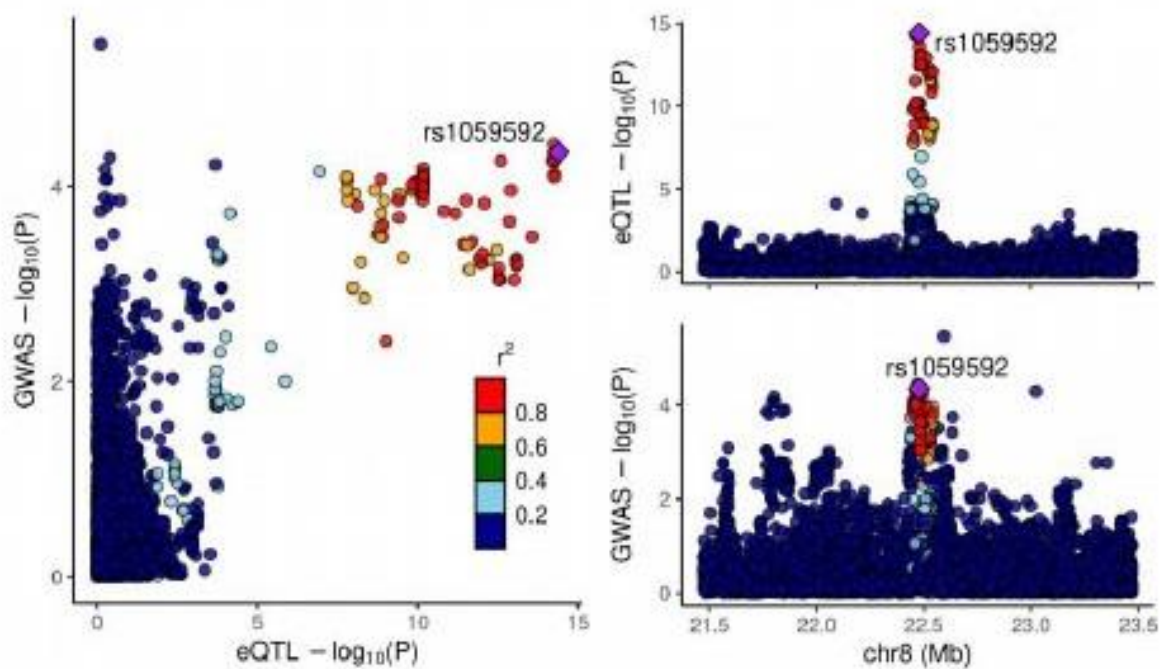
NPC1

MAF: 0.360 PP.H4.abf: 0.952 SNP.PP.H4: 0.080

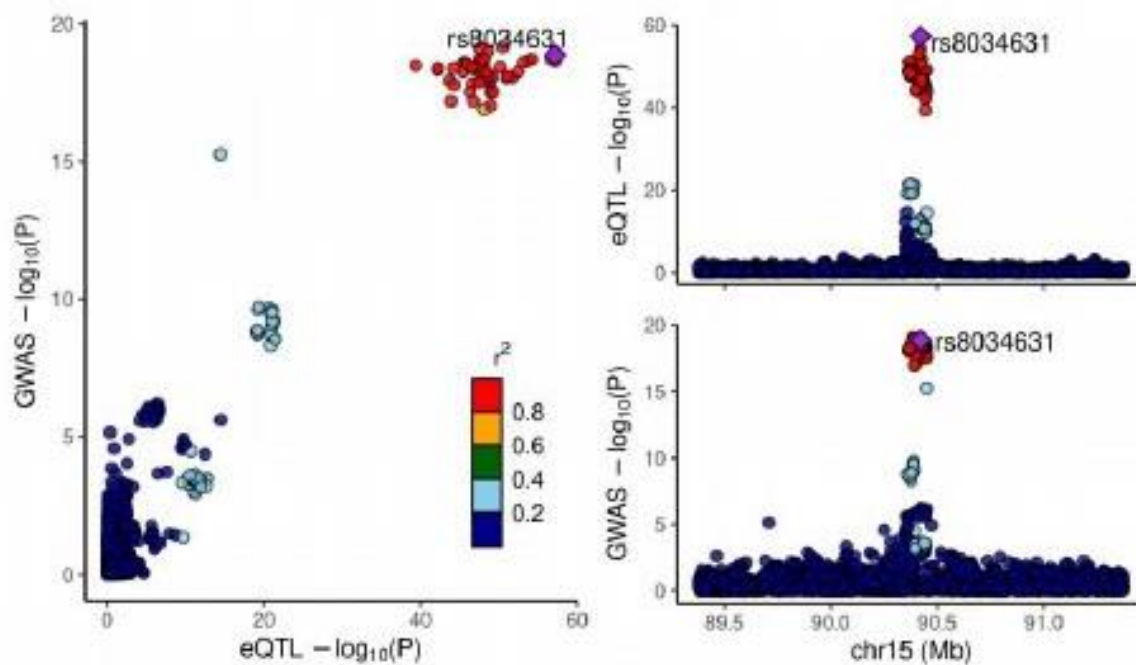




RP11-582J16.5
MAF: 0.350 PP.H4.abf: 0.806 SNP.PP.H4: 0.123

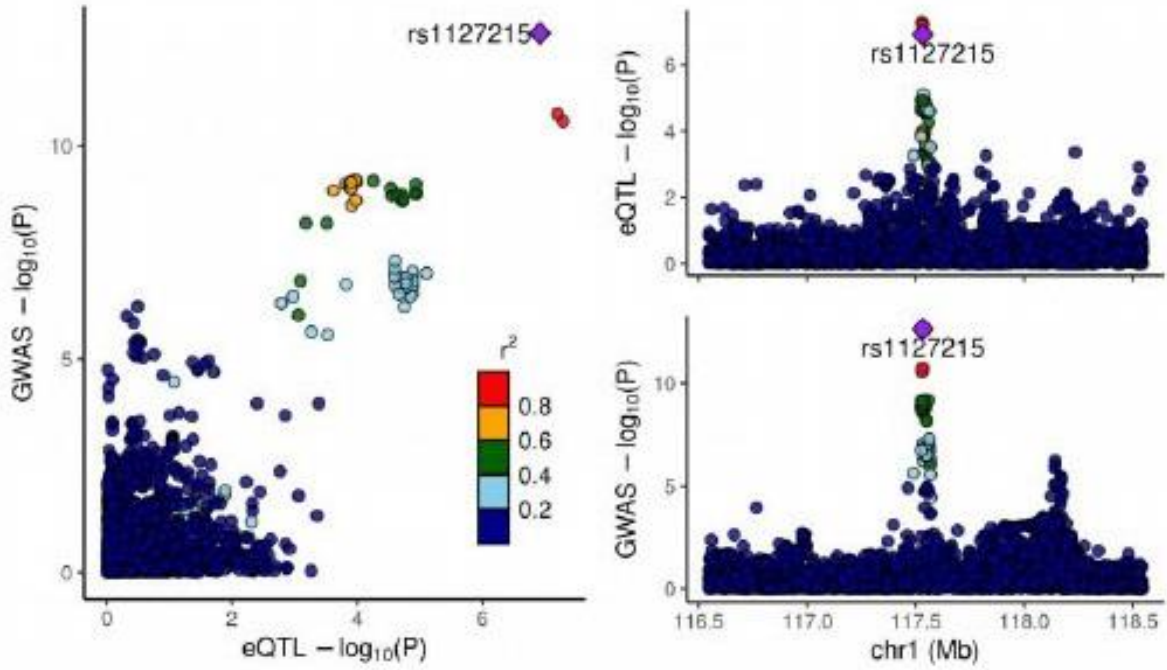


C15orf38-AP3S2
MAF: 0.280 PP.H4.abf: 0.974 SNP.PP.H4: 0.521



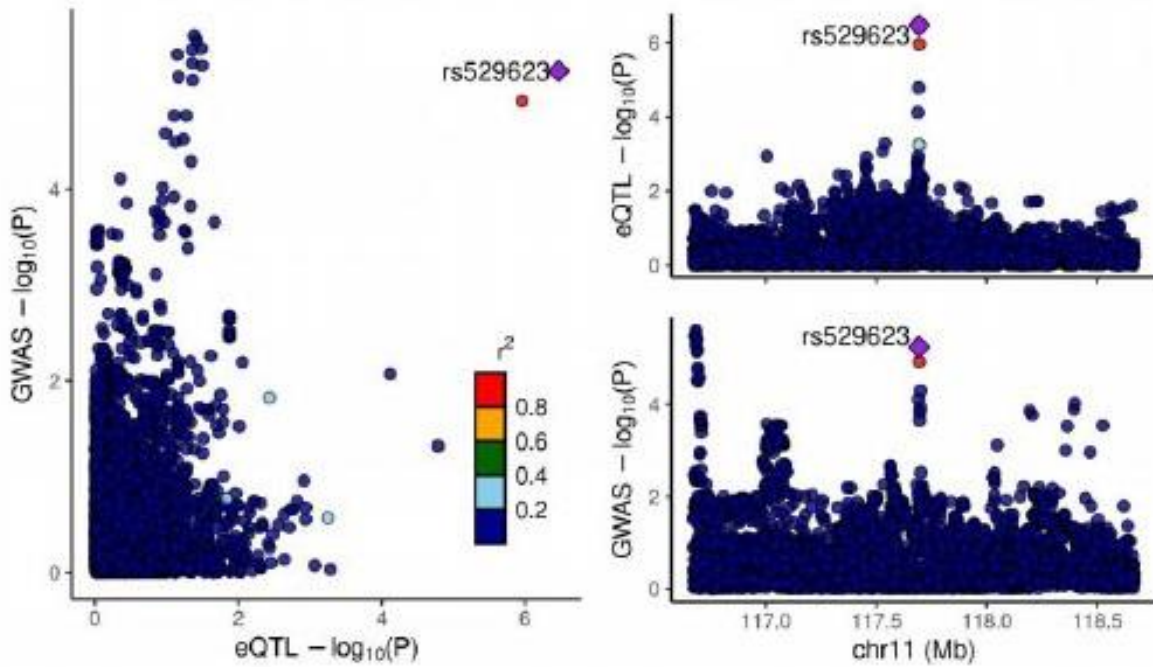
CD101

MAF: 0.420 PP.H4.abf: 0.995 SNP.PP.H4: 0.959



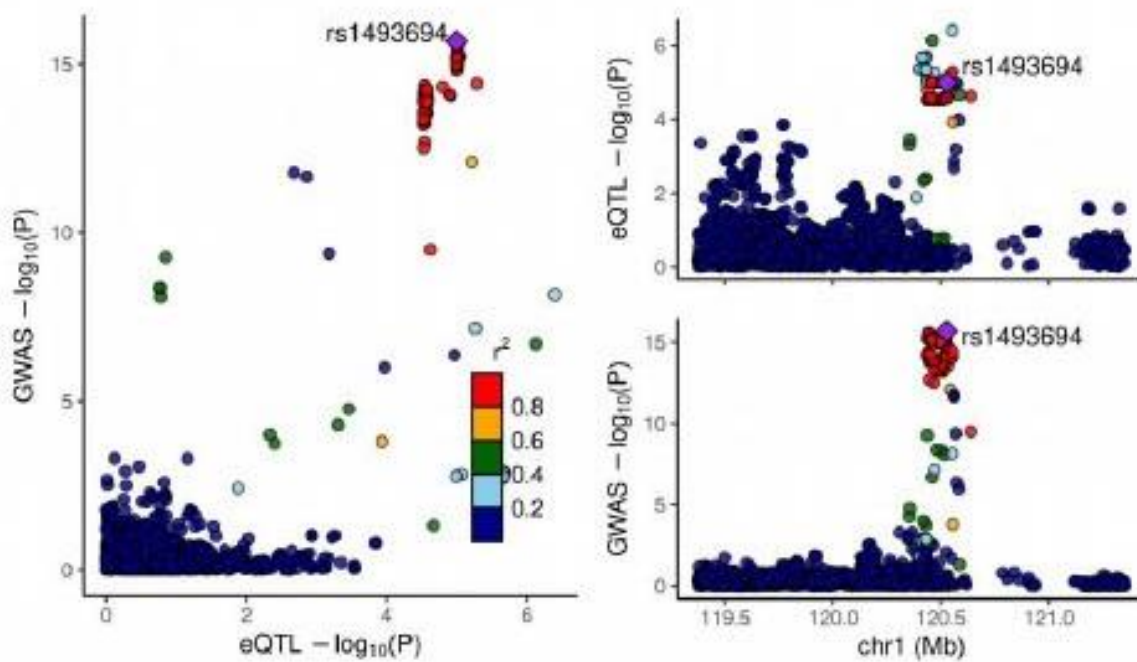
FXVD2

MAF: 0.480 PP.H4.abf: 0.923 SNP.PP.H4: 0.833



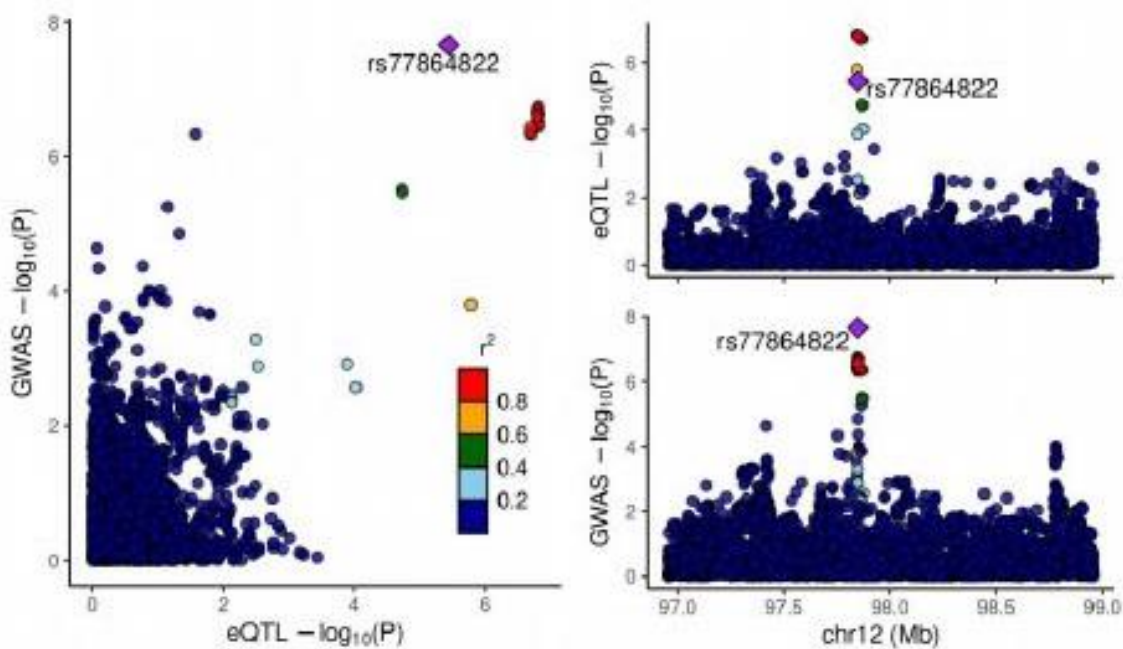
RP5-1042I8.2

MAF: 0.110 PP.H4.abf: 0.806 SNP.PP.H4: 0.092



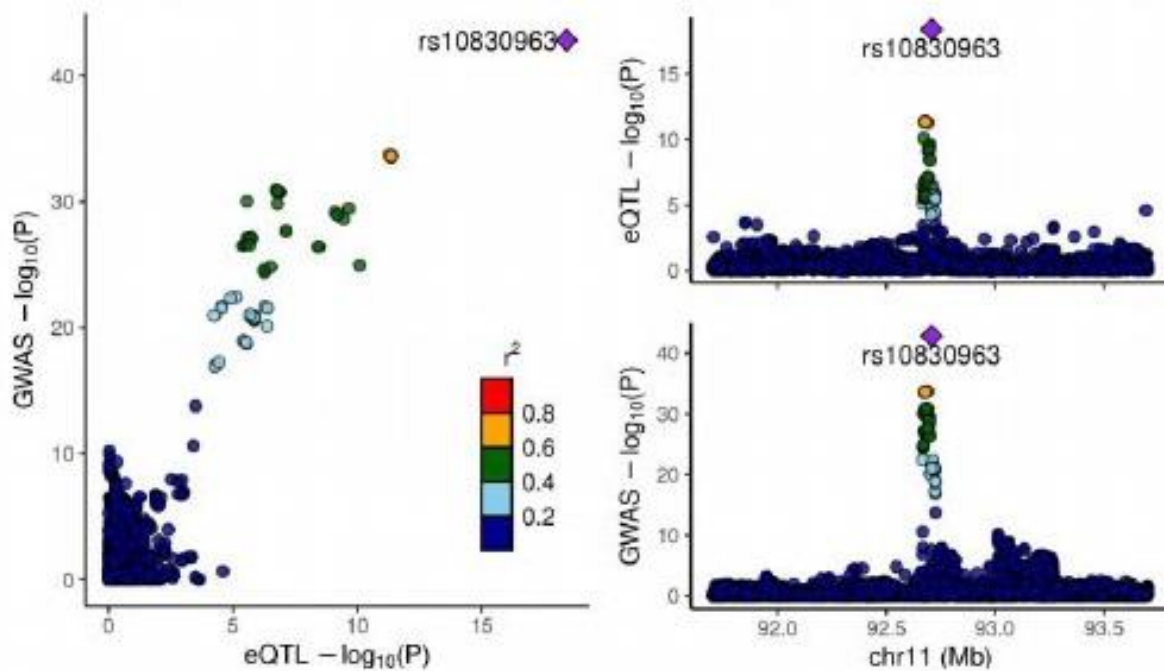
RP11-528M18.2

MAF: 0.070 PP.H4.abf: 0.948 SNP.PP.H4: 0.170



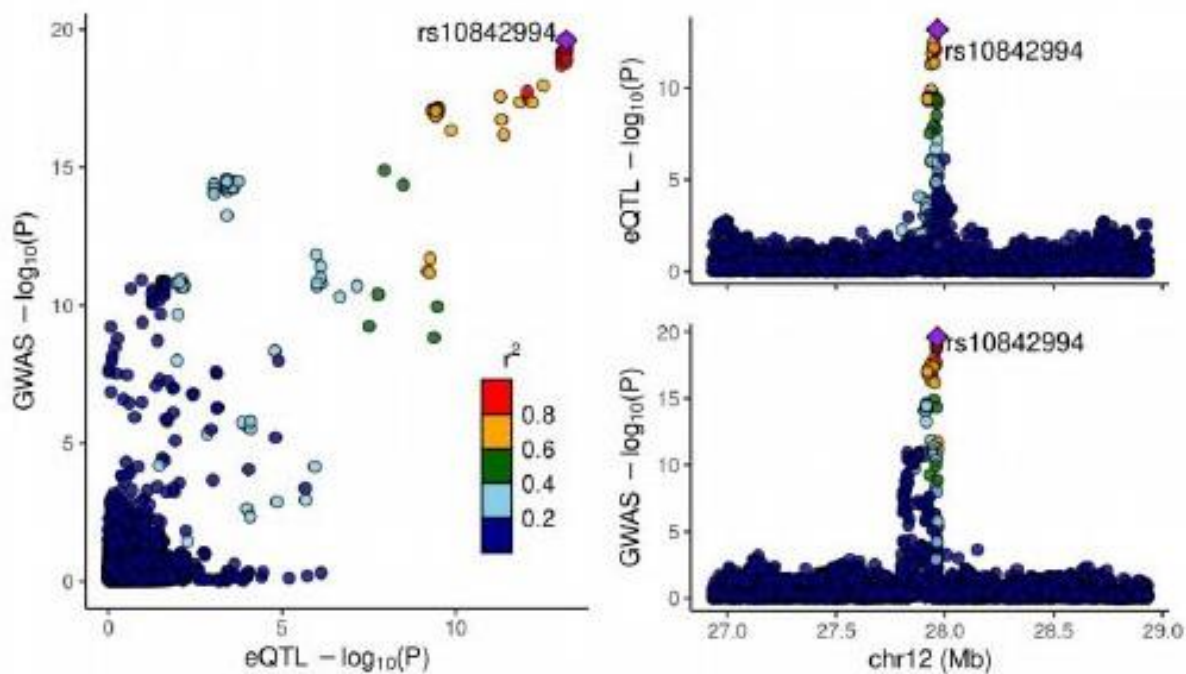
MTNR1B

MAF: 0.280 PP.H4.abf: 1.000 SNP.PP.H4: 1.000



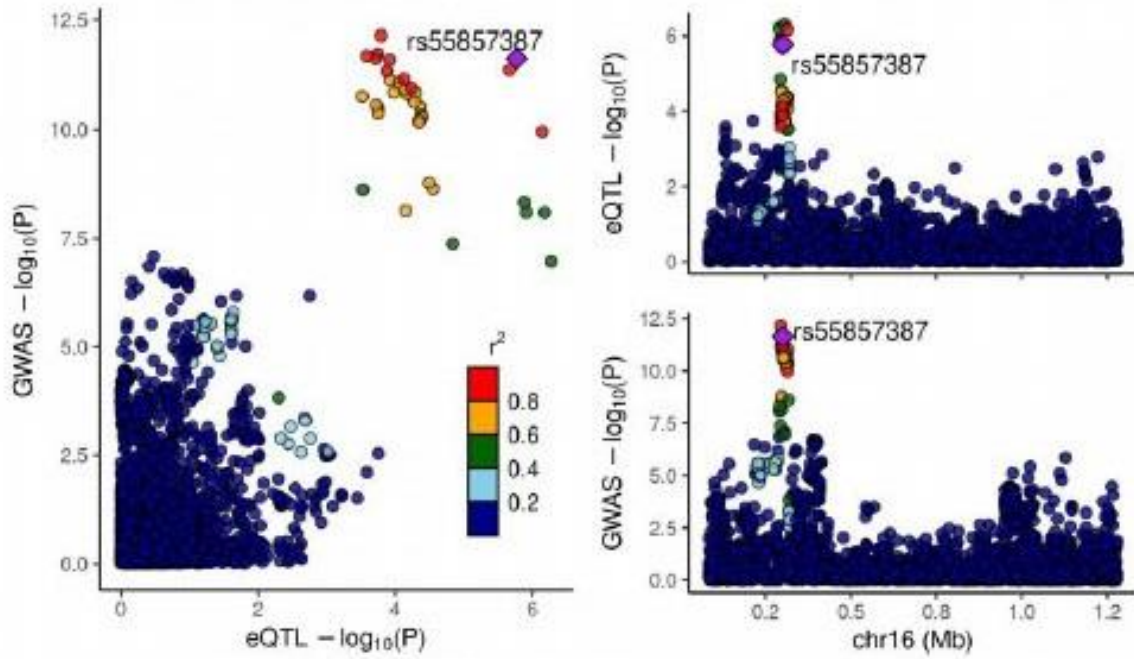
KLHL42

MAF: 0.190 PP.H4.abf: 0.979 SNP.PP.H4: 0.118



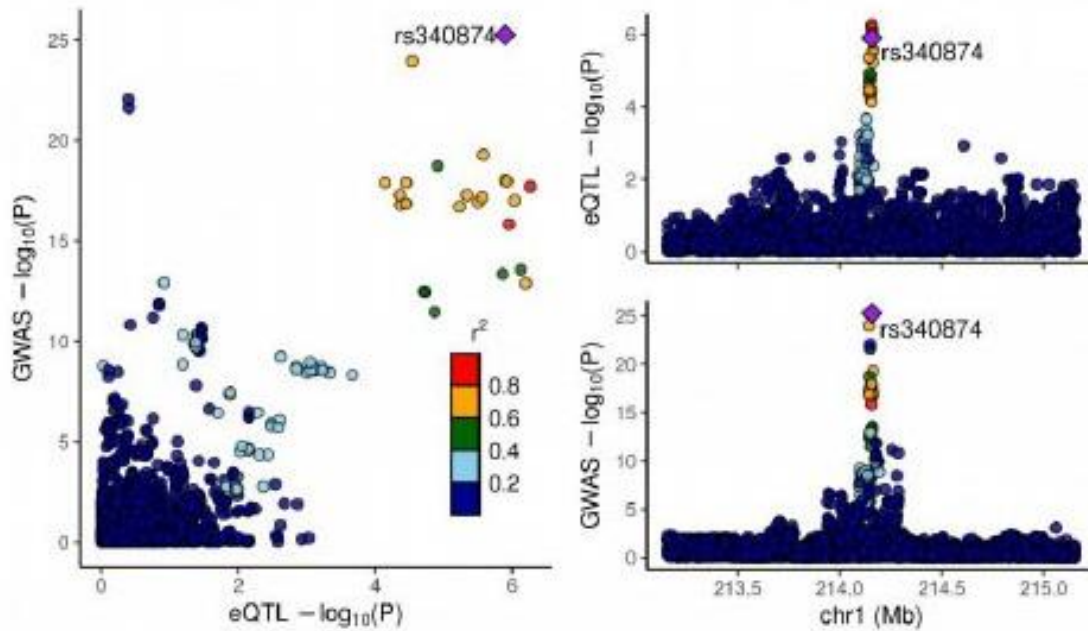
FAM234A

MAF: 0.190 PP.H4.abf: 0.857 SNP.PP.H4: 0.491



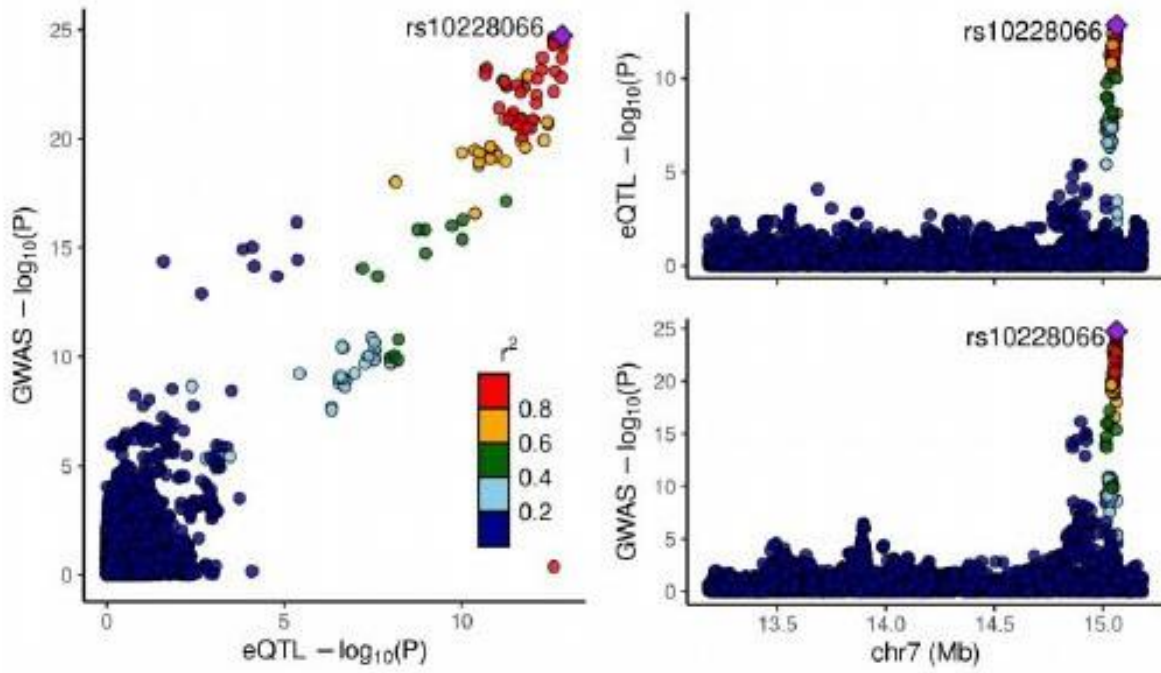
RP11-478J18.2

MAF: 0.440 PP.H4.abf: 0.983 SNP.PP.H4: 0.996



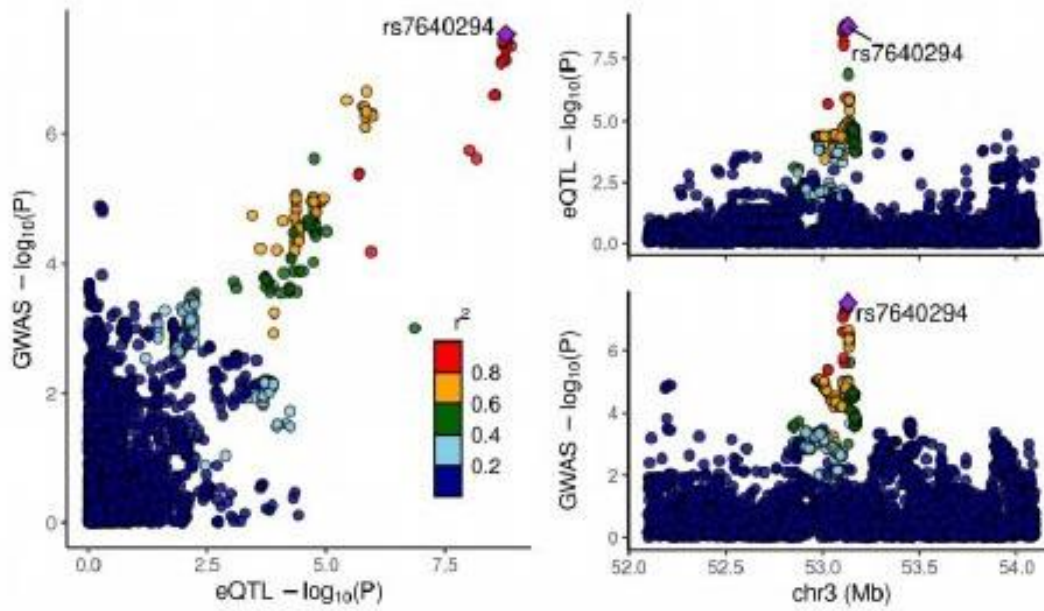
DGKB

MAF: 0.460 PP.H4.abf: 0.973 SNP.PP.H4: 0.148

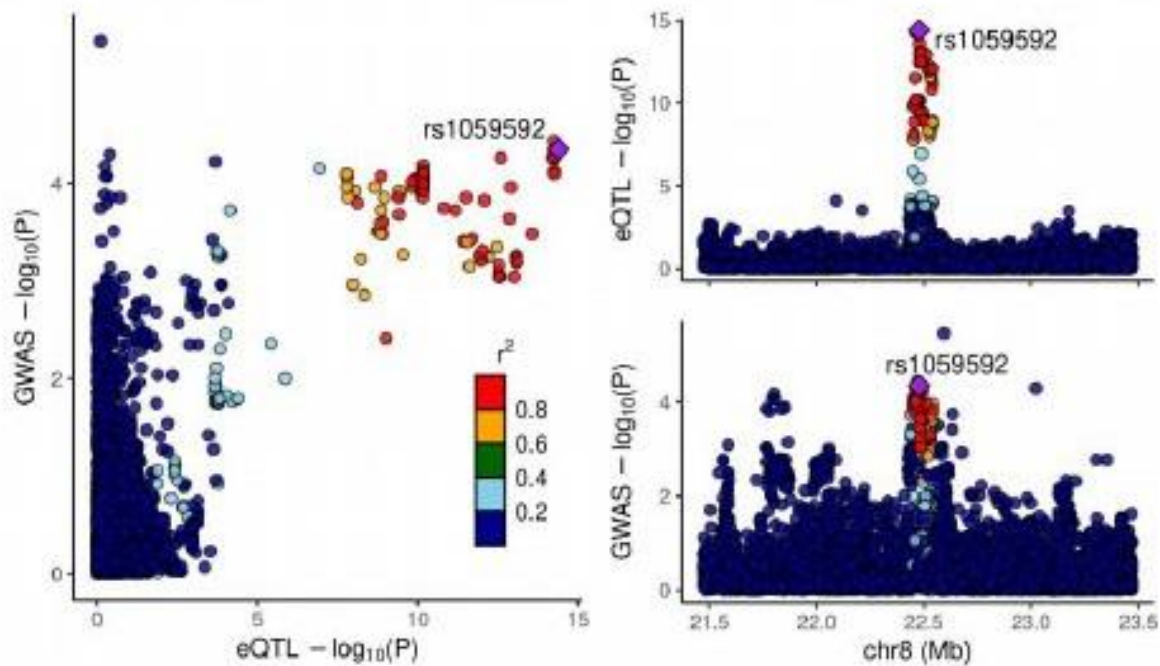


RP11-5017.2

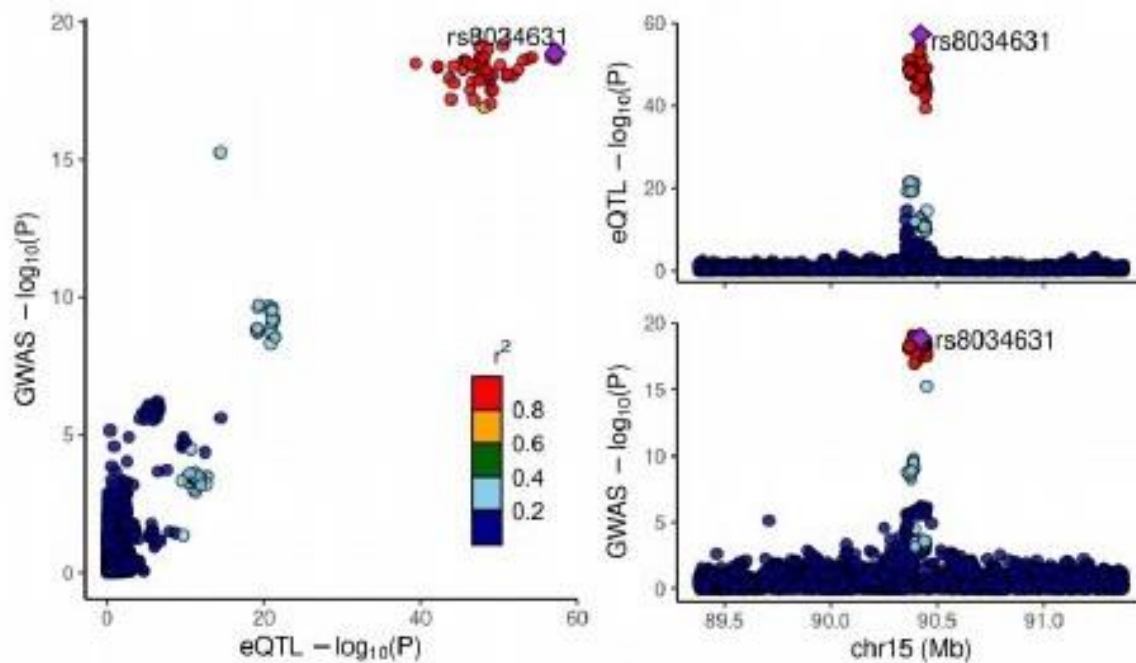
MAF: 0.440 PP.H4.abf: 0.974 SNP.PP.H4: 0.060



RP11-582J16.5
MAF: 0.350 PP.H4.abf: 0.806 SNP.PP.H4: 0.123

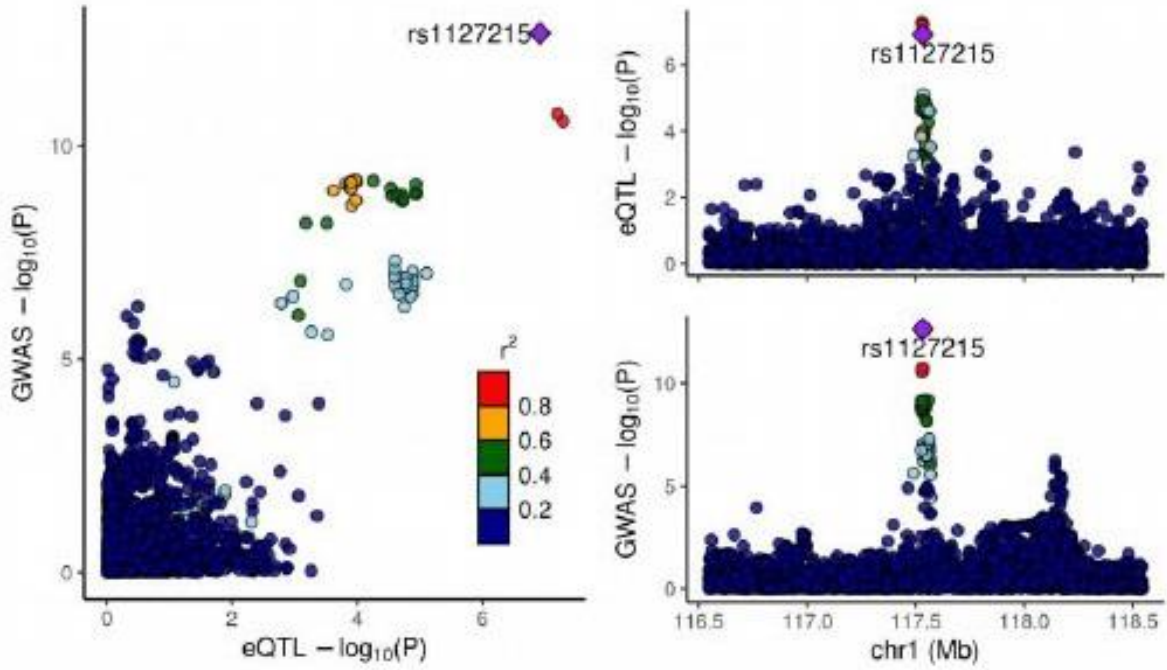


C15orf38-AP3S2
MAF: 0.280 PP.H4.abf: 0.974 SNP.PP.H4: 0.521



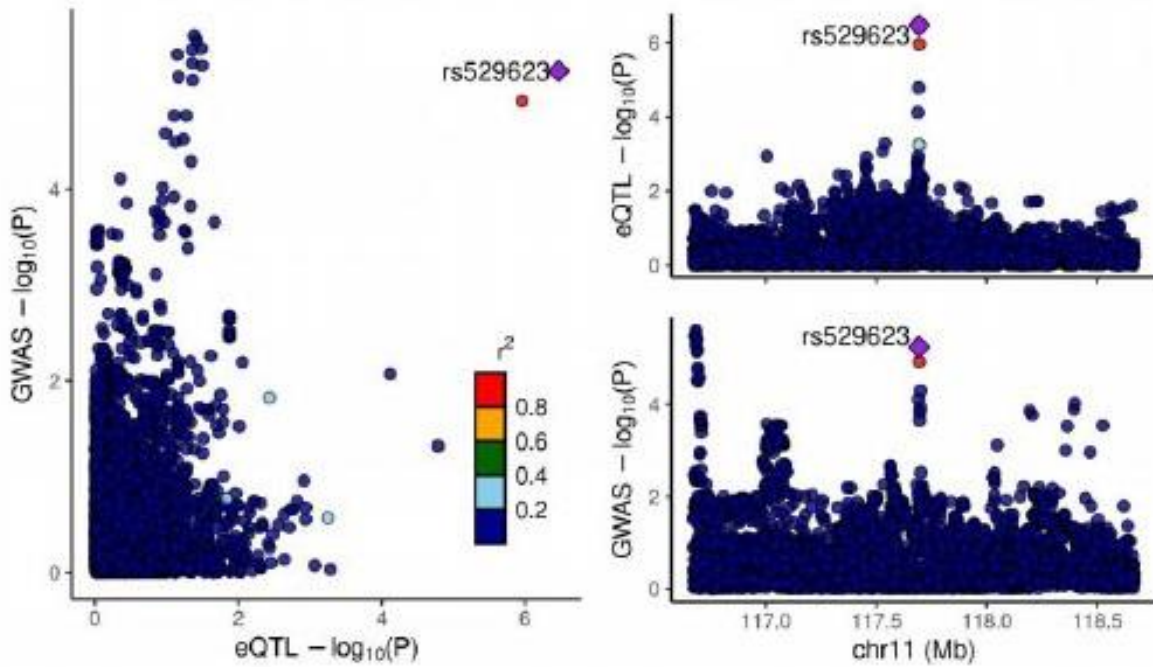
CD101

MAF: 0.420 PP.H4.abf: 0.995 SNP.PP.H4: 0.959



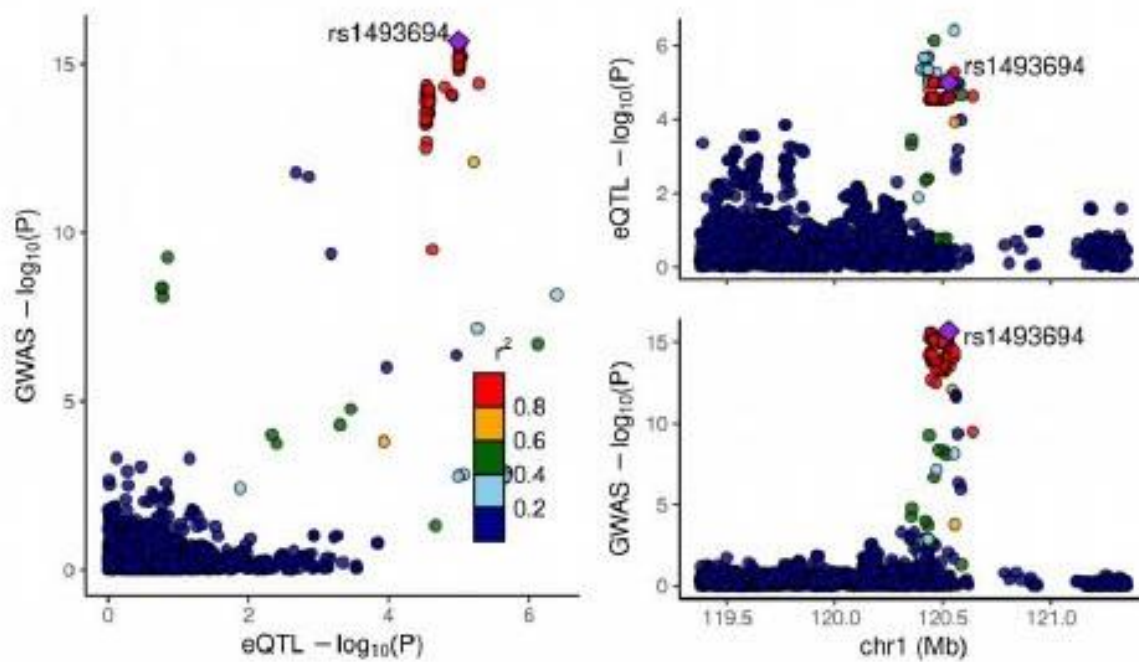
FXD2

MAF: 0.480 PP.H4.abf: 0.923 SNP.PP.H4: 0.833



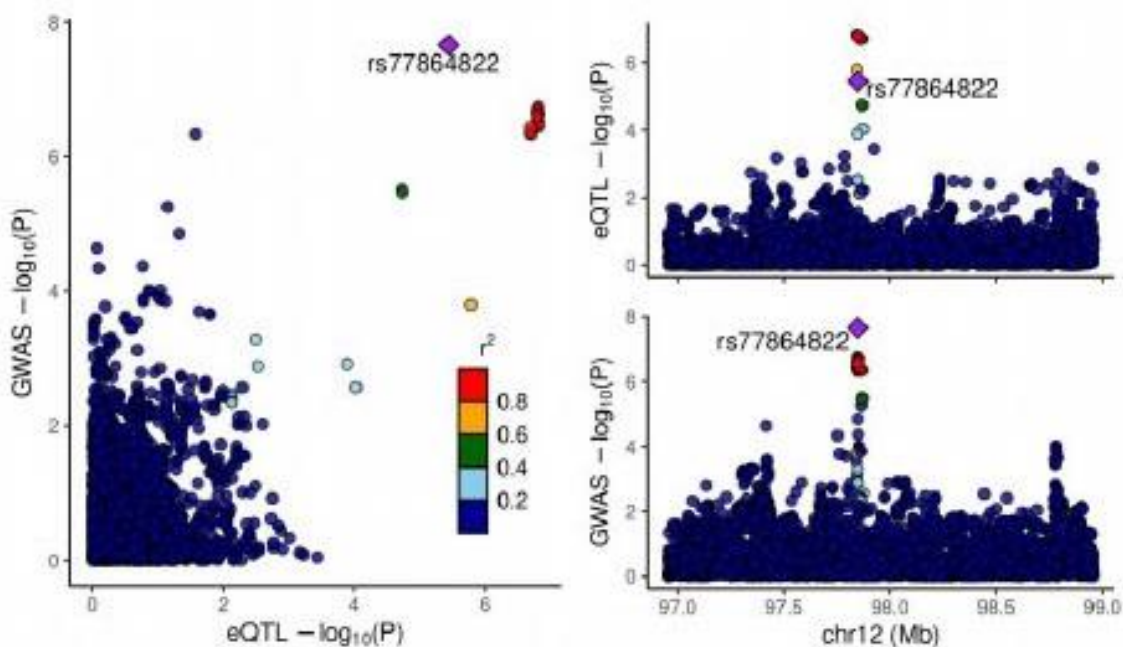
RP5-1042I8.2

MAF: 0.110 PP.H4.abf: 0.806 SNP.PP.H4: 0.092



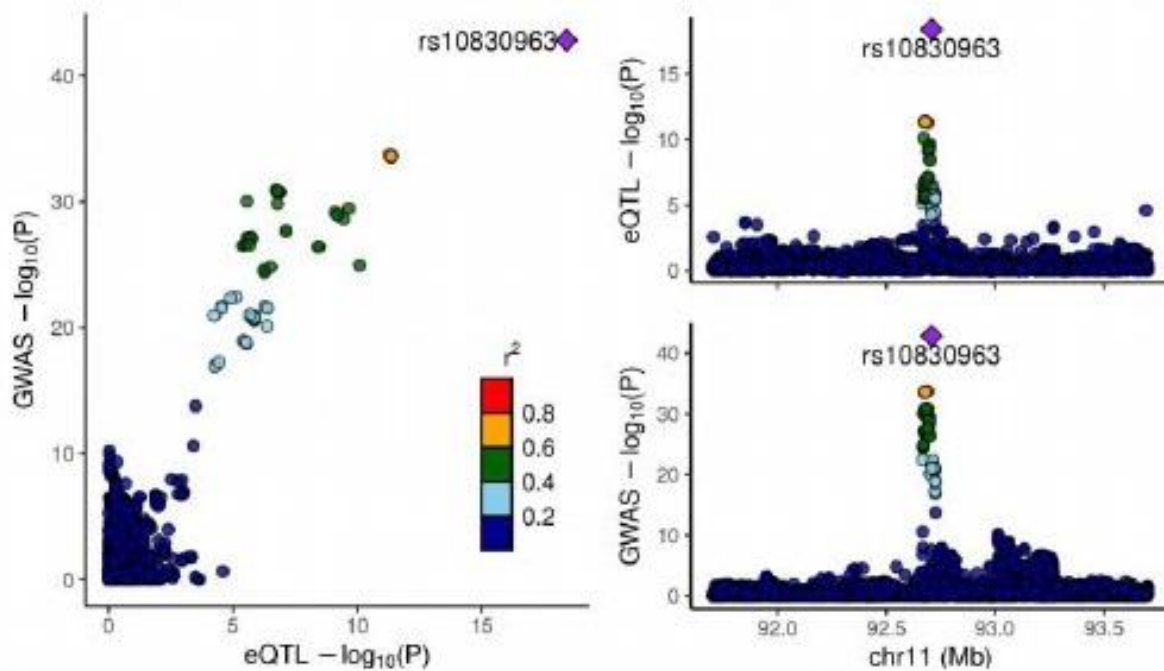
RP11-528M18.2

MAF: 0.070 PP.H4.abf: 0.948 SNP.PP.H4: 0.170



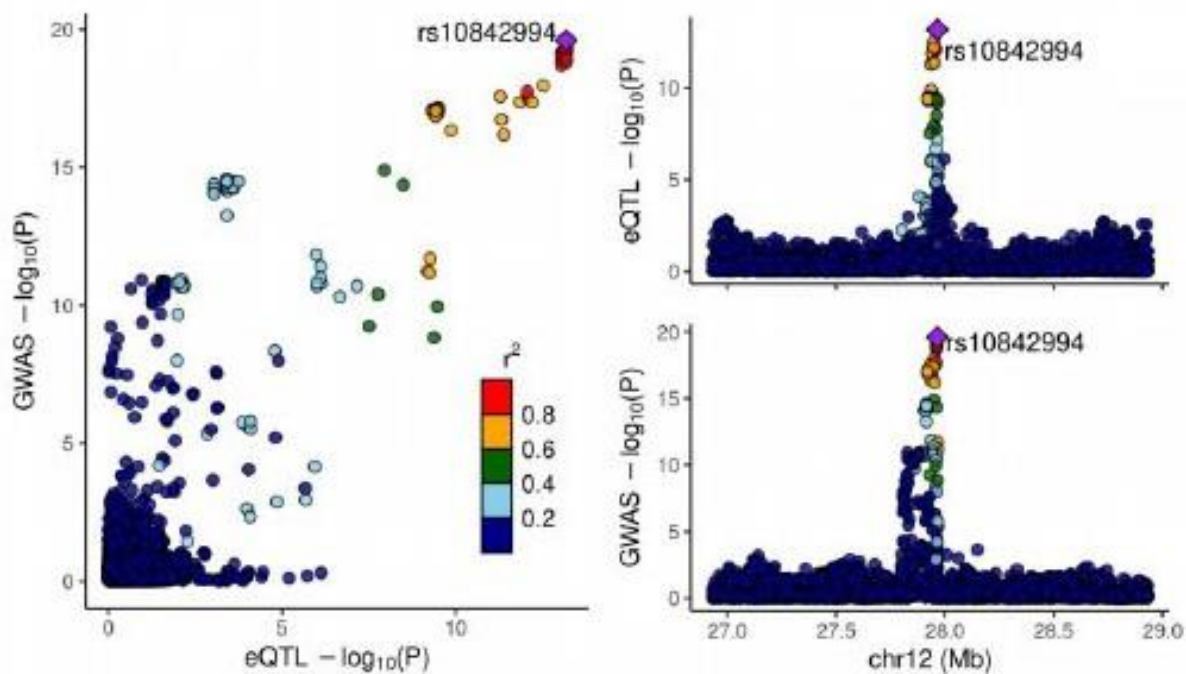
MTNR1B

MAF: 0.280 PP.H4.abf: 1.000 SNP.PP.H4: 1.000



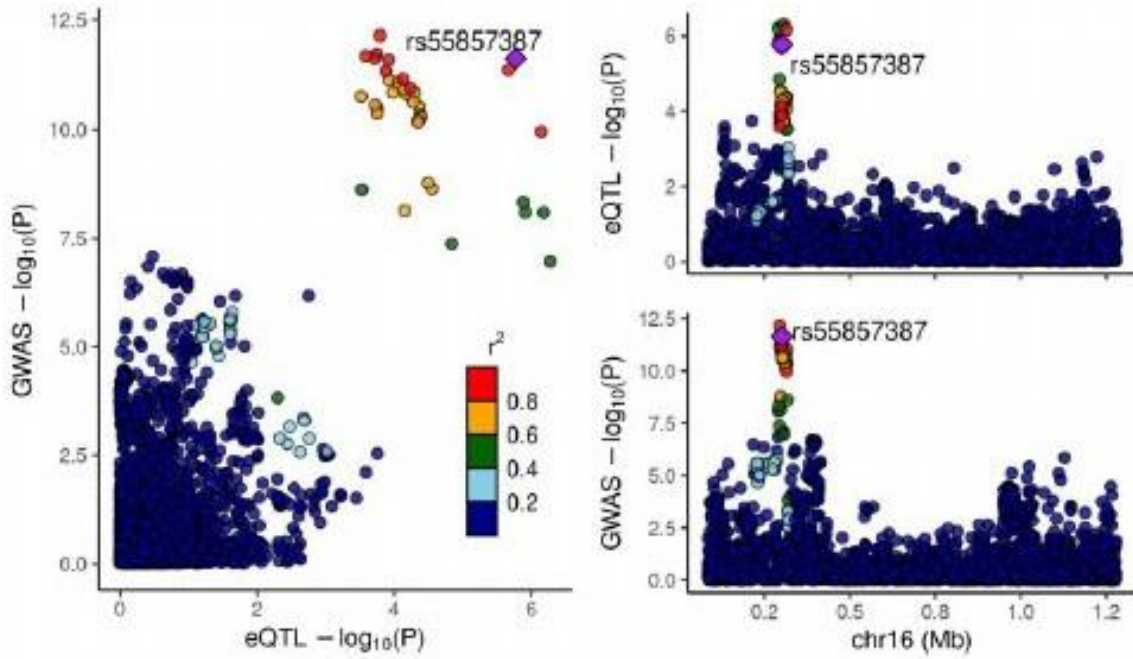
KLHL42

MAF: 0.190 PP.H4.abf: 0.979 SNP.PP.H4: 0.118



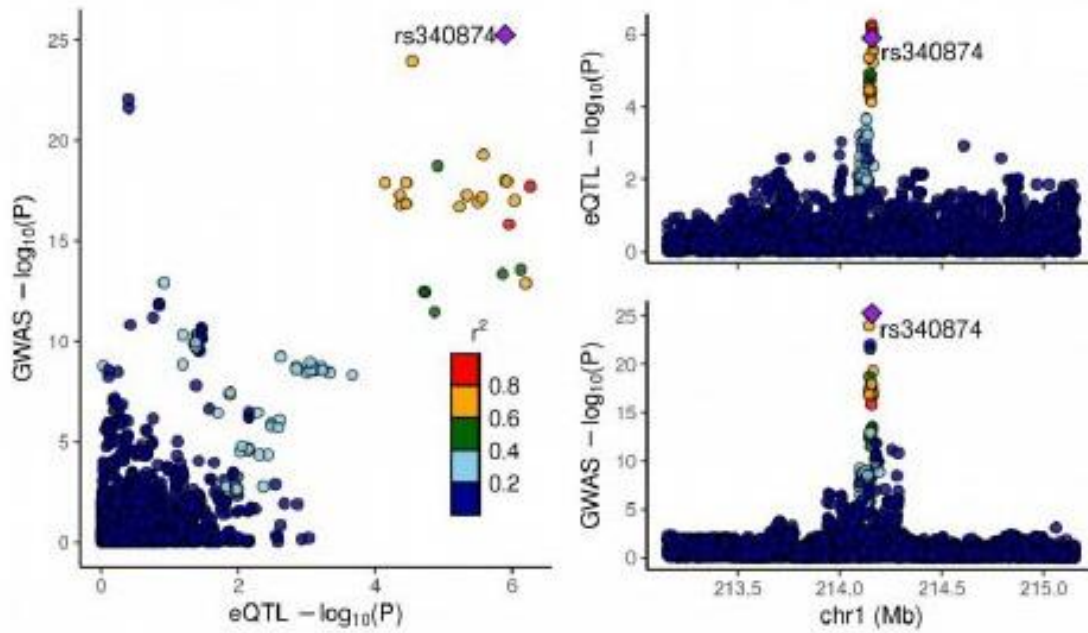
FAM234A

MAF: 0.190 PP.H4.abf: 0.857 SNP.PP.H4: 0.491



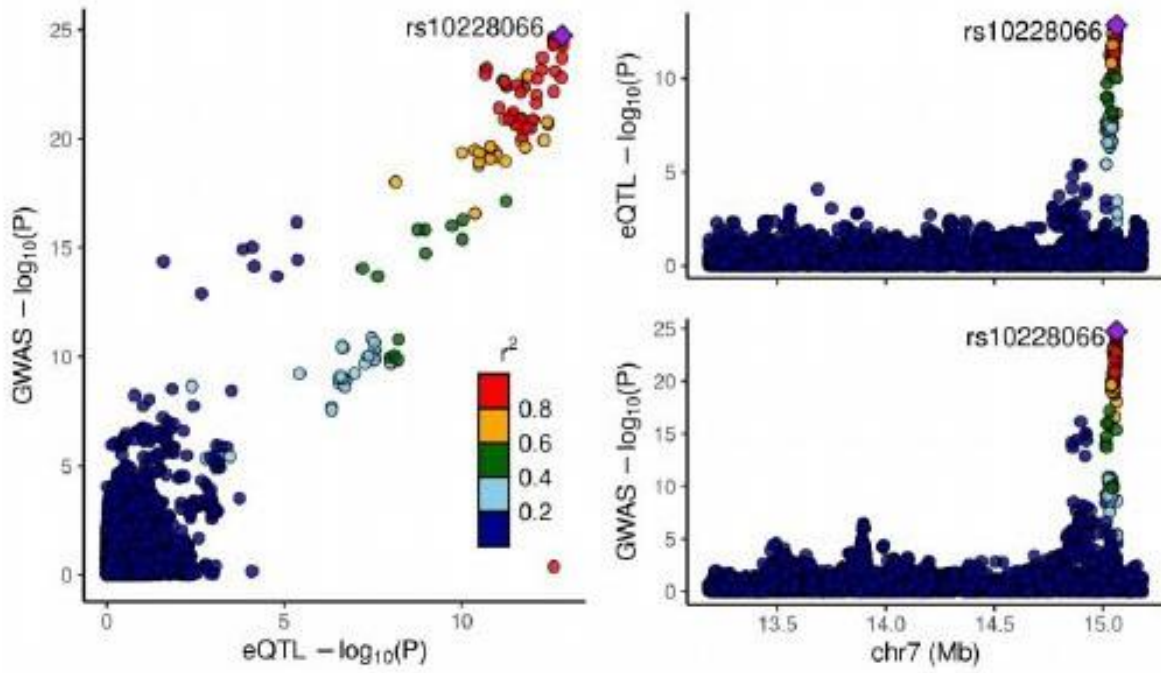
RP11-478J18.2

MAF: 0.440 PP.H4.abf: 0.983 SNP.PP.H4: 0.996



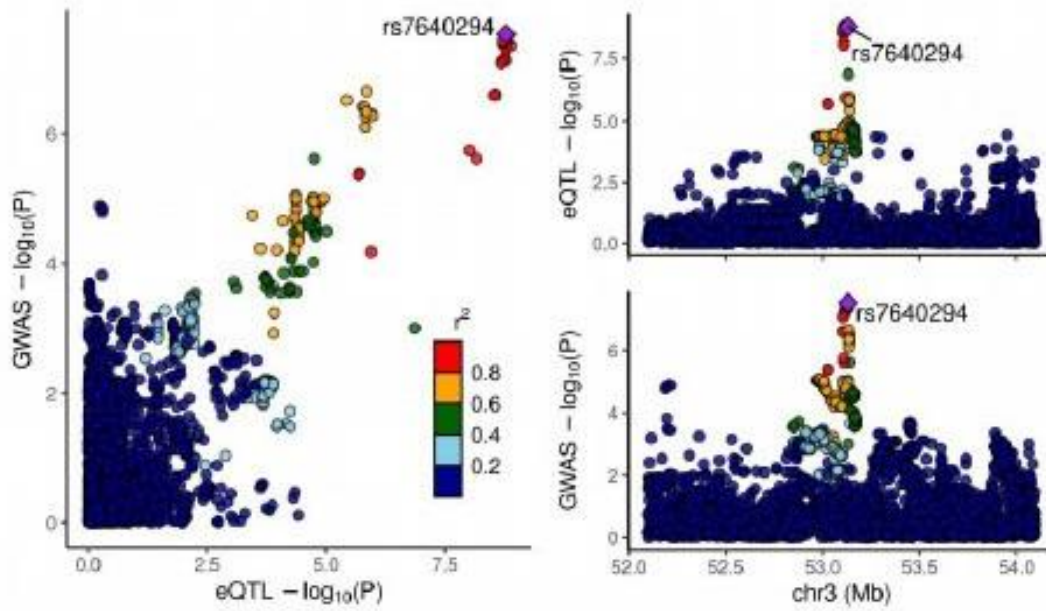
DGKB

MAF: 0.460 PP.H4.abf: 0.973 SNP.PP.H4: 0.148



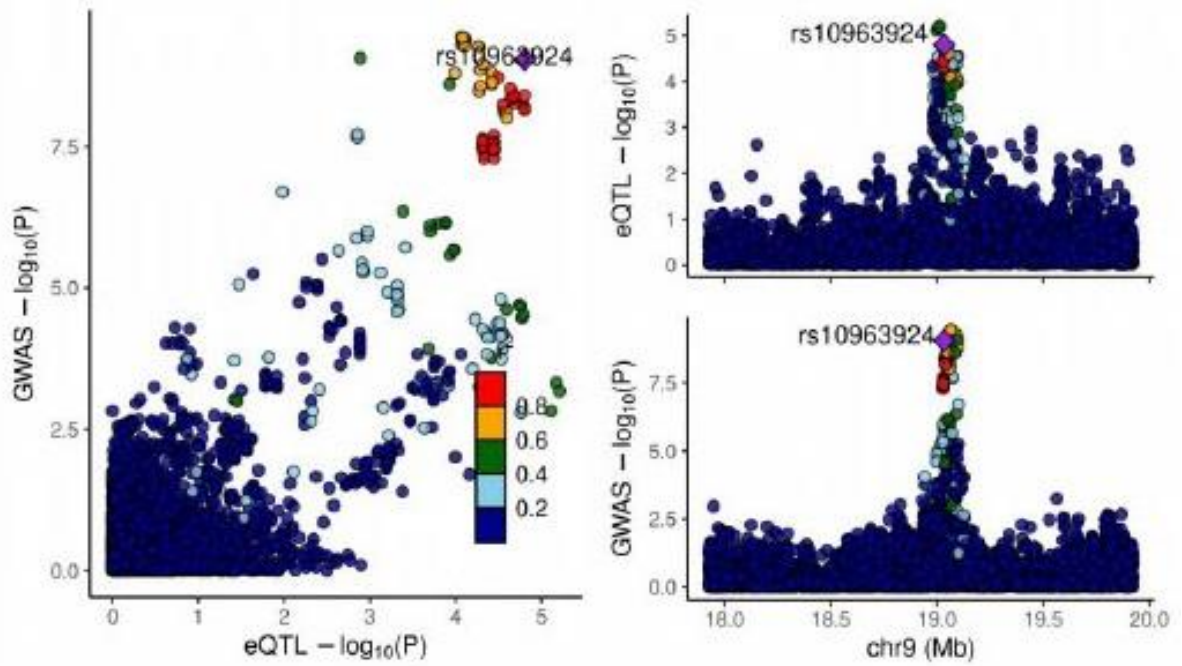
RP11-5017.2

MAF: 0.440 PP.H4.abf: 0.974 SNP.PP.H4: 0.060



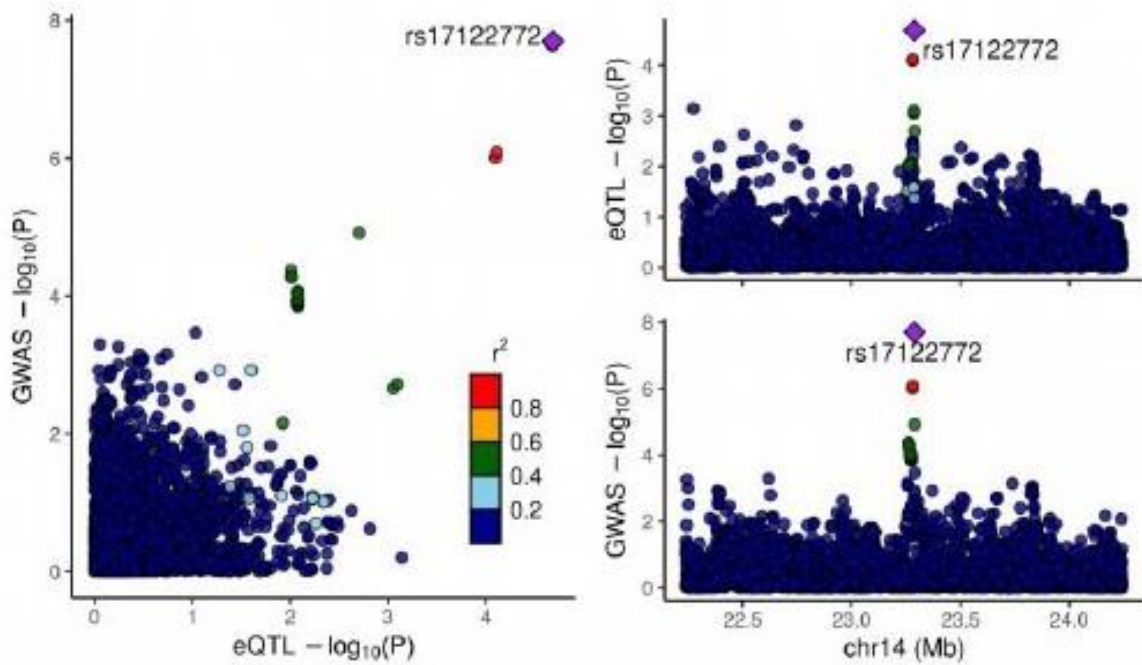
SAXO1

MAF: 0.430 PP.H4.abf: 0.825 SNP.PP.H4: 0.091



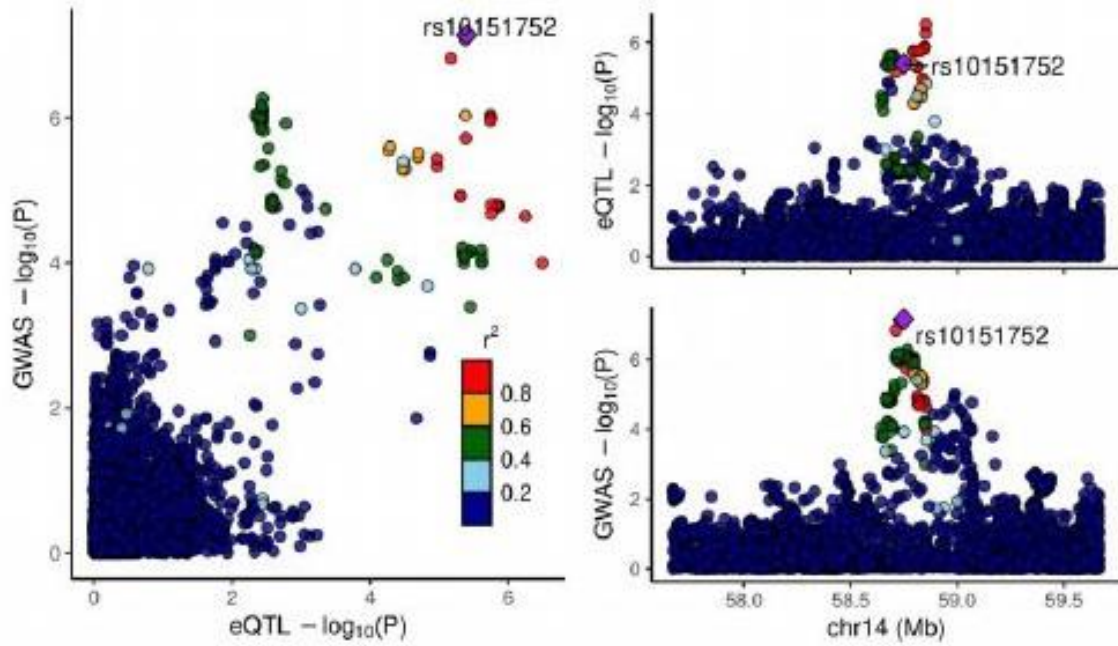
SLC7A7

MAF: 0.230 PP.H4.abf: 0.964 SNP.PP.H4: 0.273



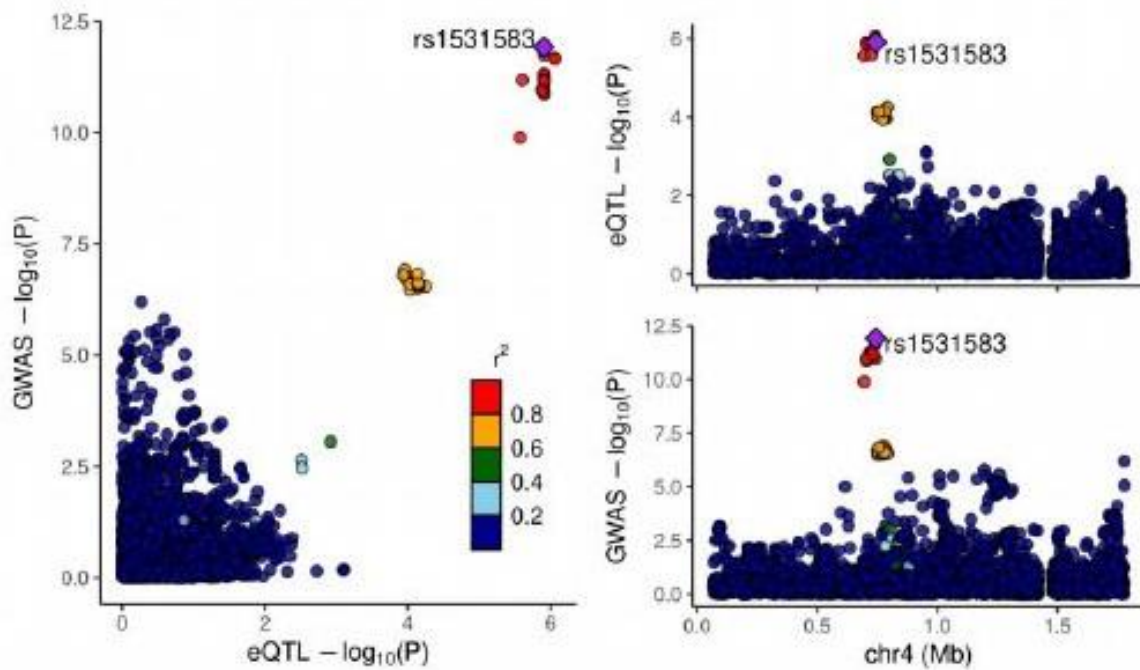
ACTR10

MAF: 0.410 PP.H4.abf: 0.863 SNP.PP.H4: 0.256



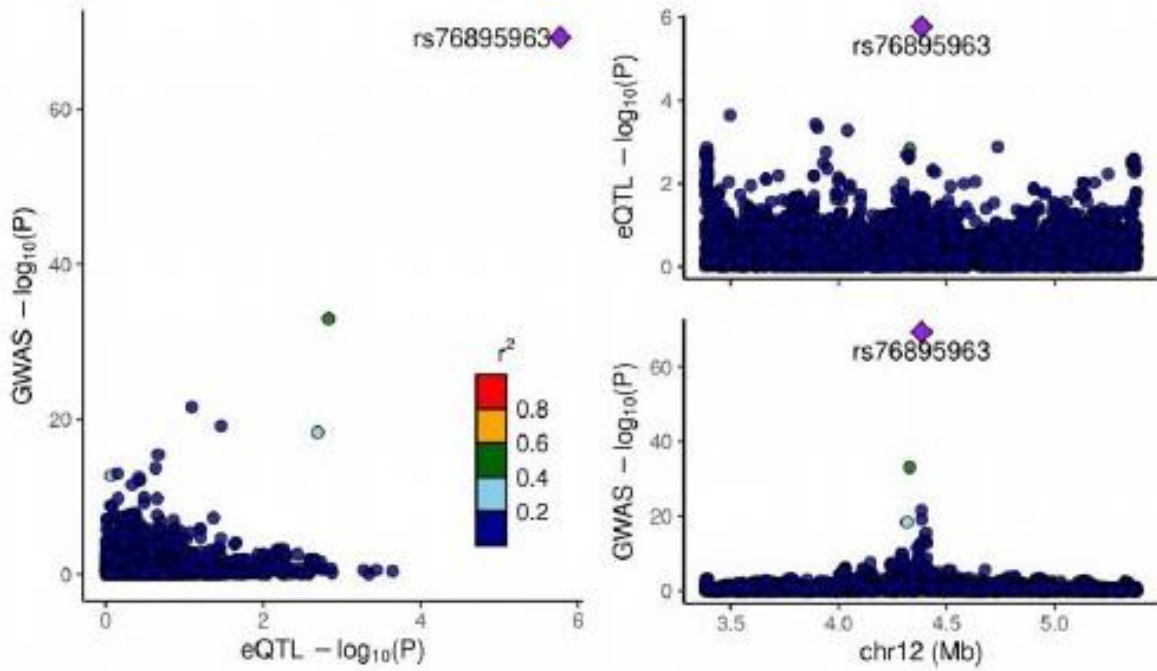
CPLX1

MAF: 0.046 PP.H4.abf: 0.875 SNP.PP.H4: 0.133



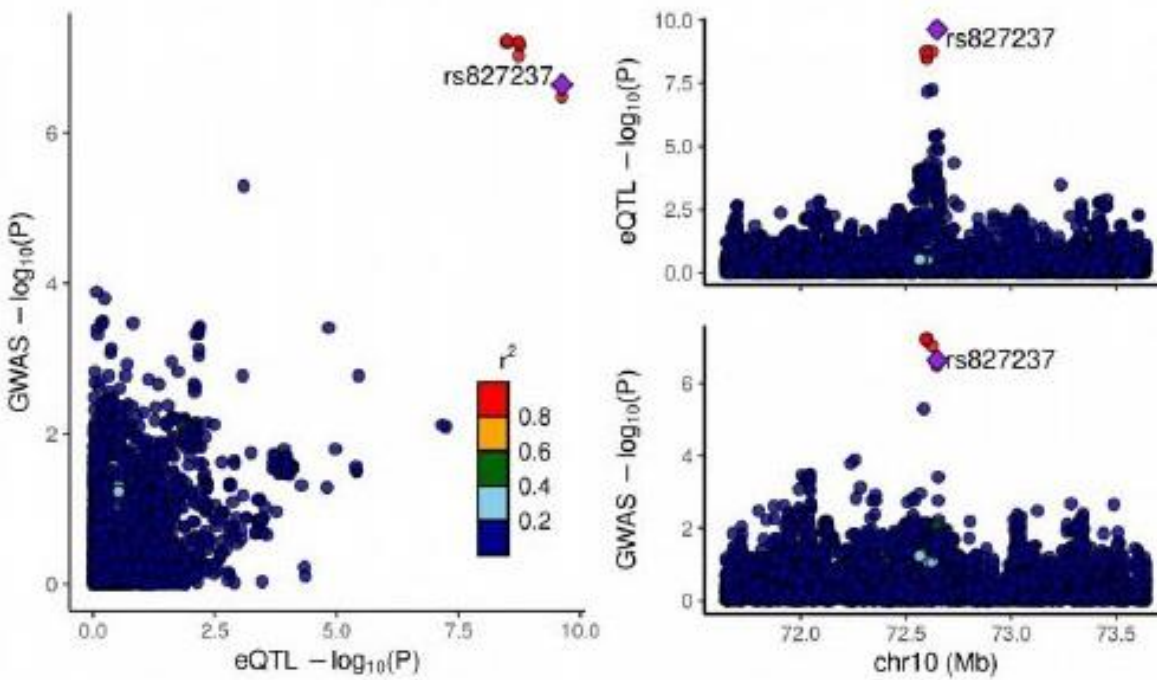
CCND2

MAF: 0.020 PP.H4.abf: 0.359 SNP.PP.H4: 1.000



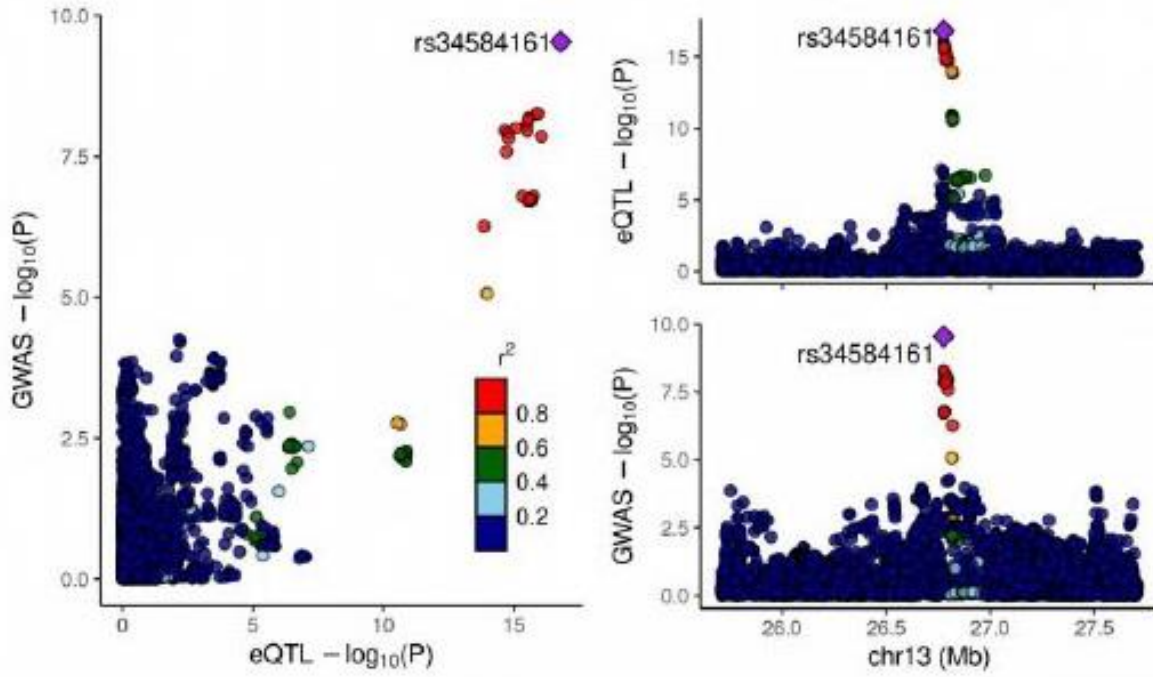
PCBD1

MAF: 0.210 PP.H4.abf: 0.988 SNP.PP.H4: 0.189



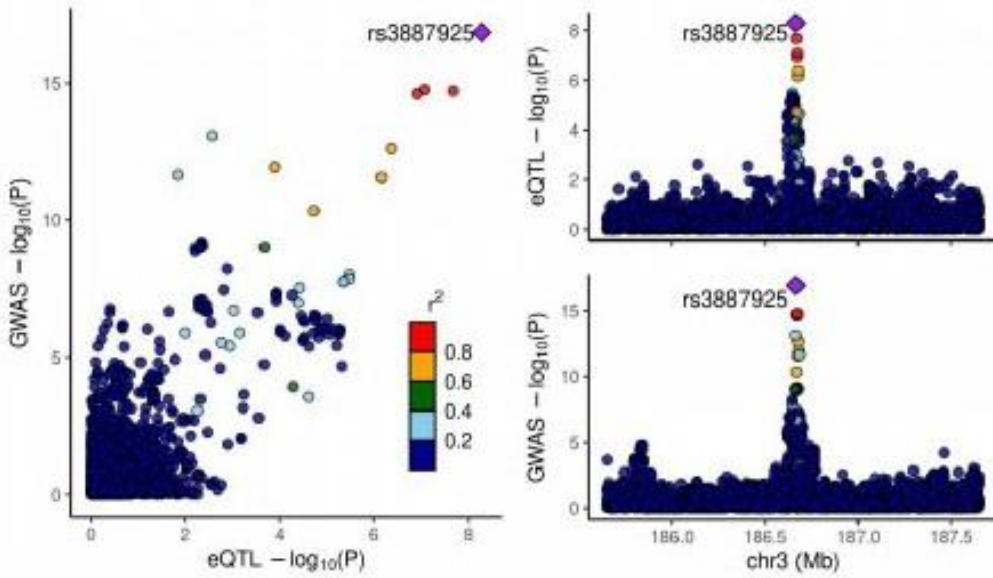
RNF6

MAF: 0.240 PP.H4.abf: 0.996 SNP.PP.H4: 0.946



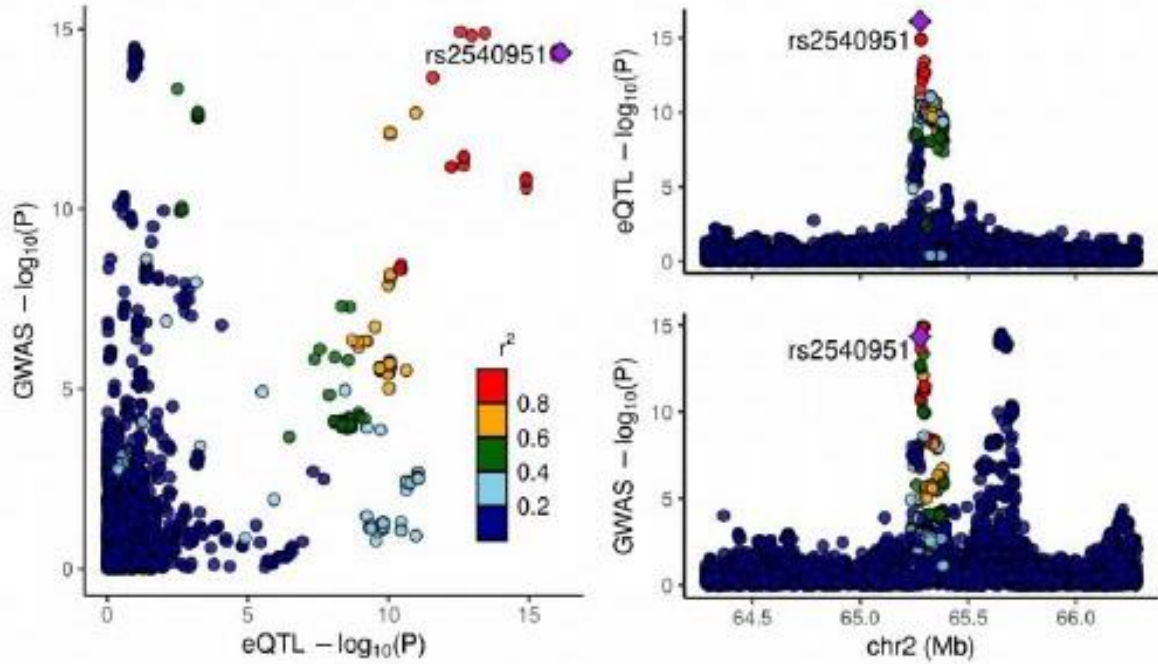
ENSG00000203632

MAF: 0.450 PP.H4.abf: 0.999 SNP.PP.H4: 0.996



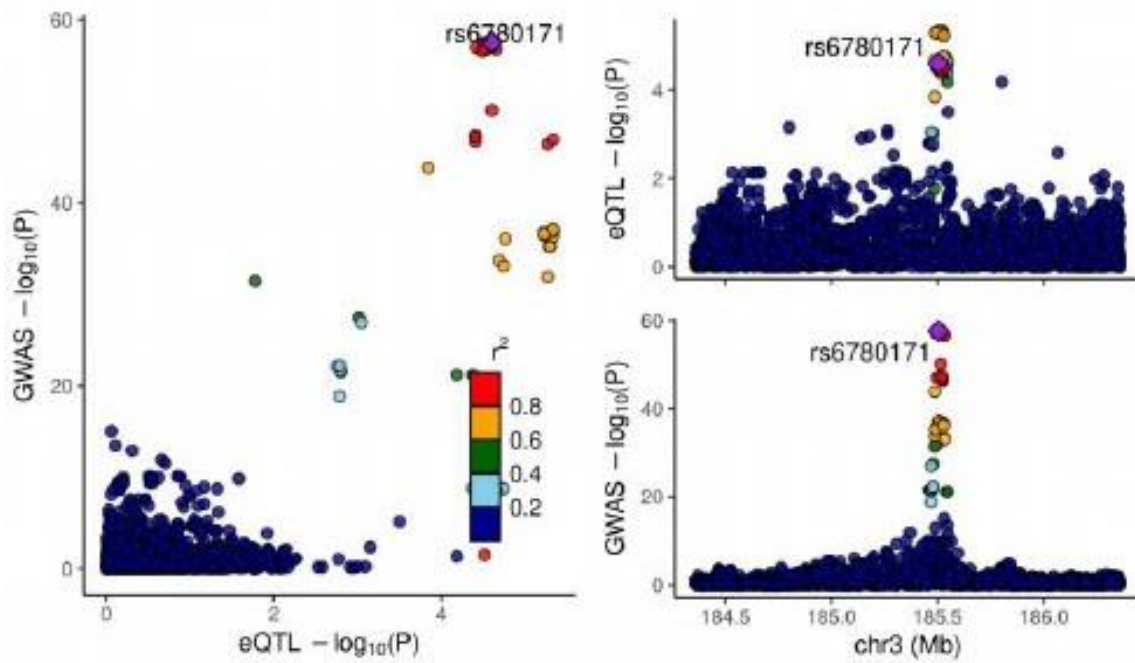
CEP68

MAF: 0.380 PP.H4.abf: 0.967 SNP,PP.H4: 0.238



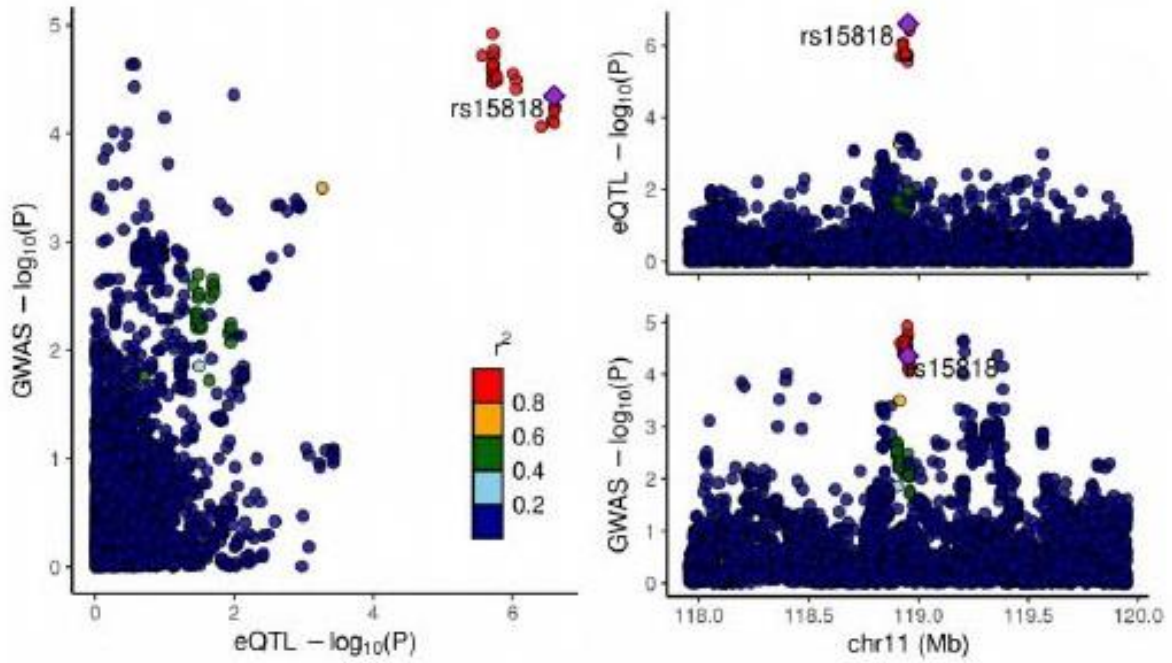
IGF2BP2

MAF: 0.310 PP.H4.abf: 0.881 SNP,PP.H4: 0.069



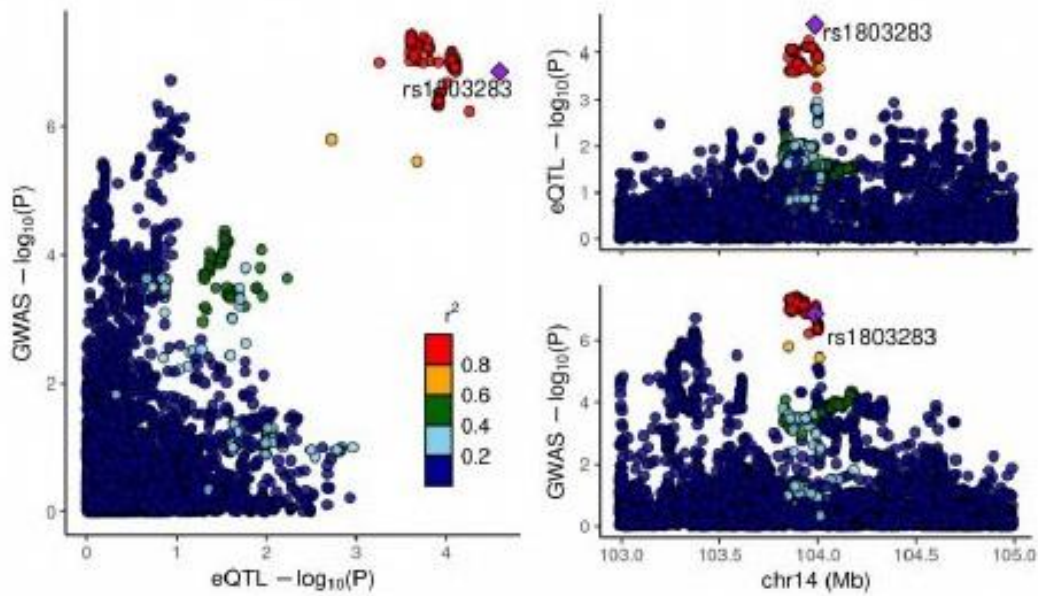
HMBS

MAF: 0.400 PP.H4.abf: 0.844 SNP.PP.H4: 0.055

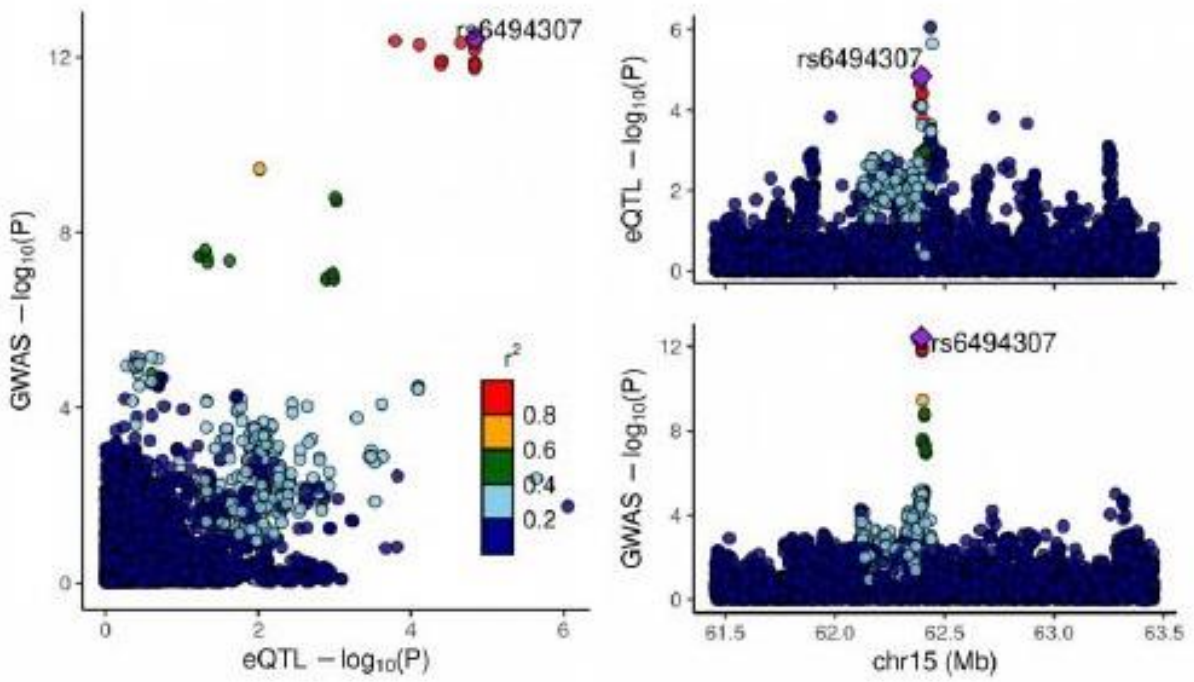


RP11-600F24.7

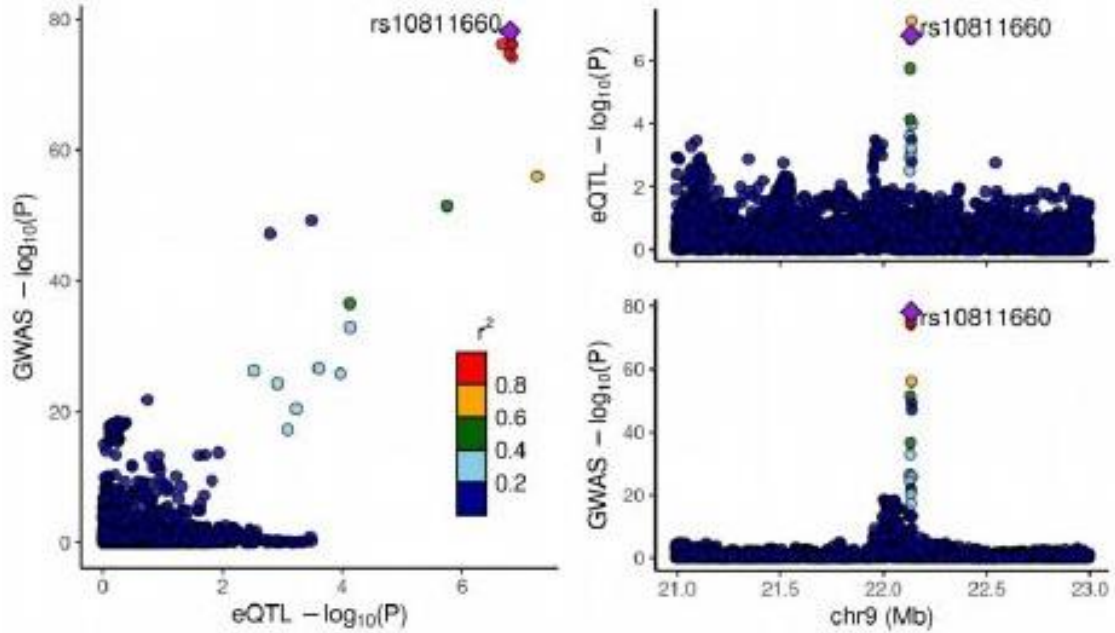
MAF: 0.350 PP.H4.abf: 0.812 SNP.PP.H4: 0.024

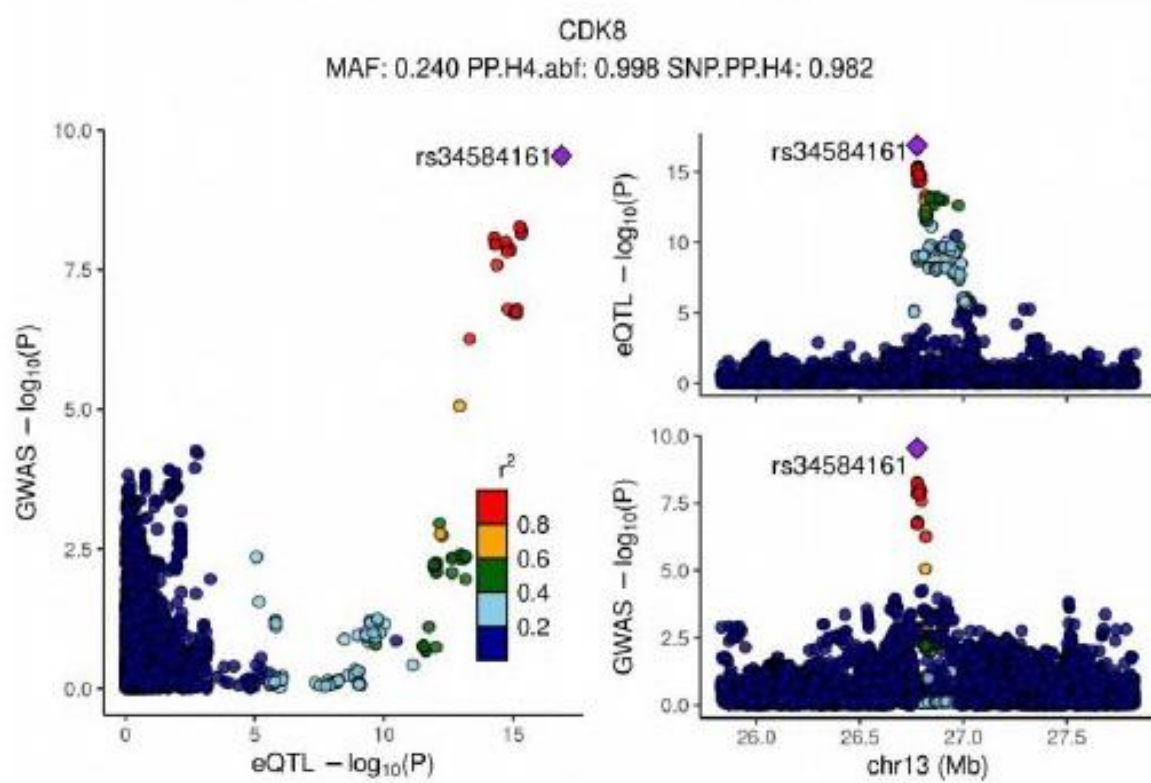
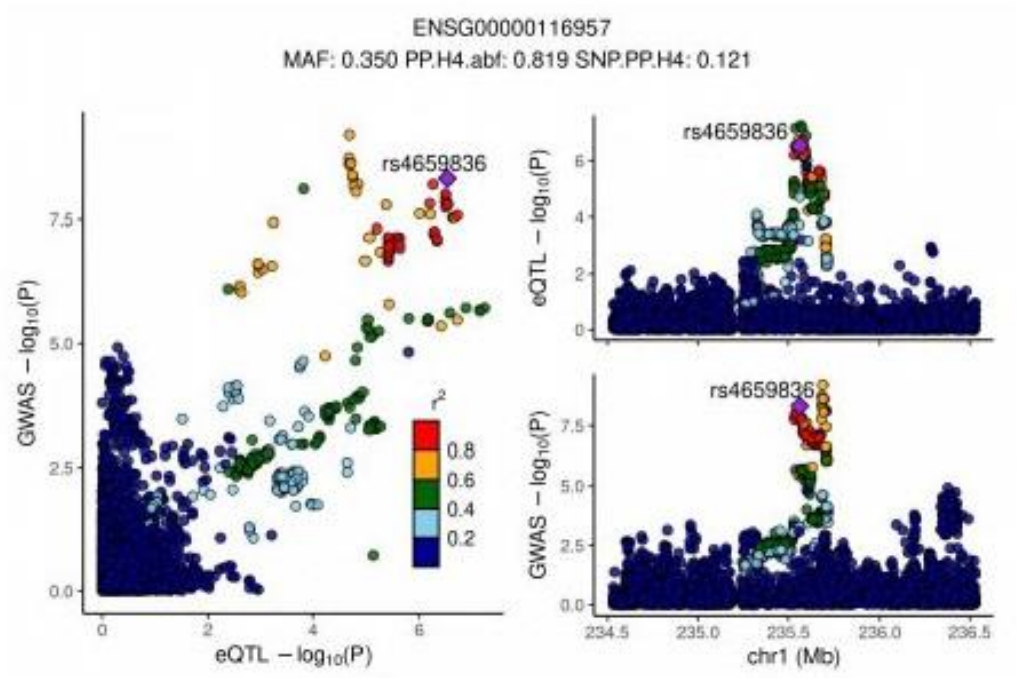


C2CD4B
MAF: 0.430 PP.H4.abf: 0.942 SNP.PP.H4: 0.157



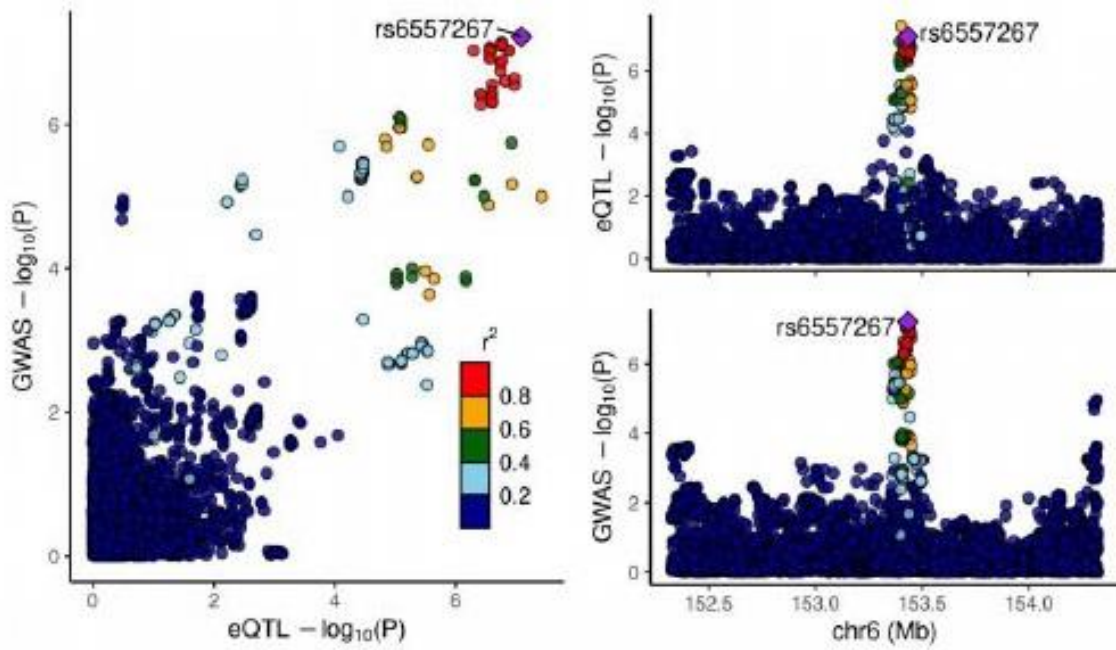
CDKN2B-AS1
MAF: 0.170 PP.H4.abf: 0.990 SNP.PP.H4: 0.476





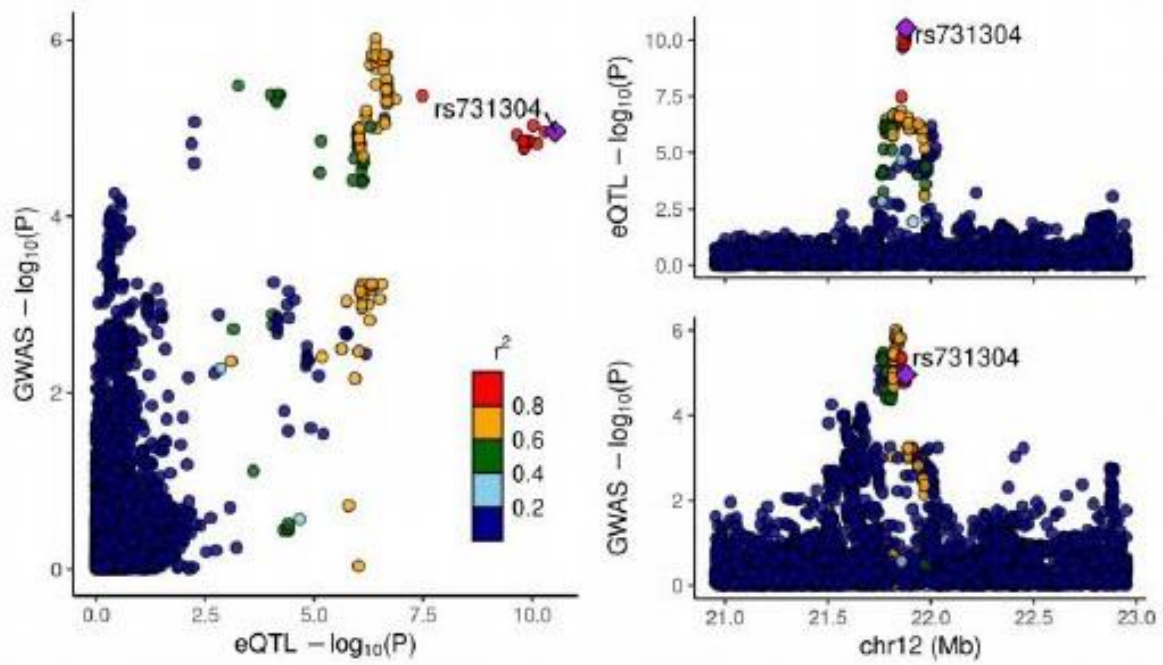
RGS17

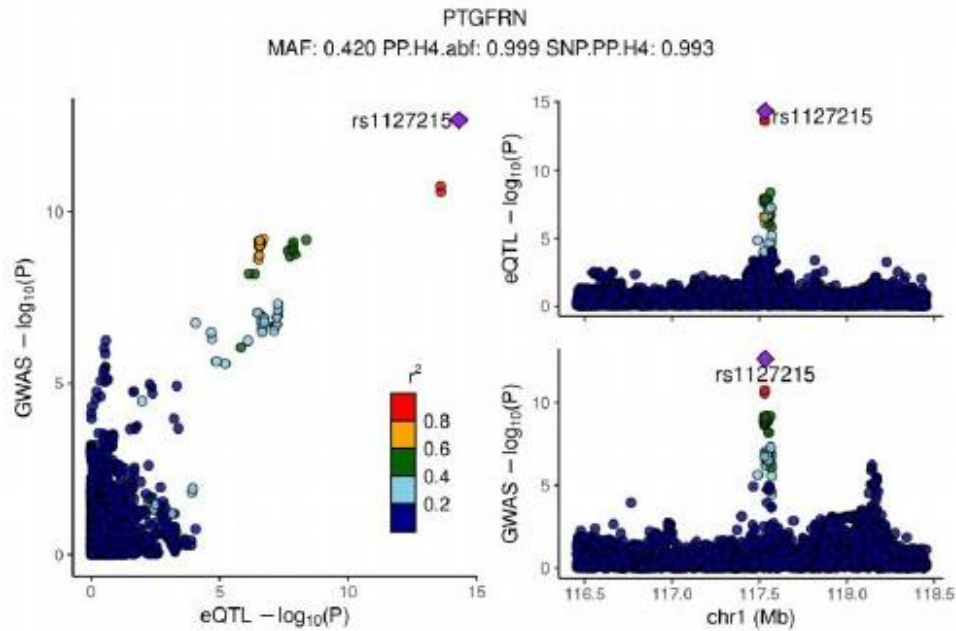
MAF: 0.420 PP.H4.abf: 0.937 SNP.PP.H4: 0.076



ABCC9

MAF: 0.240 PP.H4.abf: 0.801 SNP.PP.H4: 0.191

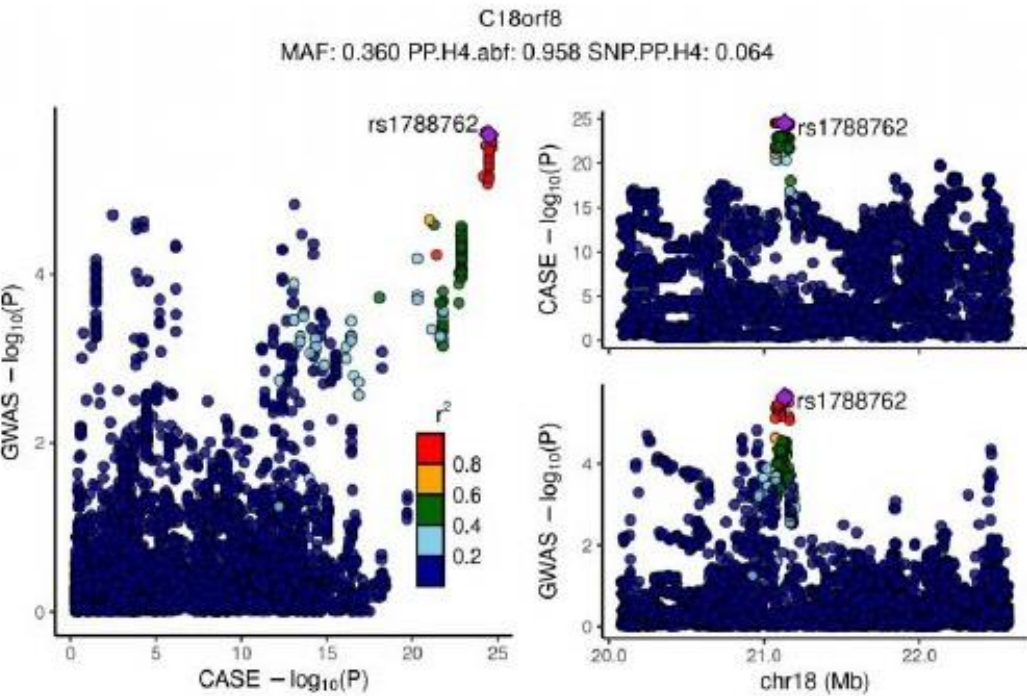
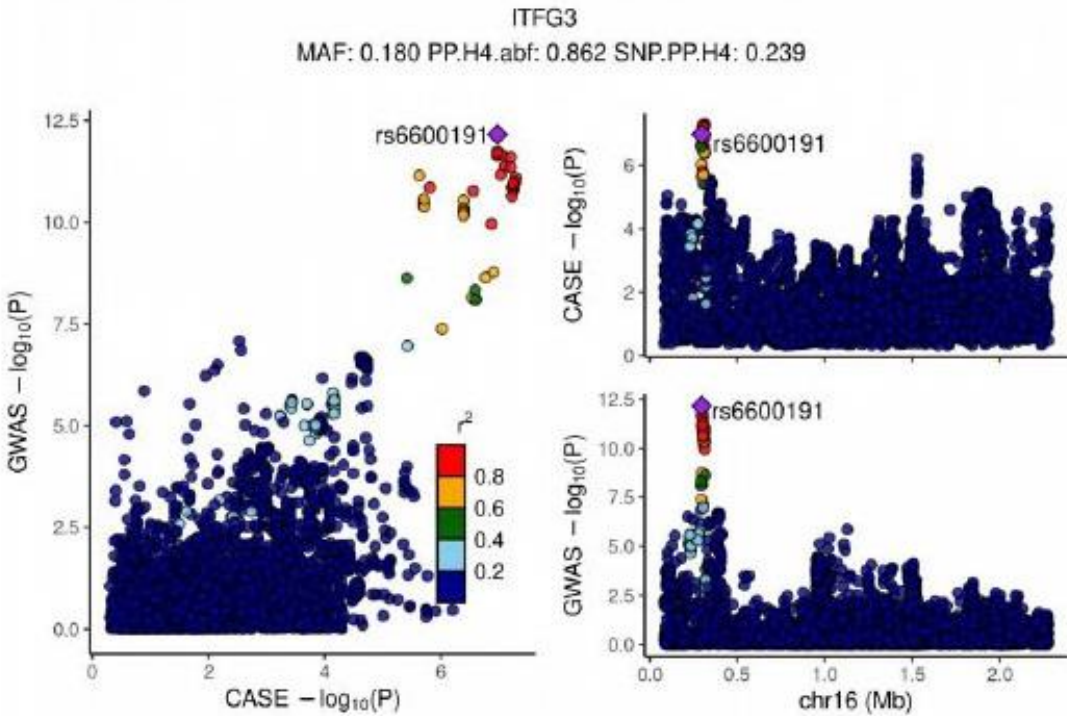




Colocalization plots of eQTL signals, related to Table 1 and STAR Methods.

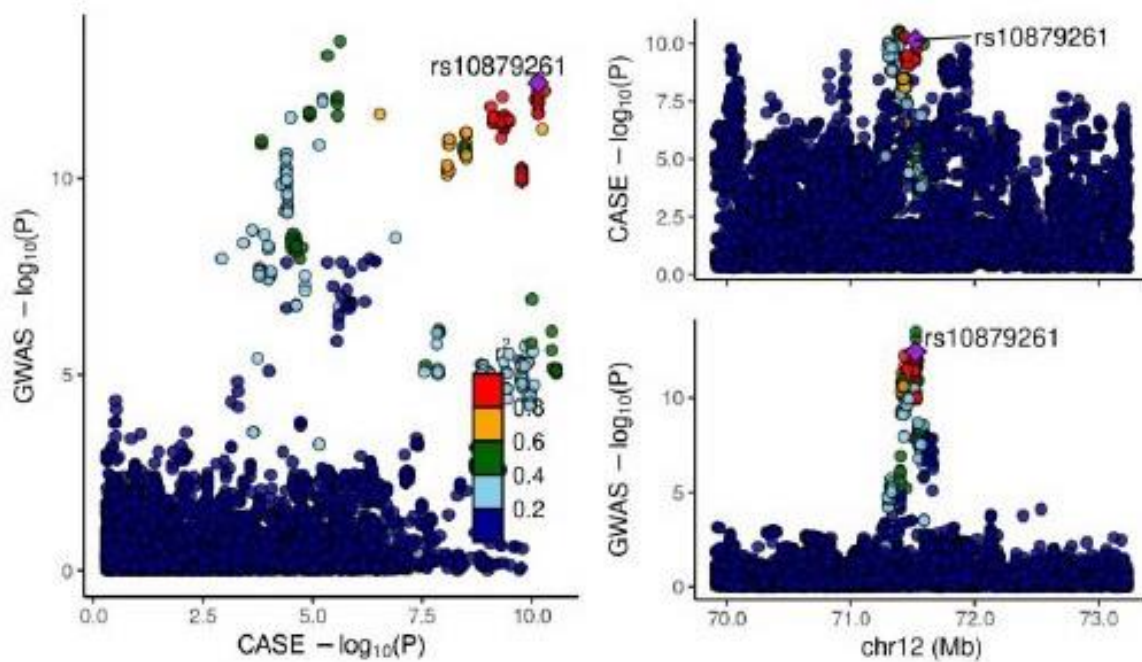
LocusCompare plots depicting all significant colocalizations between eQTL and T2D GWAS analyses. The lead variant is represented by a purple diamond. The linkage disequilibrium between the lead variant and the other variants is given as the square of the correlation coefficient r^2 and is indicated in a color scale. The $-\log_{10}(p\text{-values})$ for each variant — which are located in a region of one mega-base pair up- and downstream from the gene transcription start site — are depicted in three panels: (left) p -values of eQTL as x-axis and GWAS as y-axis, (bottom right) p -values of GWAS in the gene region and (top right) p -values of eQTL in the gene region. The title shows the gene name; MAF: minor allele frequency; PP.H4.abf: Posterior probability of colocalization; SNP.PP.H4: posterior probability of lead variant being the associated causal variant.

Data S2



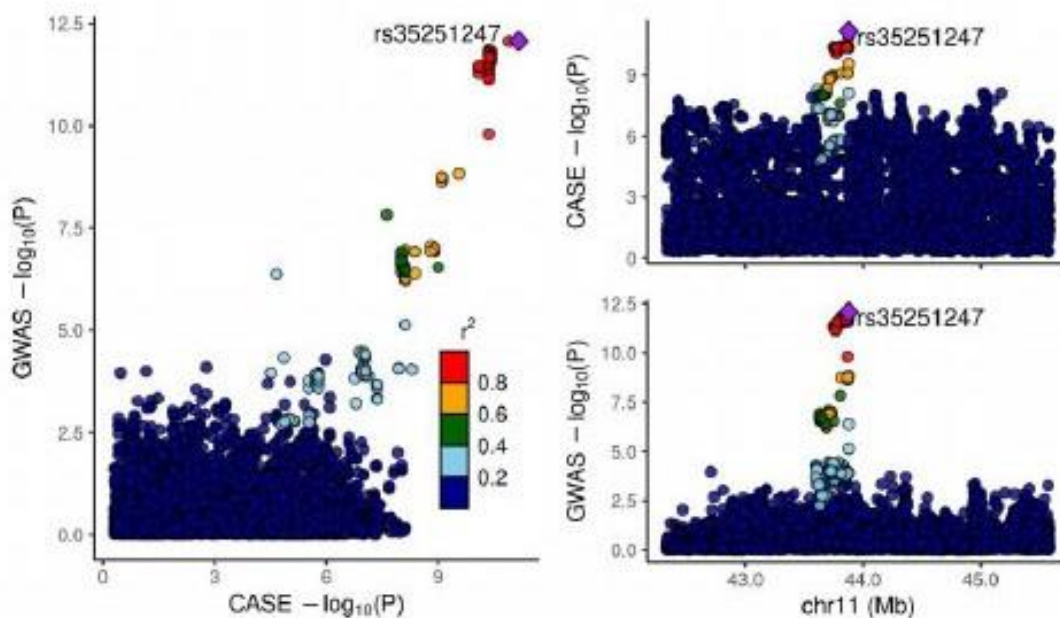
TSPAN8

MAF: 0.410 PP.H4.abf: 0.850 SNP.PP.H4: 0.081



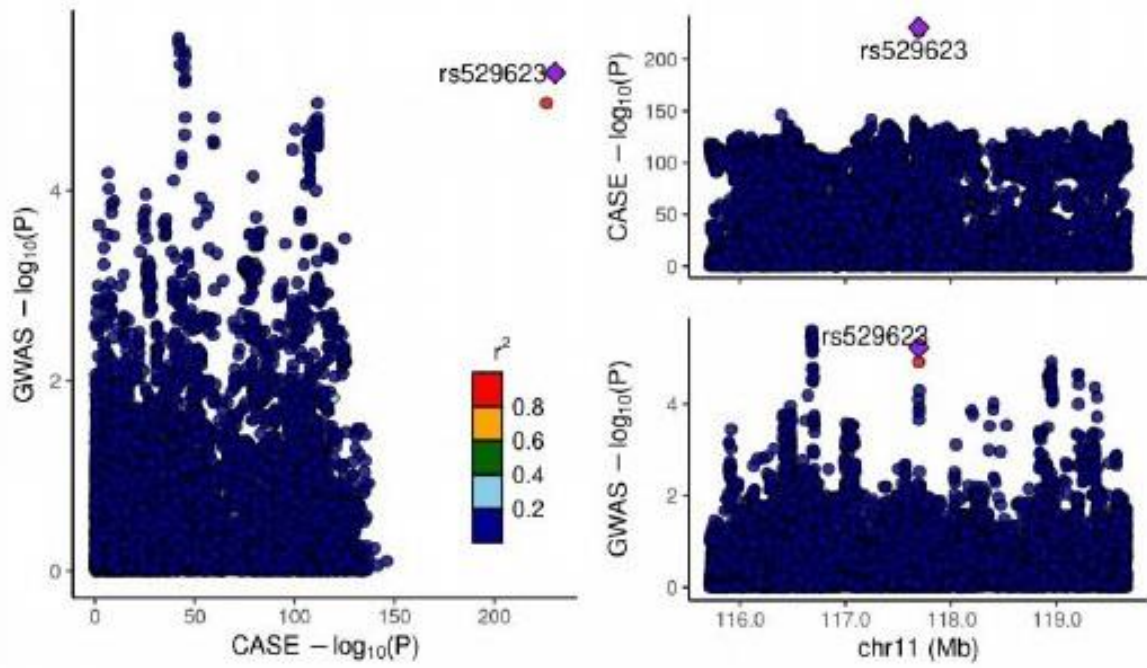
RP11-613D13.5

MAF: 0.290 PP.H4.abf: 0.930 SNP.PP.H4: 0.074



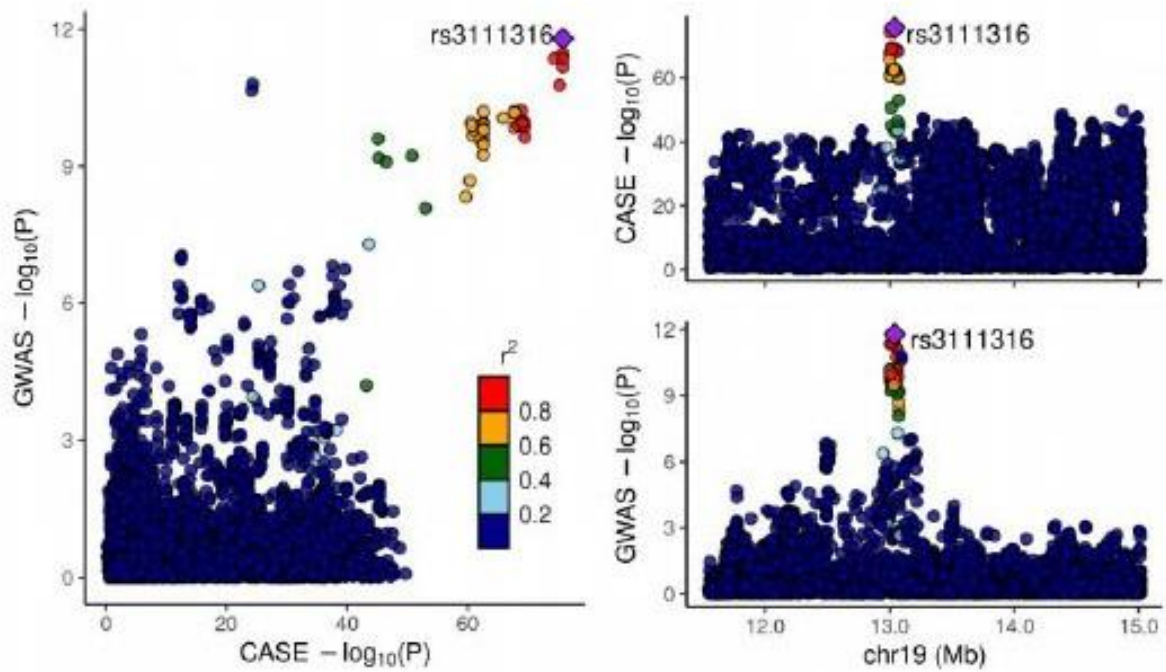
FXVD2

MAF: 0.480 PP.H4.abf: 0.945 SNP.PP.H4: 1.000



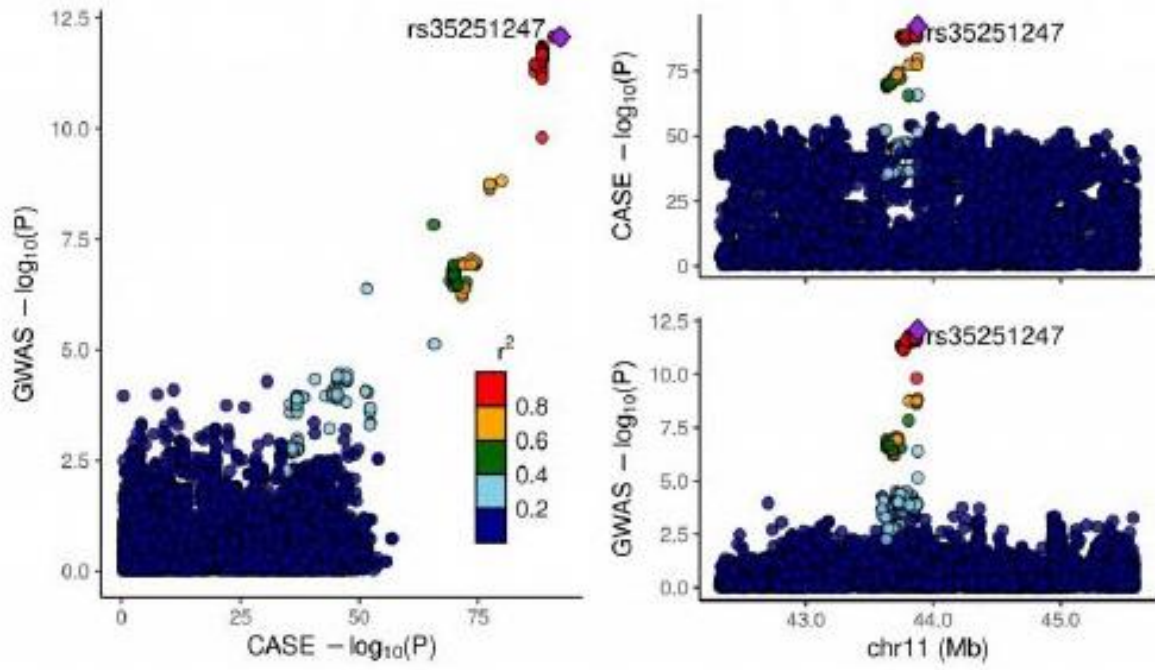
CALR

MAF: 0.410 PP.H4.abf: 0.994 SNP.PP.H4: 0.469



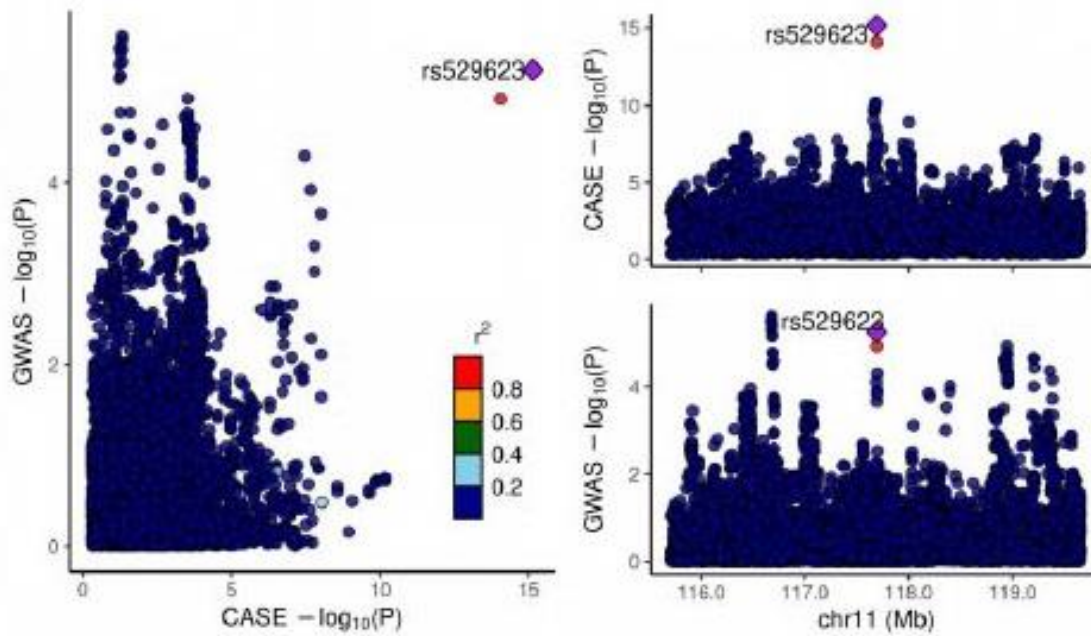
HSD17B12

MAF: 0.290 PP.H4.abf: 0.952 SNP.PP.H4: 0.211



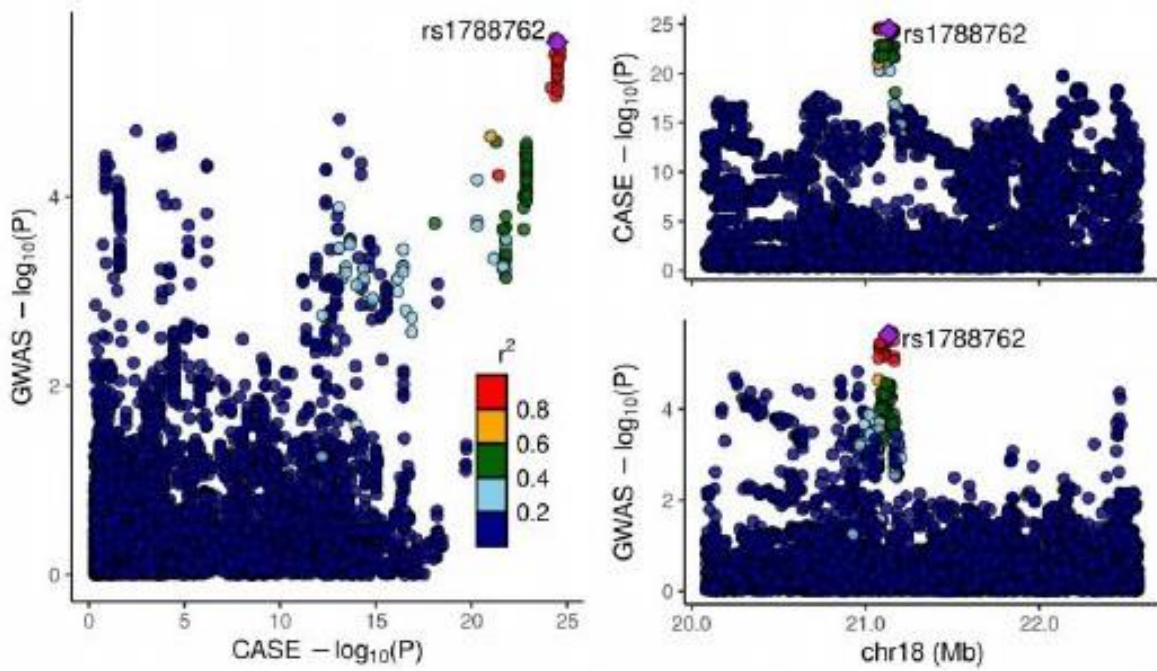
RP11-728F11.3

MAF: 0.480 PP.H4.abf: 0.910 SNP.PP.H4: 0.808



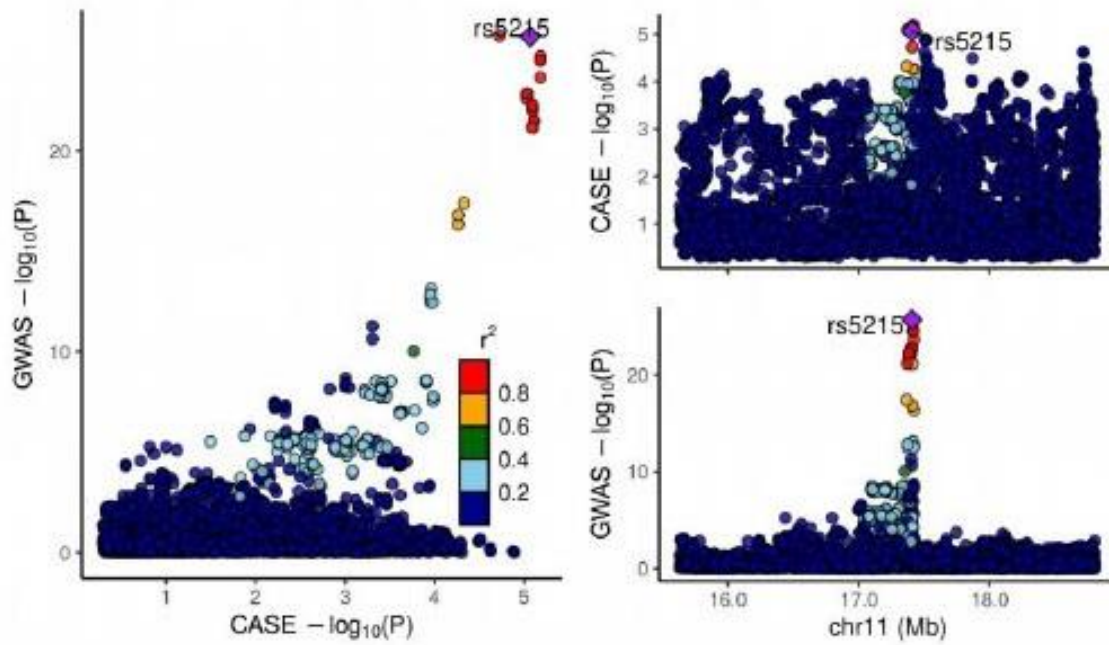
NPC1

MAF: 0.360 PP.H4.abf: 0.956 SNP.PP.H4: 0.064



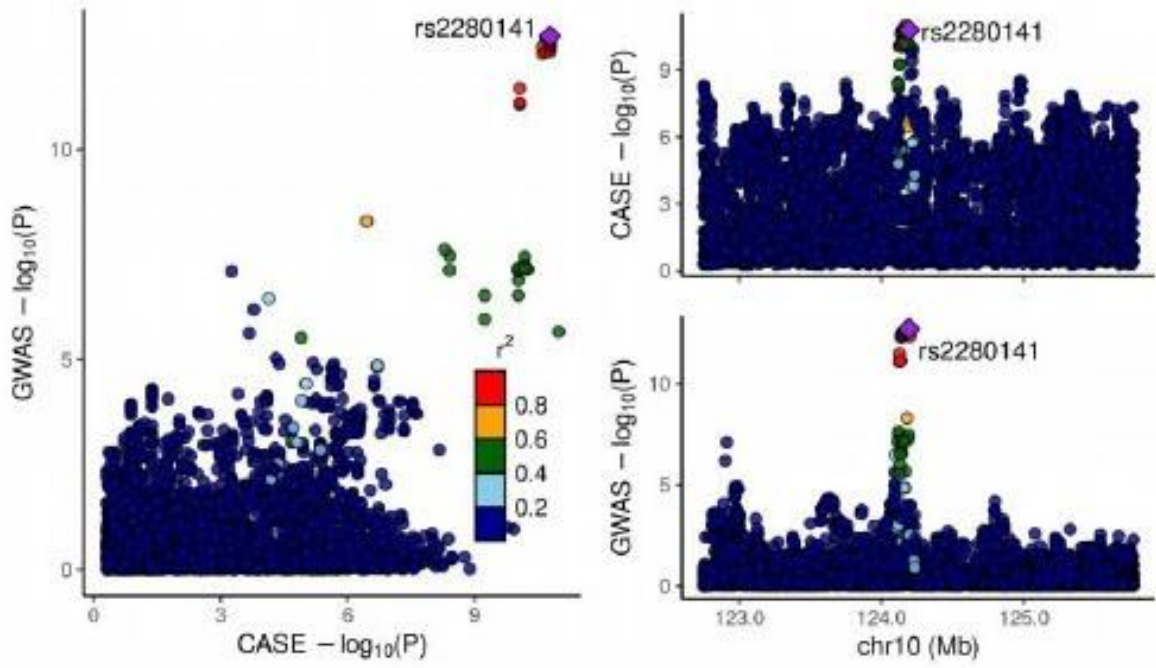
KCNJ11

MAF: 0.370 PP.H4.abf: 0.832 SNP.PP.H4: 0.361



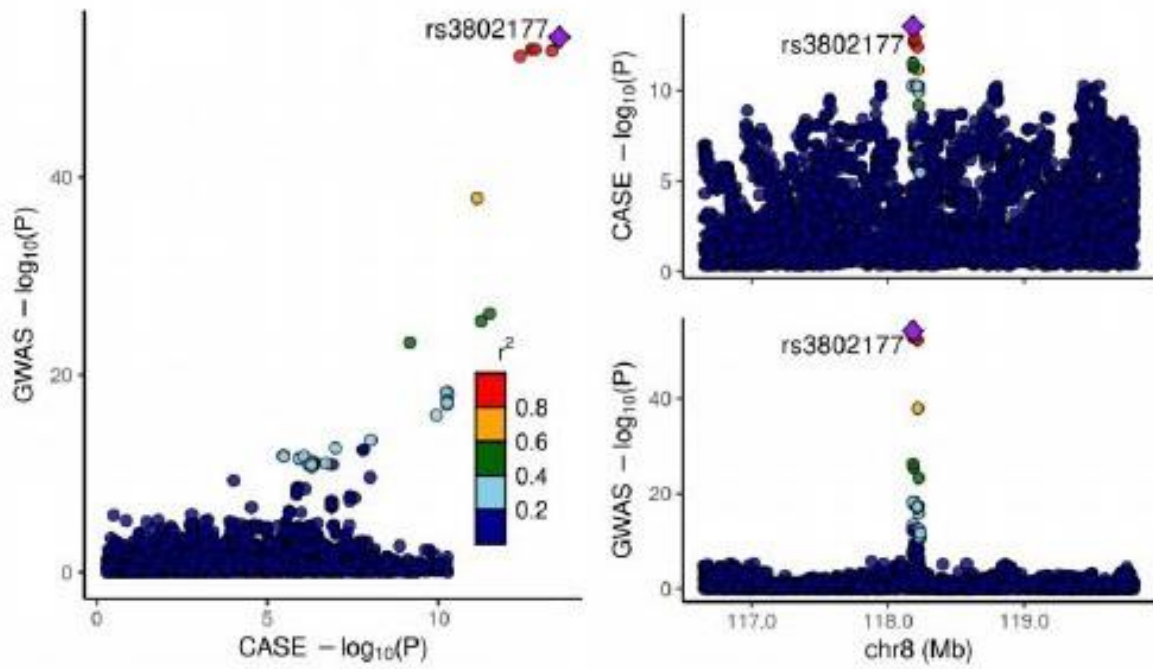
PLEKHA1

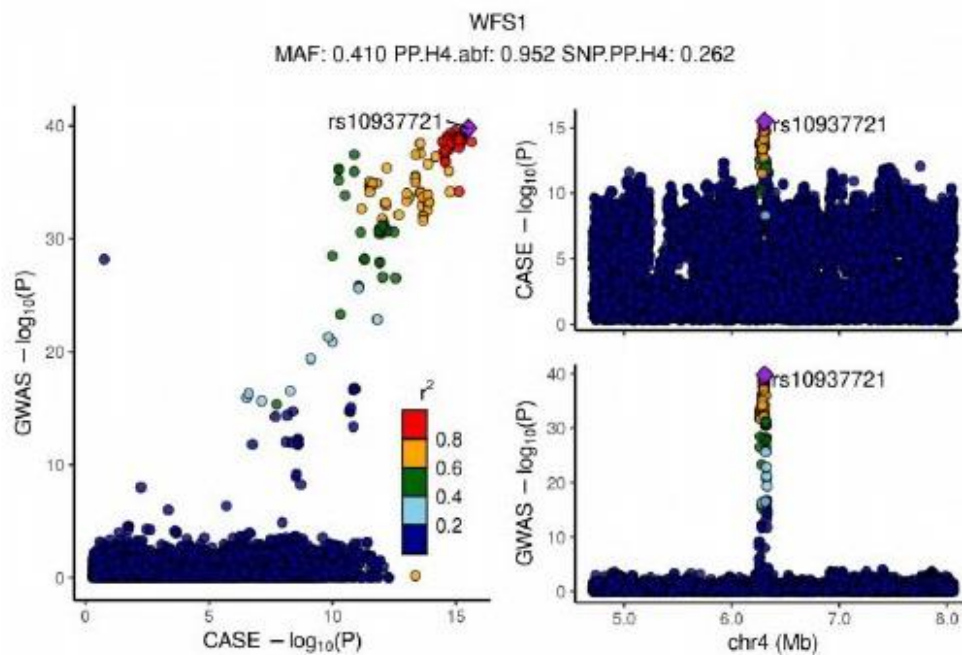
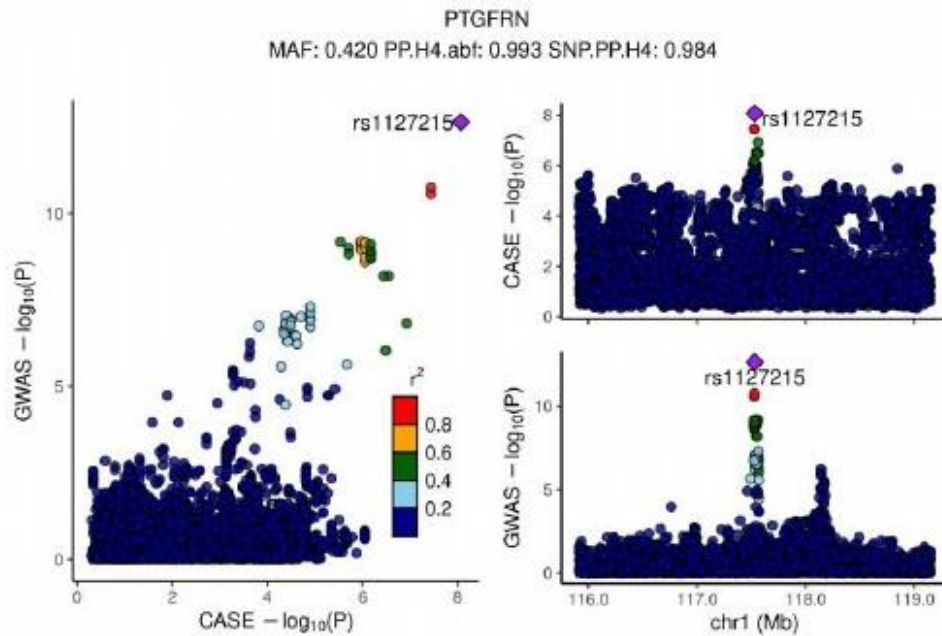
MAF: 0.480 PP.H4.abf: 0.955 SNP.PP.H4: 0.060



SLC30A8

MAF: 0.310 PP.H4.abf: 0.996 SNP.PP.H4: 0.607





Colocalization plots of cASE signals, related to Table 2 and STAR Methods.

LocusCompare plots depicting all significant colocalizations between cASE and T2D GWAS analyses. The lead variant is represented by a purple diamond. The linkage disequilibrium between the lead variant and the other variants is given as the square of the correlation coefficient r^2 and is indicated in a color scale. The $-\log_{10}(p\text{-values})$ for each variant — which are located in a region of one mega-base pair up- and downstream from the gene transcription start site — are depicted in three panels: (left) p -values of cASE as x-axis and GWAS as y-axis,

(bottom right) p -values of GWAS in the gene region and (top right) p -values of cASE in the gene region. The title shows the gene name; MAF: the minor allele frequency; PP.H4.abf: Posterior probability of colocalization; SNP.PP.H4: posterior probability of lead variant being the associated causal variant.

GLOBAL RESULTS AND DISCUSSION

7. Global Results and discussion

Despite the large catalog of variants which have been found associated with complex diseases, such as diabetes, asthma, or Alzheimer's disease, only a small fraction of the heritability has been explained, thus affecting current predictive models and its application to the clinics (Kullo et al., 2022; Kumuthini et al., 2022; Lambert et al., 2019). This is in part derived from Genome-Wide Association Studies (GWAS) limitations (Génin, 2020; Tam et al., 2019). Particularly, the evaluation of single independent variants in a background of complex diseases, where the simultaneous combination of multiple genetic and environmental factors are required to develop the disease, represents an obstacle for the discovery of variant synergies. Additionally, the outcomes from GWAS are limited to the summary statistics, which despite its relevance for understanding which are the regions involved in disease predisposition, and their effect, only can be used in predictors, thus disregarding the comprehension of the molecular mechanisms underlying variation and its association with diseases, and restricting the advance towards the discovery of new drugs and treatments.

Overall this thesis contributes to the better understanding of the genomic basis of complex diseases focusing on these two limitations as a departure point. On one hand, the analysis of epistasis constitutes a novel approach to overcome the lack of knowledge about the existence of variant-variant interactions associated with diseases, and their effect (Génin, 2020; Tam et al., 2019; Visscher et al., 2017; Wray et al., 2013). Particularly, we develop and use machine learning models to find groups of epistatic variants associated with Type 2 Diabetes (T2D). Moreover, progressive pancreatic islet dysfunction has been described to play an important role in the explanation of T2D pathophysiology and other related traits (Bartolomé, 2022; Del Guerra et al., 2005; Eizirik et al., 2020; Gloyn et al., 2022). Therefore, we analyse the *cis*-regulatory effects of variation in pancreatic islets gene expression. Additionally, we create a publicly available platform integrating the results obtained from these analyses with other functional information to facilitate the interpretation of disease susceptibility loci. In the next pages the results obtained from this thesis are discussed.

7.1. Epistasis

The multiple advances done in the genomic study of T2D have led to the discovery of more than 700 GWAS variants significantly associated with this disorder (Bonàs-Guarch et al., 2018; J. Chen et al., 2021; Mahajan, Taliun, et al., 2018; Scott et al., 2017; The DIAGRAM Consortium et al., 2014; Vujkovic et al., 2020). However, despite these efforts, the contribution of these variants to the development of the disease is evaluated in a single independent manner, therefore resulting in a poor understanding of the disease, with only a small fraction of its heritability explained (~20%) (DeForest & Majithia, 2022). Epistasis or variant-variant interactions has been suggested as one of the factors that can contribute to a better genomic explanation of complex diseases, particularly, to T2D (Génin, 2020; Tam et al., 2019; Visscher et al., 2017; Wray et al., 2013).

Capitalising the fact that each GWAS variant contributes with a small effect to disease development (McCarthy et al., 2008), polygenic risk scores (PRS) are currently broadly applied to empower GWAS based on the existence of variant synergies. Particularly, to evaluate the predisposition to disease, PRS additively combines the effects of GWAS variants. Compared with more complex approaches, such as machine learning methods, PRS represents a benefit being a cost-effective technique, which only requires the use of GWAS summary statistics to calculate the risk score of each individual genotype in a particular cohort of study. However, the use of these scores have not resulted in a great improvement compared to the predictions based on clinical measures (Padilla-Martínez et al., 2020). Moreover, although these approaches consider variant synergies, PRS ignores the possible functional interconnections between variants and its association with disease phenotypes (Mackay, 2014).

To approach the study of epistasis, and therefore, to discover variants which contribute to the development of the disease synergically, but in a dependent manner, different statistical methods and computational approaches have been applied. Nonetheless, the many computational and methodological difficulties surrounding epistatic studies have limited its progress. For example, the analysis of the complete set of pairwise interactions between only 500,000 SNPs using classical statistical methods, involves the execution of billions of tests (Marchini et al., 2005). Thus, converting the study of epistasis into something still unaffordable at a genome-wide level or, when discretely approached, reporting no evidence of epistasis as a result of the highly restrictive multiple testing thresholds that are needed to ensure the accuracy of the outcomes (Nag et al., 2020).

As a consequence of the complexity behind genome-wide epistasis analysis, diverse techniques such as multidimensionality reduction analysis, or variant prioritisation based on previous biological knowledge have been applied (Manduchi et al., 2018; Josep Maria Mercader et al., 2008; Nag et al., 2020). Remarkably, some of the studies based on the use of these techniques have reported variants which present a modest effect on the disease when evaluated independently, but an increased effect when considered jointly (Cordell, 2009; Kirino et al., 2013; Monir & Zhu, 2017). However, the small number of variants included in these studies, have reduced the discovery to a few genetic loci.

In contrast with classical methods, which are able to approach the epistatic problem at a genome-wide level (Nag et al., 2020), in this thesis, the use of a Machine Learning (ML) approach has limited the extension of the method to the analysis of a small group of variants. This problem is a consequence from the current statistical and computational limitations derived from the use of ML methods. More specifically, to ensure a good performance of the statistical models underlying ML approaches, and prevent overfitting, therefore allowing the replication in a completely independent dataset, the number of variants that can be included in the input dataset is recommended to be less than 10% of the observations (Chicco, 2017; Dey, 2016; Greener et al., 2021; Sarker, 2021). Therefore, many side analyses have been required to ensure the possibility of replication when trespassing this threshold. Moreover, the computational memory load that represents the evaluation of the multiple combinations that can be simultaneously tested for its association with the disease, has also represented a computational burden for the use of our ML approach in a genome-wide manner. Therefore, although the imputed genotype for more than 15 million variants was available for the analysis, we have largely reduced the number of input variants to only 105,896.

The reduction in the number of input variants applied in the analysis conducted in this thesis can be contrasted with other studies which rely on multi-dimensionality reduction techniques or that reduce the number of initial variants by applying filters based on previous functional knowledge (Manduchi et al., 2018; Josep Maria Mercader et al., 2008; Nag et al., 2020). Particularly, the filter based on only keeping the variants with a certain association with the disease, despite resulting in a large reduction of the number of variants included, is less restrictive than other functional filters. Therefore, resulting in a lower dimensionality reduction, and therefore facilitating a broad inspection of the synergies between a larger group of variants.

Despite applying a filter based on the marginal degree of association with the disease we expected our interacting groups to be enriched in functionally relevant variants, it is known that the vast majority of disease-associated variants lie on non-coding regions, thus difficulting the functional interpretation of the results obtained from our analyses. Particularly, to confirm that the interaction pathway between variants can be mediated by the affected genes, and to suggest islet regulatory regions and islet expression regulatory variation as some of the underlying mechanisms mediating the effect of genomic variants interaction, many side analyses integrating and evaluating our outcomes with functional annotations including T2D and related traits GWAS meta-analyses, *cis*-regulatory expression, gene functional impact, and epigenetic marks, had been required. Importantly, although

from these analyses we have found that some of the single independent variants, which were previously known to be significantly associated with T2D, can be thought as driving the effect of the interaction (Hemani et al., 2021), 75% of the epistatic groups do not include any variant previously associated with T2D, glycemic traits, or an already known susceptibility loci for islet expression, thus evidencing the relevance of interactions between different genomic regions to improve the understanding of the disease.

Remarkably, the use of ML models, in this thesis, compared with more classical methods (Cordell, 2009; Kirino et al., 2013; Monir & Zhu, 2017; Nag et al., 2020), has facilitated the suppression of the very restrictive multiple testing significance threshold (Marchini et al., 2005). The avoidance of this restriction, which usually results in poor detection power and limits the discovery to a few significant loci, has allowed the detection of 367 single variants, 980 pairs, 1,952 triplets, and 3,607 quadruplets which contain variants synergically associated with the predisposition to T2D. However, the combination of this ML methodology with classical logistic regression to explore the existence of epistatic variants inside these groups, and to measure their effect, in the same line as the cited studies, although successful, has reduced the detection power of our analysis, mainly because of the need of applying multiple testing corrections to ensure the significance of the tests performed. Fortunately, the reduced number of final tests has resulted in a less restrictive threshold to ensure the significance, and despite this statistical burden has reduced the discovery to a few loci, we have been able to find 10 pairs, 1 triplet, and 1 quadruplet of epistatic variants associated with T2D, which would have been impossible to find by applying current methods.

Finally, in this thesis we have explored the simultaneous effect of multiple variants and its association with T2D. Particularly, we have taken advantage of the use of Machine Learning (ML) approaches which, in contrast with other methods, such as PRS, have facilitated not only the prediction of disease predisposition based on the combination of the effects of multiple genomic variants, but also the discovery of variant synergies. Remarkably, these synergies include both the additive and epistatic ways of variant interactions. Notably, while measuring the effect of variant synergies, we observed that there were significant differences between the marginal effects of the variants under the logistic regression additive model and the full model including interactions. More specifically, we found some variants from which effect not only varied in module but also in the sign, thus changing for example from being protective to represent a risk for the development of the disease. This finding represents a new challenge for current PRS which sum the marginal effect of variants without accounting for the possible changes in their effects derived from their synergies. Additionally, the creation of an input dataset with paired metadata case-control individuals, which although it can be argued that can result in a loose of detection power due to the reduction of individuals, has enhanced the discovery of genomic loci that, apart of synergically contributing to the development of the disease, are less representative of clinical disease-related measures. Thus, overcoming the limitations of the prediction of PRS, which are still far from improving the predictions based on clinical measures (Padilla-Martínez et al., 2020). Particularly, the use of PRS in our prediction dataset, which includes 2,280 cases and 2,280 controls with paired metadata, therefore, individuals from which less variance explanation is expected from the clinical measures, results in a ~50% of precision, which is far from the ~60% of precision obtained from the use of our ML method.

However, despite the potential of the results obtained in our study, there are some limitations that can be improved in future epistatic analysis. First, the restriction of the analysis to European ancestry individuals affects the extension of the results to non-European populations, limiting its explanation to common shared ancestry loci (Josep M. Mercader et al., 2017; Spracklen et al., 2020; Vujkovic et al., 2020). Second, the still computational challenge of analysing millions of variants simultaneously has limited our study to variants with a higher probability to be associated with T2D. The increase of computational power, or the use of other approaches, can facilitate the discovery in future epistatic studies. Third, the number of individuals included in the study has represented an

additional layer of complexity related to the methodology applied in our study. Nevertheless, as the number of individuals is increasing in current studies, in the future, better results can be obtained using the same approaches. Fourth, all the analyses performed were under the additive inheritance model, thus limiting the discovery to variants falling under this model or non-additive models with a higher effect (Guindo-Martínez et al., 2021). Fifth, although chromosome X has been included in this study, there are many details that need to be considered for its appropriate analysis (Bonàs-Guarch et al., 2018). For this reason, future epistatic studies relying on the same methodology applied in this project will need to improve the approaches presented to enhance the discovery power in this chromosome. Finally, the work presented in this thesis shows just the first preliminary results of the study. Therefore, there are some plans to improve the analyses performed previous to its publication, which include the replication of our results in a completely independent dataset. In addition to this, in a background of personalised medicine, this study can be observed as a first step to understand the effects of epistasis in T2D. Thus, opening a new avenue for the analysis of epistasis in other complex diseases, and to reveal the epistatic differences between subgroups of patients (Ahlqvist et al., 2018, 2020; Dimas et al., 2014; H. Kim et al., 2022; Mahajan, Wessel, et al., 2018; Mansour Aly et al., 2021; McCarthy, 2017; Scott et al., 2017; Udler et al., 2018).

7.2. TIGER

The great advances produced by the use of GWAS for the genomic study of complex traits and diseases have led to the discovery of a large number of genetic variants statistically associated with the disorder (Beck et al., 2014; Buniello et al., 2019; K. Watanabe et al., 2019). Particularly, for the case of T2D, more than 700 loci have been found significantly associated with the disease (Bonàs-Guarch et al., 2018; J. Chen et al., 2021; Mahajan, Taliun, et al., 2018; Scott et al., 2017; The DIAGRAM Consortium et al., 2014; Vujkovic et al., 2020). However, the lack of functional interpretation of these signals has complicated the understanding of their underlying molecular mechanisms and its relation with disease. The use of genomic, transcriptomic, and epigenetic information to evaluate the overlap between disease associated loci and function has been suggested as one of the ways to improve disease knowledge (Cano-Gamez & Trynka, 2020; Lichou & Trynka, 2020; Manolio, 2013).

Remarkably, although gene expression can be ubiquitous or cell-type specific, some of the regulatory elements such as gene expression signatures, enhancers, and promoters are cell-type specific (Long et al., 2016; Nica & Dermitzakis, 2013; Pope & Medzhitov, 2018). Thus, suggesting the relevance of the study of disease related cell-type or tissue-specific regulatory elements to improve the understanding of the mechanisms mediating disease. Particularly, progressive pancreatic islet dysfunction has been described to play an important role in the explanation of T2D pathophysiology (Bartolomé, 2022; Del Guerra et al., 2005; Eizirik et al., 2020; Gloyn et al., 2022). More specifically, pancreatic beta-cells deterioration or death can lead to insulin secretory dysfunctions, usually resulting in hyperglycemia. Thus, converting pancreatic islets in a very relevant tissue for the study of T2D and other related traits. However, there are many restrictions which limit the access to human pancreatic islets and also convert their analysis into a challenge (Gloyn et al., 2022).

In addition to this challenge, the fact that the vast majority of variants significantly associated with a complex disease lie in non-coding regions and that the relationship between variation and transcription factors cannot always be inferred from the proximity with a gene binding site (Deplancke et al., 2016), adds a layer of complexity to the functional interpretation of genomic variation. Thus, converting the *cis* inspection of the transcriptome of genomic variation in a powerful tool. As a result, some transcriptomic techniques, which are broadly used to understand the effect of genetic variation on gene expression, such as expression quantitative trait loci (eQTL) or allele-specific expression (ASE), have become crucial for the functional understanding of genomic variation (Albert & Kruglyak, 2015b; Cleary & Seoighe, 2021; Nica & Dermitzakis, 2013). Nevertheless, despite large databases

have been generated containing the outcomes from the study of the effects of variation in the transcriptome of different tissues, such as the GTEx initiative (The GTEx Consortium, 2020), which can be complemented by many expression studies in pancreatic islets (Fadista et al., 2014; Solimena et al., 2018; van de Bunt et al., 2015; Viñuela et al., 2020), these studies only recapitulate the effects of the groups of variants that have been analysed in their studies, which although representing a large amount of variation, are still incomplete.

Complementarily to the genomic and transcriptomic analysis of the effects of variation, epigenetic assays such as chromatin immunoprecipitation followed by sequencing (ChIP-seq) or assays for transposase-accessible chromatin sequencing (ATAC-seq) have been suggested to play a key role for the identification of transcription factor binding sites, and the identification of enhancers, and therefore for the *cis*-regulatory interpretation of GWAS outcomes (Buccitelli & Selbach, 2020; T. K. Kim & Shiekhattar, 2015; Lambert et al., 2018; Smith et al., 2012). As a result, in a same manner than expression, *cis*-regulatory maps have been broadly studied in different cell types (The ENCODE Project Consortium, 2012), including pancreatic islets (Hall et al., 2014; Miguel-Escalada et al., 2019; Pasquali et al., 2014; Thurner et al., 2018). Overall, many efforts have been devoted to generate large islets transcriptomic and epigenetic databases, which are the promise to promote the genetic understanding of T2D and other islet related diseases. However, although many efforts have been devoted to the generation of genomic browsers and other public platforms which facilitate the access to this valuable information (Beck et al., 2014; Buniello et al., 2019; Haeussler et al., 2019; The GTEx Consortium, 2020; K. Watanabe et al., 2019), only a few of these resources are specific for T2D, such as the T2D Knowledge Portal (Flannick & Florez, 2016), or for pancreatic islets (Mularoni, Ramos-Rodríguez, & Pasquali, 2017). Remarkably, despite the vast majority of genomic studies highlighting the relevance of the integration of different omic layers to improve the understanding of disease development, none of them has yet analysed and integrated diverse pancreatic islets omics in a unique publicly accessible database.

In contrast with previous pancreatic islets studies, which were boosted from one independent research centre, in this thesis we have benefited from the collaboration of a large consortia, the T2DSysTems, which involved, among many other participants, five research centres with wide expertise in the analysis of human pancreatic islets. As a result from this huge collaboration, we have been procured access to the largest pancreatic human islet cohort, which included the RNA-seq, the genotype and the metadata of 514 pancreatic islets samples, from which 307 samples were novel. This collaboration reduced some of the problems that can be derived from the access to this valuable resource of data (Gloyn et al., 2022), and facilitated the collection, harmonisation and quality control of the data. As a result of this process, although some of the samples being discarded, 404 islet samples were kept, thus representing a large increase in the sample size compared with previous islets expression studies (Fadista et al., 2014; van de Bunt et al., 2015), and therefore, a potential increase in the association detection power derived not only from the study of cell-type specific expression but also from the increment of samples (Long et al., 2016; Nica & Dermitzakis, 2013). However, in parallel to these efforts and during the development of this thesis, another large cohort of pancreatic islets was created accounting with 420 samples and with an overlap of 206 samples with our cohort (Viñuela et al., 2020).

Remarkably, current islets studies including the above mentioned recently published study from Viñuela (Fadista et al., 2014; Miguel-Escalada et al., 2019; Pasquali et al., 2014; Solimena et al., 2018; Thurner et al., 2018; van de Bunt et al., 2015; Viñuela et al., 2020), only focus in one type of analysis. More specifically, capitalising on the benefits of the study of this particular cell-type and its relevance to improve the functional explanation of islet-related diseases, most of these previous projects targeted the transcriptomic analysis of gene expression or the study of epigenetic marks. In contrast, in this thesis, we aimed to generate the Translational Human Pancreatic Islets Genotype-Tissue Expression Resource (TIGER), a unique platform which integrates the outcomes obtained

from homogeneous islets gene expression, one of the biggest, if not the biggest, islet eQTL meta-analysis, and a new trustworthy method to measure allele specific expression, combined with already published epigenetic marks, and T2D GWAS meta-analysis summary statistics, in a publicly available database, thus constituting a unique and formidable resource for the functional interpretation of pancreatic islets and related diseases.

In this thesis, we have taken advantage of this large resource of pancreatic islets to calculate and homogenise islets gene expression, and to include this information in the public platform in a visual way so that it facilitates the comparison between the expression in islets of a given gene with the rest of the genes in the genome. Although this information is also available in other platforms, such as the GTEx (The GTEx Consortium, 2020), the GTEx platform does not allow the comparison with other genes and, most importantly, do not include pancreatic islets expression. Additionally, as it can be argued that the GTEx project includes the gene expression counts for a wide diversity of tissues while we are only recapitulating this information for a specific tissue, we have scaled islets expression to be compared with other reference tissues. As a result, the TIGER platform not only shows if a gene is expressed in pancreatic islets but also allows the comparison of expression across all the GTEx tissues. Remarkably, despite there is a high order of eQTL similarity between different tissues (The GTEx Consortium, 2020), the study of cell dysfunction based on eQTL tissue-specificity can lead to a better disease interpretation. Therefore, the integration of islets with other reference tissues in TIGER facilitates the comparison between the different T2D-related tissues (pancreas, brain, intestine, adipose tissue, muscle, kidney, liver and pancreatic islets) (Cnop et al., 2005; Cornell, 2015; DeFronzo, 2009; Del Guerra et al., 2005; Eizirik et al., 2020; Galicia-Garcia et al., 2020; Gilon, 2020; Rhodes, 2005) and, therefore, promotes the detection of the best tissue to functional interpret disease susceptibility loci.

In comparison with previous and the most recent islet eQTL studies (Fadista et al., 2014; van de Bunt et al., 2015; Viñuela et al., 2020), in this thesis we benefited from an improved imputation using GUIDANCE (Guindo-Martínez et al., 2021). Particularly, these studies imputed the genotype using 1000 Genomes reference panel (The 1000 Genomes Project Consortium, 2015), while we used multiple reference panels including 1000 Genomes, UK10K, GoNI and HRC (Boomsma et al., 2014; The 1000 Genomes Project Consortium, 2015; The Haplotype Reference Consortium, 2016; The UK10K Consortium, 2015). As after the imputation we merged the results to recover each variant from the panel reporting the best imputation quality (INFO>0.7), this allowed us to include a higher number of good quality genetic markers, compared with the previous published studies. More specifically, while previous published studies included between 5.8 million and 8 million variants, we imputed over 22 million unique genetic variants with high-quality across all of the samples, of which approximately 10% are Indels and small SVs, more than 1.05 million variants in chromosome X, above 4 million low-frequency variants, and over 10 million rare variants. Notably, only in the last study (Viñuela et al., 2020) and this thesis rare variants were included, while in the rest of previous studies those were disregarded (Fadista et al., 2014; van de Bunt et al., 2015), despite their interest given their expected higher effect on the risk of developing the disease (McCarthy et al., 2008). This maximisation of genetic variants improved the detection power of the expression analyses resulting in over 1 million eQTLs and 256,981 ASE associated variants.

Notably, current variation expression analyses use a wide variety of tools to colocalise their outcomes with GWAS summary statistics to find possible connections with disease or to check the overlap with regulatory elements. However, this type of analyses are computationally expensive and even, in some cases, it is complex to get granted access to the data. As a result, for example, it is common that colocalisation analyses only use the summary statistics from the latest published study, thus disregarding the signals that have been only captured in other cohorts. In this thesis, we have facilitated the colocalisation analysis by aggregating the results from the largest T2D GWAS meta-analyses from European ancestry (Bonàs-Guarch et al., 2018; Mahajan, Taliun, et al., 2018; Scott et

al., 2017; The DIAGRAM Consortium et al., 2014). Moreover, the integration in the platform of a genomic browser (Down, Piipari, & Hubbard, 2011), containing different islets epigenetic marks, and a wide diversity of elements from the human islet regulome (Hall et al., 2014; Miguel-Escalada et al., 2019; Pasquali et al., 2014; Thurner et al., 2018) not only promotes the easy and fast check for the overlap with islet regulatory annotations but also allows the comparison with unpublished tracks.

In summary, the large number of expression regulatory variation results obtained in human pancreatic islets in this project, as well as the database and the platform created during this thesis, represent a valuable resource for the study of diabetes, related traits, and other disorders where pancreatic islets have a central pathogenic role (**Figure 11.A**). Particularly, from the last 90 days report obtained from the website (8th July 2022), we know that 325 users from all over the world have been accessing the portal, with over 500 sessions during this period (**Figure 11.B-C**). Interestingly, most of these users seem to be familiarised with the platform, as they have accessed it directly through the URL. However, we are still capturing new users through Google organic search and other referrals such as ncbi.nlm.nih.gov (**Figure 11.D**). Thus suggesting a real interest on the platform and all the results that it includes. More specifically, this portal has been proved successful to provide support to many recently published genetic studies (Bone et al., 2021; Dorsey-Trevino, Kaur, Mercader, Florez, & Leong, 2022; O'Connor et al., 2022; Sulaiman et al., 2022; Zheng et al., 2020).

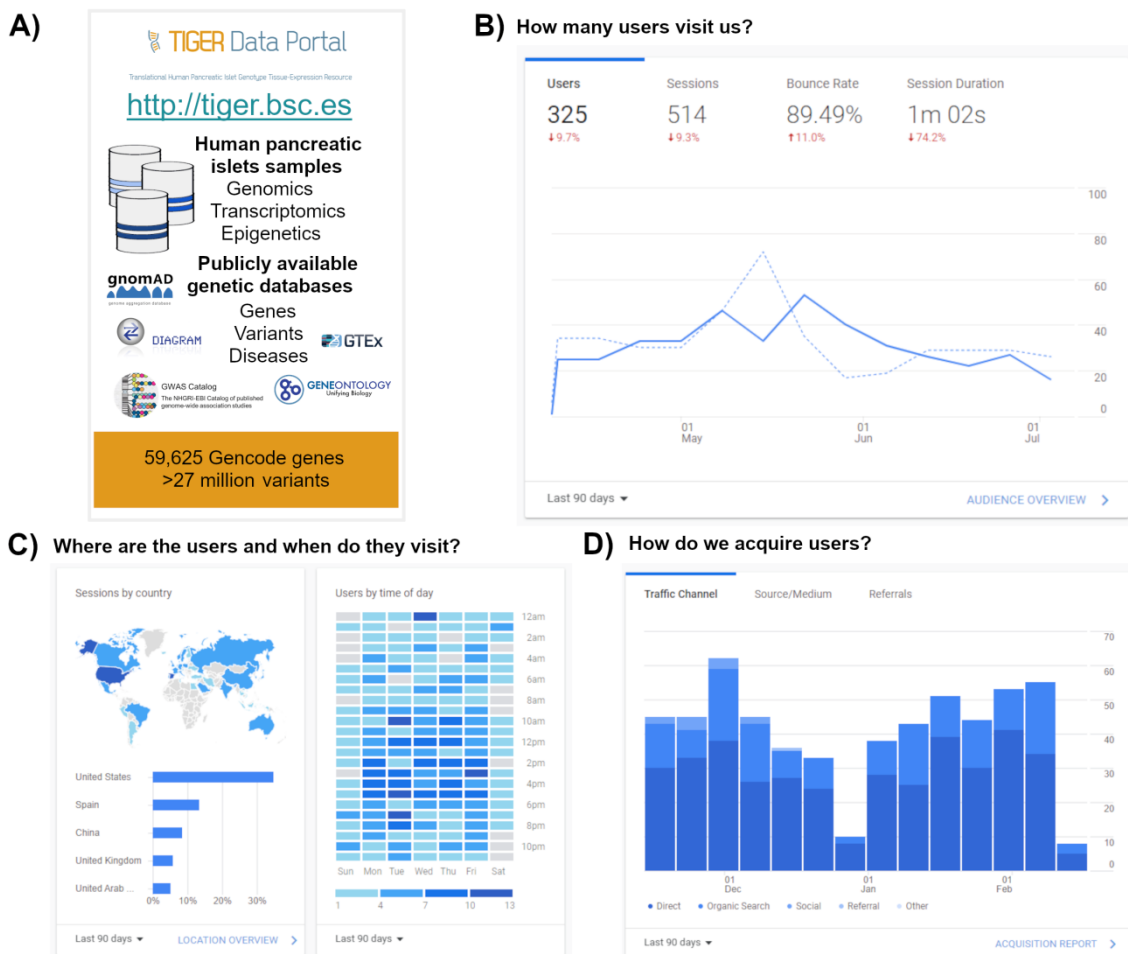


Figure 11. TIGER data portal. We integrated genomic, transcriptomic and epigenetic human islet results, with other publicly available disease, gene, and variant information in the publicly available TIGER Data Portal. The different graphs represent:

A) The general overview of the content inside the TIGER resource.

B) The number of users accessing the portal during the last 90 days (8th July 2022) and the average duration of the session. The straight blue line represents the number of users (y axis) accessing the platform by week (x

axis). The dashed blue line represents the number of users that accessed the platform in the same week 3 months ago.

C) Sessions by country (left) and by time of day (right). The map on the (top left) has coloured in blue the countries with users accessing the platform during the last 90 days (8th July 2022). The blue scale represents the countries with more (dark blue) or less (light blue) access to the portal. The bar plots (bottom left) show the distribution of sessions (x axis) on the top 5 of the countries (y axis). The heatmap (right) displays the accession times (y axis) by the day of the week (x axis) to the platform. The blue scale represents the range of time with more (dark blue) or less (light blue) sessions detected.

D) How users are acquired. The bar plots represent the number of users (y axis) accessing the platform by week (x axis) during the last 90 days (8th July 2022). The different blue colours represent the way of accessing the platform: direct access through the URL tiger.bsc.es (dark blue), Google organic search (medium blue), or other referrals (light blue).

However, despite the potential interest of the outcomes generated in this study, there are many limitations that should be focused in the future. First, although the overlap between the samples analysed in a previously published eQTL study and TIGER facilitates the confirmation of some results, it also complicates replication (Viñuela et al., 2020). For this reason, future studies should only focus on non-overlapping samples. Second, despite this being presumably one of the largest, if not the largest, pancreatic islets datasets analysed for the effects of expression, the integration of additional datasets in the study will increase the prediction power of the analysis. Third, the samples included in the study were only from European ancestry, thus complicating the extension of the results to non-European populations, and limiting it to the shared variants between populations (Josep Maria Mercader & Florez, 2017; Spracklen et al., 2020; Vujkovic et al., 2020). For this reason, future studies should collect data from different ancestries. Fourth, despite pancreatic islets being made by a heterogeneous group of cells, the use of bulk RNA-seq data in our study limits the discovery to only capture the effects of expression of the more representative cells or the average between the different groups of cells. Hence, the use of single-cell sequencing will enhance the expression study in each particular group of cells and allow its comparison (Kawasaki, 2004). Fifth, the fact that the largest T2D GWAS meta-analysis (Mahajan, Taliun, et al., 2018) doesn't include SVs or Indels limited our colocalization study. Therefore, the inclusion of Indels and SVs in future T2D GWAS will improve the understanding of T2D pathophysiology. Sixth, despite the expression analyses included the study of a large fraction of coding elements there are still some elements that are uncovered, such as microRNA. Therefore, the inclusion of these elements in future expression studies can be useful to gain insight of T2D pathophysiology (Taylor et al., 2022).

CONCLUSIONS

8. Conclusions

8.1. Epistasis

- 1) The analysis of epistasis, using machine learning approaches, revealed thousands of groups of variants which combined additively or in a synergic dependent manner have an effect on disease development.
- 2) The study of variants interaction was crucial to find 75% novel loci associated with complex diseases (20 out of 27), thus improving the genetic understanding of T2D.
- 3) By analysing the effect of epistasis under a full logistic regression model we found 30% of the variants inside the epistatic groups (8 out of 27) changing the sign of its individual effect, therefore, affecting current detection and prevention protocols.
- 4) The regulation of gene expression of disease-associated genes is suggested as one the putative underlying mechanisms of epistasis and its association with complex diseases.

8.2. TIGER

- 5) The study of pancreatic islets promotes the translation of genomic variation in gene function and, therefore, the better understanding of T2D and other islets related disorders pathophysiology.
- 6) The use of integrative approaches in expression analyses has been crucial to improve the identification of additional genetic markers and to discovery over 1.05 million eQTLs and 256,981 cASE variants.
- 7) The combination of T2D GWAS results with eQTL and cASE is necessary to support the expression findings, and to facilitate the functional interpretation of GWAS.
- 8) The creation of a publicly available database that integrates different omic layers of information is essential to ensure the shareability of the results, and to provide the research community with powerful and useful tools to complement and support their studies.

REFERENCES

10. References

- Abbott, A. (2016). Scientists bust myth that our bodies have more bacteria than human cells. *Nature*. Retrieved from <https://doi.org/10.1038/NATURE.2016.19136>
- Abbott, S., & Fairbanks, D. J. (2016). Experiments on Plant Hybrids by Gregor Mendel. *Genetics*, 204(2), 407–422. Retrieved 20 September 2021 from <https://doi.org/10.1534/GENETICS.116.195198>
- Ahlqvist, E., Prasad, R. B., & Groop, L. (2020). Subtypes of Type 2 Diabetes Determined From Clinical Parameters. *Diabetes*, 69(10), 2086–2093. Retrieved 21 July 2021 from <https://doi.org/10.2337/DBI20-0001>
- Ahlqvist, E., Storm, P., Käräjämäki, A., Martinell, M., Dorkhan, M., Carlsson, A., ... Groop, L. (2018). Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. *The Lancet. Diabetes & Endocrinology*, 6(5), 361–369. Retrieved 27 February 2019 from [https://doi.org/10.1016/S2213-8587\(18\)30051-2](https://doi.org/10.1016/S2213-8587(18)30051-2)
- Akerman, I., Tu, Z., Beucher, A., Rolando, D. M. Y., Sauty-Colace, C., Benazra, M., ... Ferrer, J. (2017). Human Pancreatic β Cell lncRNAs Control Cell-Specific Regulatory Networks. *Cell Metabolism*, 25(2), 400–411. Retrieved 21 January 2021 from <https://doi.org/10.1016/j.cmet.2016.11.016>
- Albert, F. W., & Kruglyak, L. (2015a). The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics*, 16(4), 197–212. Retrieved 27 February 2019 from <https://doi.org/10.1038/nrg3891>
- Albert, F. W., & Kruglyak, L. (2015b). The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics*, 16(4), 197–212. Retrieved 18 February 2022 from <https://doi.org/10.1038/nrg3891>
- Alliance, G., Screening, & Services, T. N. Y.-M.-A. C. for G. and N. (2009). Inheritance Patterns. Retrieved 30 April 2021 from <https://www.ncbi.nlm.nih.gov/books/NBK115561/>
- Alonso, L., Morán, I., Salvoró, C., & Torrents, D. (2021). In Search of Complex Disease Risk through Genome Wide Association Studies. *Mathematics*, 9(23), 3083. Retrieved 18 February 2022 from <https://doi.org/10.3390/MATH9233083>
- Alonso, L., Piron, A., Morán, I., & et al. (2021). TIGER: The gene expression regulatory variation landscape of human pancreatic islets. *Cell Reports*, 37(2), 109807. Retrieved 29 November 2021 from <https://doi.org/10.1016/j.celrep.2021.109807>
- Arendt, D. (2008). The evolution of cell types in animals: emerging principles from molecular studies. *Nature Reviews Genetics* 2008 9:11, 9(11), 868–882. Retrieved 18 February 2022 from <https://doi.org/10.1038/nrg2416>
- Avery, O. T., Macleod, C. M., & McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. *The Journal of Experimental Medicine*, 79(2), 137. Retrieved 18 February 2022 from <https://doi.org/10.1084/JEM.79.2.137>
- Band, G., & Marchini, J. (2018). BGEN: A binary file format for imputed genotype and haplotype data. *BioRxiv*, 18, 308296. Retrieved 11 March 2021 from <https://doi.org/10.1101/308296>
- Bartolomé, A. (2022). Stem Cell-Derived beta Cells: A Versatile Research Platform to Interrogate the Genetic Basis of beta Cell Dysfunction. *International Journal of Molecular Sciences* 2022, Vol. 23, Page 501, 23(1), 501. Retrieved 18 August 2022 from <https://doi.org/10.3390/IJMS23010501>
- Beck, T., Hastings, R. K., Gollapudi, S., Free, R. C., & Brookes, A. J. (2014). GWAS Central: A comprehensive resource for the comparison and interrogation of genome-wide association studies. *European Journal of Human Genetics*, 22(7), 949–952. Retrieved 30 April 2021 from <https://doi.org/10.1038/ejhg.2013.274>
- Behravan, H., Hartikainen, J. M., Tengström, M., Pylkäs, K., Winqvist, R., Kosma, V., & Mannermaa, A. (2018). Machine learning identifies interacting genetic variants contributing to breast cancer risk: A case study in Finnish cases and controls. *Scientific Reports*, 8(1), 13149. Retrieved 27 February 2019 from <https://doi.org/10.1038/s41598-018-31573-5>
- Bomba, L., Walter, K., & Soranzo, N. (2017). The impact of rare and low-frequency genetic variants in common disease. *Genome Biology* 2017 18:1, 18(1), 1–17. Retrieved 18 February 2022 from <https://doi.org/10.1186/S13059-017-1212-4>
- Bonàs-Guarch, S., Guindo-Martínez, M., Miguel-Escalada, I., Grarup, N., Sebastian, D., Rodríguez-Fos, E., ... Torrents, D. (2018). Re-analysis of public genetic data reveals a rare X-chromosomal variant associated with type 2 diabetes. *Nature Communications*, 9(1), 321. Retrieved 27 February 2019 from <https://doi.org/10.1038/s41467-017-02380-9>

- Bone, R. N., Tong, X., Weaver, S. A., Muralidharan, C., Krishnan, P., Kono, T., & Evans-Molina, C. (2021). Loss of Secretory Pathway Ca²⁺ ATPase (SPCA1) Impairs Insulin Secretion and Reduces Autophagy in the Pancreatic Islet. *BioRxiv*, 2021.08.30.458203. Retrieved 18 August 2022 from <https://doi.org/10.1101/2021.08.30.458203>
- Bookman, E. B., McAllister, K., Gillanders, E., Wanke, K., Balshaw, D., Rutter, J., ... Gunnar, M. R. (2011). Gene-environment interplay in common complex diseases: Forging an integrative model-Recommendations from an NIH workshop. *Genetic Epidemiology*, 35(4), 217–225. Retrieved 31 May 2021 from <https://doi.org/10.1002/gepi.20571>
- Boomsma, D. I., Wijmenga, C., Slagboom, E. P., Swertz, M. A., Karszen, L. C., Abdellaoui, A., ... Van Duijn, C. M. (2014). The Genome of the Netherlands: Design, and project goals. *European Journal of Human Genetics*, 22(2), 221–227. Retrieved 21 January 2021 from <https://doi.org/10.1038/ejhg.2013.118>
- Buccitelli, C., & Selbach, M. (2020). mRNAs, proteins and the emerging principles of gene expression control. *Nature Reviews Genetics*, 21(10), 630–644. Retrieved 18 February 2022 from <https://doi.org/10.1038/s41576-020-0258-4>
- Budi, E. H., Hoffman, S., Gao, S., Zhang, Y. E., & Derynck, R. (2019). Integration of TGF- β -induced Smad signaling in the insulin-induced transcriptional response in endothelial cells. *Scientific Reports*, 9(1), 1–16. Retrieved 2 March 2022 from <https://doi.org/10.1038/s41598-019-53490-x>
- Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., ... Parkinson, H. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, 47(D1), D1005–D1012. Retrieved 27 February 2019 from <https://doi.org/10.1093/nar/gky1120>
- Burton, P. R., Clayton, D. G., Cardon, L. R., Craddock, N., Deloukas, P., Duncanson, A., ... Compston, A. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145), 661–678. Retrieved 11 March 2021 from <https://doi.org/10.1038/nature05911>
- Burton, P. R., Tobin, M. D., & Hopper, J. L. (2005). Key concepts in genetic epidemiology. *The Lancet*, 366(9489), 941–951. Retrieved 18 February 2022 from [https://doi.org/10.1016/S0140-6736\(05\)67322-9](https://doi.org/10.1016/S0140-6736(05)67322-9)
- Cano-Gamez, E., & Trynka, G. (2020, May 13). From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Frontiers in Genetics*. Frontiers Media S.A. Retrieved 30 April 2021 from <https://doi.org/10.3389/fgene.2020.00424>
- Chang, C. C., Chow, C. C., Tellier, L. C. A. M., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, 4(1), 7. Retrieved 21 February 2022 from <https://doi.org/10.1007/s11892-022-01462-3>
- Chargaff, E. (1950). Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. *Experientia*, 6(6), 201–209. Retrieved 18 February 2022 from <https://doi.org/10.1007/BF02173653>
- Chen, J., Spracklen, C. N., Marenne, G., Varshney, A., Corbin, L. J., Luan, J., ... Barroso, I. (2021). The trans-ancestral genomic architecture of glycemic traits. *Nature Genetics*, 1–21. Retrieved 7 June 2021 from <https://doi.org/10.1038/s41588-021-00852-9>
- Chen, M.-H., Raffield, L. M., Mousas, A., Sakaue, S., Huffman, J., Moscati, A., ... Lettre, G. (2020). Trans-ethnic and Ancestry-Specific Blood-Cell Genetics in 746,667 Individuals from 5 Global Populations. *Cell*, 182(5), 1198–1213.e14. Retrieved 16 July 2021 from <https://doi.org/10.1016/J.CELL.2020.06.045>
- Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System*. Retrieved 27 February 2019 from <https://github.com/dmlc/xgboost>
- Chicco, D. (2017). Ten quick tips for machine learning in computational biology. *BioData Mining*, 10(1), 35. Retrieved 18 February 2022 from <https://doi.org/10.1186/S13040-017-0155-3>
- Cho, S. B., Kim, S. C., & Chung, M. G. (2019). Identification of novel population clusters with different susceptibilities to type 2 diabetes and their impact on the prediction of diabetes. *Scientific Reports*, 9(1), 1–9. Retrieved 18 February 2022 from <https://doi.org/10.1038/s41598-019-40058-y>
- Cho, Y. M., Ritchie, M. D., Moore, J. H., Park, J. Y., Lee, K. U., Shin, H. D., ... Park, K. S. (2004). Multifactor-dimensionality reduction shows a two-locus interaction associated with Type 2 diabetes mellitus. *Diabetologia*, 47(3), 549–554. Retrieved 18 February 2022 from <https://doi.org/10.1007/S00125-003-1321-3>
- Cleary, S., & Seoighe, C. (2021). Perspectives on Allele-Specific Expression. *Annual Review of Biomedical Data Science*, 4(1), 101–122. Retrieved 18 February 2022 from <https://doi.org/10.1146/ANNUREV-BIODATASCI-021621-122219>

- Cnop, M., Welsh, N., Jonas, J. C., Jörns, A., Lenzen, S., & Eizirik, D. L. (2005, December). Mechanisms of pancreatic β -cell death in type 1 and type 2 diabetes: Many differences, few similarities. *Diabetes*. Retrieved 11 March 2021 from https://doi.org/10.2337/diabetes.54.suppl_2.S97
- Cobb, M. (2014). Oswald Avery, DNA, and the transformation of biology. *Current Biology*, 24(2), R55–R60. Retrieved from <https://doi.org/10.1016/J.CUB.2013.11.060>
- Colditz, G. A., & Hankinson, S. E. (2005, May). The nurses' health study: Lifestyle and health among women. *Nature Reviews Cancer*. Nature Publishing Group. Retrieved 11 March 2021 from <https://doi.org/10.1038/nrc1608>
- Collins, F. S., Doudna, J. A., Lander, E. S., & Rotimi, C. N. (2021). Human Molecular Genetics and Genomics — Important Advances and Exciting Possibilities. *New England Journal of Medicine*, 384(1), 1–4. Retrieved 9 August 2022 from https://doi.org/10.1056/NEJMP2030694/SUPPL_FILE/NEJMP2030694_DISCLOSURES.PDF
- Cookson, W., Liang, L., Abecasis, G., Moffatt, M., & Lathrop, M. (2009). Mapping complex disease traits with global gene expression. *Nature Reviews Genetics*, 10(3), 184–194. Retrieved 18 February 2022 from <https://doi.org/10.1038/nrg2537>
- Cordell, H. J. (2009). Detecting gene–gene interactions that underlie human diseases. *Nature Reviews Genetics*, 10(6), 392–404. Retrieved 27 February 2019 from <https://doi.org/10.1038/nrg2579>
- Cornell, S. (2015). Continual evolution of type 2 diabetes: an update on pathophysiology and emerging treatment options. *Therapeutics and Clinical Risk Management*, 11, 621–632. Retrieved 18 February 2022 from <https://doi.org/10.2147/TCRM.S67387>
- Craig, J. (2008). Complex Diseases: Research and Applications | Learn Science at Scitable. Retrieved 18 February 2022, from <https://www.nature.com/scitable/topicpage/complex-diseases-research-and-applications-748/>
- Crick, F. H. (1958). On protein synthesis. *Symposia of the Society for Experimental Biology*, 12, 138–163.
- Croft, D., O'Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., ... Stein, L. (2011). Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Research*, 39(Database issue), D691. Retrieved 18 February 2022 from <https://doi.org/10.1093/NAR/GKQ1018>
- Dalkılıç, M. M. (2009). Gene Expression Arrays. *Encyclopedia of Database Systems*, 1218–1221. Retrieved 18 February 2022 from https://doi.org/10.1007/978-0-387-39940-9_1435
- DeForest, N., & Majithia, A. R. (2022). Genetics of Type 2 Diabetes: Implications from Large-Scale Studies. *Current Diabetes Reports*, 22(5), 227–235. Retrieved 30 June 2022 from <https://doi.org/10.1007/s11892-022-01462-3>
- DeFronzo, R. A. (2009). From the triumvirate to the ominous octet: A new paradigm for the treatment of type 2 diabetes mellitus. In *Diabetes* (Vol. 58, pp. 773–795). American Diabetes Association. Retrieved 1 February 2021 from <https://doi.org/10.2337/db09-9028>
- Del Guerra, S., Lupi, R., Marselli, L., Masini, M., Bugliani, M., Sbrana, S., ... Marchetti, P. (2005). Functional and molecular defects of pancreatic islets in human type 2 diabetes. *Diabetes*, 54(3), 727–735. Retrieved 11 March 2021 from <https://doi.org/10.2337/diabetes.54.3.727>
- Dempfle, A., Scherag, A., Hein, R., Beckmann, L., Chang-Claude, J., & Schäfer, H. (2008, June 4). Gene-environment interactions for complex traits: Definitions, methodological requirements and challenges. *European Journal of Human Genetics*. Retrieved 31 May 2021 from <https://doi.org/10.1038/ejhg.2008.106>
- Deplancke, B., Alpern, D., & Gardeux, V. (2016). The Genetics of Transcription Factor DNA Binding Variation. *Cell*, 166(3), 538–554. Retrieved 18 February 2022 from <https://doi.org/10.1016/J.CELL.2016.07.012>
- Dey, A. (2016). Machine Learning Algorithms: A Review. *International Journal of Computer Science and Information Technologies*, 7(3), 1174–1179.
- Dimas, A. S., Lagou, V., Barker, A., Knowles, J. W., Mägi, R., Hivert, M. F., ... Prokopenko, I. (2014). Impact of type 2 diabetes susceptibility variants on quantitative glycemic traits reveals mechanistic heterogeneity. *Diabetes*, 63(6), 2158–2171. Retrieved 18 February 2022 from <https://doi.org/10.2337/DB13-0949>
- Dorsey-Trevino, E. G., Kaur, V., Mercader, J. M., Florez, J. C., & Leong, A. (2022). Association of GLP1R Polymorphisms With the Incretin Response. *The Journal of Clinical Endocrinology & Metabolism*. Retrieved 18 August 2022 from <https://doi.org/10.1210/CLINEM/DGAC374>
- Down, T. A., Piipari, M., & Hubbard, T. J. P. (2011). Dalliance: Interactive genome viewing on the web. *Bioinformatics*, 27(6), 889–890. Retrieved 21 January 2021 from

- <https://doi.org/10.1093/bioinformatics/btr020>
- Eichler, E. E. (2019). Genetic Variation, Comparative Genomics, and the Diagnosis of Disease. *The New England Journal of Medicine*, 381(1), 64. Retrieved 18 February 2022 from <https://doi.org/10.1056/NEJMRA1809315>
- Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., ... Turner, S. (2009). Real-time DNA sequencing from single polymerase molecules. *Science*, 323(5910), 133–138. Retrieved 18 February 2022 from <https://doi.org/10.1126/science.1162986>
- Eizirik, D. L., Pasquali, L., & Cnop, M. (2020, July 1). Pancreatic β -cells in type 1 and type 2 diabetes mellitus: different pathways to failure. *Nature Reviews Endocrinology*. Nature Research. Retrieved 22 January 2021 from <https://doi.org/10.1038/s41574-020-0355-7>
- Escaramís, G., Docampo, E., & Rabionet, R. (2015). A decade of structural variants: description, history and methods to detect structural variation. *Briefings in Functional Genomics*, 14(5), 305–314. Retrieved 18 February 2022 from <https://doi.org/10.1093/BFGP/ELV014>
- Fadista, J., Vikman, P., Laakso, E. O., Mollet, I. G., Esguerra, J. Lou, Taneera, J., ... Groop, L. (2014). Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism. *Proceedings of the National Academy of Sciences of the United States of America*, 111(38), 13924–9. Retrieved 27 February 2019 from <https://doi.org/10.1073/pnas.1402665111>
- Flannick, J., & Florez, J. C. (2016). Type 2 diabetes: genetic data sharing to advance complex disease research. *Nature Reviews Genetics*. Retrieved from <https://doi.org/10.1038/nrg.2016.56>
- Flannick, J., Johansson, S., & Njølstad, P. R. (2016). Common and rare forms of diabetes mellitus: towards a continuum of diabetes subtypes. *Nature Reviews Endocrinology*, 12(7), 394–406. Retrieved 18 February 2022 from <https://doi.org/10.1038/nrendo.2016.50>
- Frankish, A., Diekhans, M., Ferreira, A. M., Johnson, R., Jungreis, I., Loveland, J., ... Flicek, P. (2019). GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research*, 47(D1), D766–D773. Retrieved 21 January 2021 from <https://doi.org/10.1093/nar/gky955>
- Galicia-García, U., Benito-Vicente, A., Jebari, S., Larrea-Sebal, A., Siddiqi, H., Uribe, K. B., ... Martín, C. (2020). Pathophysiology of Type 2 Diabetes Mellitus. *International Journal of Molecular Sciences*, 21(17), 1–34. Retrieved 18 February 2022 from <https://doi.org/10.3390/IJMS21176275>
- García-Sánchez, A., & Marqués-García, F. (2016). Review of Methods to Study Gene Expression Regulation Applied to Asthma. *Methods in Molecular Biology*, 1434, 71–89. Retrieved 18 February 2022 from https://doi.org/10.1007/978-1-4939-3652-6_6
- Génin, E. (2020). Missing heritability of complex diseases: case solved? *Human Genetics*, 139(1), 103–113. Retrieved 27 October 2021 from <https://doi.org/10.1007/S00439-019-02034-4>
- Ghosh, S., Watanabe, R. M., Valle, T. T., Hauser, E. R., Magnuson, V. L., Langefeld, C. D., ... Boehnke, M. (2000). The Finland-United States investigation of non-insulin-dependent diabetes mellitus genetics (FUSION) study. I. An autosomal genome scan for genes that predispose to type 2 diabetes. *American Journal of Human Genetics*, 67(5), 1174–1185. Retrieved 11 March 2021 from [https://doi.org/10.1016/S0002-9297\(07\)62948-6](https://doi.org/10.1016/S0002-9297(07)62948-6)
- Gilon, P. (2020). The Role of α -Cells in Islet Function and Glucose Homeostasis in Health and Type 2 Diabetes. *Journal of Molecular Biology*, 432(5), 1367–1394. Retrieved from <https://doi.org/10.1016/J.JMB.2020.01.004>
- Gloyn, A. L., Ibberson, M., Marchetti, P., Powers, A. C., Rorsman, P., Sander, M., & Solimena, M. (2022). Every islet matters: improving the impact of human islet research. *Nature Metabolism* 2022, 1–8. Retrieved 18 August 2022 from <https://doi.org/10.1038/s42255-022-00607-8>
- González, J. R., Ruiz-Arenas, C., Cáceres, A., Morán, I., López-Sánchez, M., Alonso, L., ... Pérez-Jurado, L. A. (2020). Polymorphic Inversions Underlie the Shared Genetic Susceptibility of Obesity-Related Diseases. *American Journal of Human Genetics*, 106(6), 846–858. Retrieved from <https://doi.org/10.1016/j.ajhg.2020.04.017>
- Gonzalo, R., & Sánchez, A. (2018). Introduction to Microarrays Technology and Data Analysis. *Comprehensive Analytical Chemistry*, 82, 37–69. Retrieved from <https://doi.org/10.1016/BS.COAC.2018.08.002>
- Gottesman, O., Kuivaniemi, H., Tromp, G., Faucett, W. A., Li, R., Manolio, T. A., ... Williams, M. S. (2013, October). The Electronic Medical Records and Genomics (eMERGE) Network: Past, present, and future. *Genetics in Medicine*. Genet Med. Retrieved 11 March 2021 from <https://doi.org/10.1038/gim.2013.72>
- Greener, J. G., Kandathil, S. M., Moffat, L., & Jones, D. T. (2021). A guide to machine learning for biologists. *Nature Reviews Molecular Cell Biology*, 23(1), 40–55. Retrieved 18 February 2022

- from <https://doi.org/10.1038/s41580-021-00407-0>
- Grundberg, E., Small, K. S., Hedman, Å. K., Nica, A. C., Buil, A., Keildson, S., ... Spector, T. D. (2012). Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nature Genetics*, 44(10), 1084–1089. Retrieved 18 February 2022 from <https://doi.org/10.1038/ng.2394>
- Guindo-Martínez, M., Amela, R., & et al. (2021). The impact of non-additive genetic associations on age-related complex diseases. *Nature Communications*, 12(1), 2436. Retrieved 26 April 2021 from <https://doi.org/10.1038/s41467-021-21952-4>
- Haeussler, M., Zweig, A. S., Tyner, C., Speir, M. L., Rosenbloom, K. R., Raney, B. J., ... Kent, W. J. (2019). The UCSC Genome Browser database: 2019 update. *Nucleic Acids Research*, 47(D1), D853–D858. Retrieved 21 January 2021 from <https://doi.org/10.1093/nar/gky1095>
- Hall, E., Volkov, P., Dayeh, T., Esguerra, J. L. S., Salö, S., Eliasson, L., ... Ling, C. (2014). Sex differences in the genome-wide DNA methylation pattern and impact on gene expression, microRNA levels and insulin secretion in human pancreatic islets. *Genome Biology*, 15(12), 522. Retrieved 27 February 2019 from <https://doi.org/10.1186/s13059-014-0522-z>
- Han, H., Shim, H., Shin, D., Shim, J. E., Ko, Y., Shin, J., ... Lee, I. (2015). TRRUST: a reference database of human transcriptional regulatory interactions. *Scientific Reports*, 5(1), 1–11. Retrieved 18 February 2022 from <https://doi.org/10.1038/srep11432>
- Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., ... Hubbard, T. J. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Research*, 22(9), 1760–1774. Retrieved 18 February 2022 from <https://doi.org/10.1101/GR.135350.111>
- Hemani, G., Shakhbazov, K., Westra, H. J., Esko, T., Henders, A. K., McRae, A. F., ... Powell, J. E. (2021). Retraction Note: Detection and replication of epistasis influencing transcription in humans. *Nature*, 596(7871), 306–306. Retrieved 21 February 2022 from <https://doi.org/10.1038/s41586-021-03766-y>
- Herzig, A. F., Clerget-Darpoux, F., & Génin, E. (2022). The False Dawn of Polygenic Risk Scores for Human Disease Prediction. *Journal of Personalized Medicine* 2022, Vol. 12, Page 1266, 12(8), 1266. Retrieved 9 August 2022 from <https://doi.org/10.3390/JPM12081266>
- Horiba, N., Masuda, S., Ohnishi, C., Takeuchi, D., Okuda, M., & Inui, K. I. (2003). Na⁺-dependent fructose transport via rNaGLT1 in rat kidney. *FEBS Letters*, 546(2–3), 276–280. Retrieved 12 August 2022 from [https://doi.org/10.1016/S0014-5793\(03\)00600-8](https://doi.org/10.1016/S0014-5793(03)00600-8)
- Hu, F. B. (2011). Globalization of diabetes: the role of diet, lifestyle, and genes. *Diabetes Care*, 34(6), 1249–1257. Retrieved 18 February 2022 from <https://doi.org/10.2337/DC11-0442>
- International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature* 2001 409:6822, 409(6822), 860–921. Retrieved 20 September 2021 from <https://doi.org/10.1038/35057062>
- International Human Genome Sequencing Consortium. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011), 931–945. Retrieved 31 May 2021 from <https://doi.org/10.1038/nature03001>
- Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., ... Loose, M. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology*, 36(4), 338–345. Retrieved 18 February 2022 from <https://doi.org/10.1038/NBT.4060>
- Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., ... D'Eustachio, P. (2020). The reactome pathway knowledgebase. *Nucleic Acids Research*, 48(D1), D498–D503. Retrieved 21 January 2021 from <https://doi.org/10.1093/nar/gkz1031>
- Ji, Z., Lu, M., Xie, H., Yuan, H., & Chen, Q. (2022). β cell regeneration and novel strategies for treatment of diabetes (Review). *Biomedical Reports*, 17(3), 1–5. Retrieved 11 July 2022 from <https://doi.org/10.3892/BR.2022.1555>
- Jiang, C., Xuan, Z., Zhao, F., & Zhang, M. Q. (2007). TRED: a transcriptional regulatory element database, new entries and other development. *Nucleic Acids Research*, 35(Database issue), D137. Retrieved 18 February 2022 from <https://doi.org/10.1093/NAR/GKL1041>
- Kaku, K. (2010). Pathophysiology of Type 2 Diabetes and Its Treatment Policy. *Research and Reviews*, 53, 41–46.
- Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., ... MacArthur, D. G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581(7809), 434–443. Retrieved 21 January 2021 from <https://doi.org/10.1038/s41586-020-2308-7>
- Kawasaki, E. S. (2004). Microarrays and the gene expression profile of a single cell. *Annals of the New York Academy of Sciences*, 1020, 92–100. Retrieved 18 February 2022 from <https://doi.org/10.1196/ANNALS.1310.010>

- Kim, H., Westerman, K. E., Smith, K., Chiou, J., Cole, J. B., Majarian, T., ... Udler, M. S. (2022). High-throughput Genetic Clustering of Type 2 Diabetes Loci Reveals Heterogeneous Mechanistic Pathways of Metabolic Disease. *MedRxiv*, 2022.07.11.22277436. Retrieved 18 August 2022 from <https://doi.org/10.1101/2022.07.11.22277436>
- Kim, T. K., & Shiekhhattar, R. (2015). Architectural and Functional Commonalities between Enhancers and Promoters. *Cell*, 162(5), 948–959. Retrieved 18 February 2022 from <https://doi.org/10.1016/J.CELL.2015.08.008>
- Kirino, Y., Bertias, G., Ishigatsubo, Y., Mizuki, N., Tugal-Tutkun, I., Seyahi, E., ... Kastner, D. L. (2013). Genome-wide association analysis identifies new susceptibility loci for Behçet's disease and epistasis between HLA-B*51 and ERAP1. *Nature Genetics*, 45(2), 202–207. Retrieved 27 February 2019 from <https://doi.org/10.1038/ng.2520>
- Klein, N. de, Dijk, F. van, Deelen, P., Urzua, C. G., Claringbould, A., Vösa, U., ... Franke, L. (2020). Imbalanced expression for predicted high-impact, autosomal-dominant variants in a cohort of 3,818 healthy samples. *BioRxiv*, 2020.09.19.300095. Retrieved 18 February 2022 from <https://doi.org/10.1101/2020.09.19.300095>
- Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J.-Y., Sackler, R. S., Haynes, C., ... Hoh., J. (2005). Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science*, 308(5720), 385. Retrieved 21 July 2021 from <https://doi.org/10.1126/SCIENCE.1109557>
- Ku, C. S., Loy, E. Y., Salim, A., Pawitan, Y., & Chia, K. S. (2010). The discovery of human genetic variations and their use as disease markers: past, present and future. *Journal of Human Genetics* 2010 55:7, 55(7), 403–415. Retrieved 18 February 2022 from <https://doi.org/10.1038/jhg.2010.55>
- Kullo, I. J., Lewis, C. M., Inouye, M., Martin, A. R., Ripatti, S., & Chatterjee, N. (2022). Polygenic scores in biomedical research. *Nature Reviews Genetics* 2022 23:9, 23(9), 524–532. Retrieved 17 August 2022 from <https://doi.org/10.1038/s41576-022-00470-z>
- Kumuthini, J., Zick, B., Balasopoulou, A., Chalikiopoulou, C., Dandara, C., El-Kamah, G., ... Abramowicz, M. (2022). The clinical utility of polygenic risk scores in genomic medicine practices: a systematic review. *Human Genetics* 2022, 1–8. Retrieved 9 August 2022 from <https://doi.org/10.1007/S00439-022-02452-X>
- Kvale, M. N., Hesselton, S., Hoffmann, T. J., Cao, Y., Chan, D., Connell, S., ... Risch, N. (2015). Genotyping informatics and quality control for 100,000 subjects in the genetic epidemiology research on adult health and aging (GERA) cohort. *Genetics*, 200(4), 1051–1060. Retrieved 11 March 2021 from <https://doi.org/10.1534/genetics.115.178905>
- LaFramboise, T. (2009). Single nucleotide polymorphism arrays: A decade of biological, computational and technological advances. *Nucleic Acids Research*. Oxford University Press. Retrieved 31 May 2021 from <https://doi.org/10.1093/nar/gkp552>
- Lambert, S. A., Abraham, G., & Inouye, M. (2019, November 21). Towards clinical utility of polygenic risk scores. *Human Molecular Genetics*. Oxford University Press. Retrieved 3 May 2021 from <https://doi.org/10.1093/hmg/ddz187>
- Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., ... Weirauch, M. T. (2018). The Human Transcription Factors. *Cell*, 172(4), 650–665. Retrieved 18 February 2022 from <https://doi.org/10.1016/j.cell.2018.01.029>
- Lamy, P., Grove, J., & Wiuf, C. (2011, May 1). A review of software for microarray genotyping. *Human Genomics*. BioMed Central Ltd. Retrieved 31 May 2021 from <https://doi.org/10.1186/1479-7364-5-4-304>
- Lee, C., Kang, E. Y., Gandal, M. J., Eskin, E., & Geschwind, D. H. (2019). Profiling allele-specific gene expression in brains from individuals with autism spectrum disorder reveals preferential minor allele usage. *Nature Neuroscience*, 22(9), 1521–1532. Retrieved 18 February 2022 from <https://doi.org/10.1038/s41593-019-0461-9>
- Lettre, G., Lange, C., & Hirschhorn, J. N. (2007). Genetic model testing and statistical power in population-based association studies of quantitative traits. *Genetic Epidemiology*, 31(4), 358–362. Retrieved 18 February 2022 from <https://doi.org/10.1002/GEPI.20217>
- Levene, P., Biochem, Z., Levene, P. A., Jacobs, W. A., Ber, T., Ges ; Levene, P. A., ... Germann, W. (1919). *The Structure of Yeast Nucleic Acid. IV. Ammonia Hydrolysis. Ber. them. Ges* (Vol. 608).
- Levi, M., Myakala, K., & Wang, X. (2018). SRGAP2a: A New Player That Modulates Podocyte Cytoskeleton and Injury in Diabetes. *Diabetes*, 67(4), 550. Retrieved 11 August 2022 from <https://doi.org/10.2337/DBI17-0050>
- Lichou, F., & Trynka, G. (2020, December 1). Functional studies of GWAS variants are gaining momentum. *Nature Communications*. Nature Research. Retrieved 11 March 2021 from <https://doi.org/10.1038/s41467-020-20188-y>

- Lin, Z., Feng, W., Liu, Y., Ma, C., Arefan, D., Zhou, D., ... Qu, S. (2021). Machine Learning to Identify Metabolic Subtypes of Obesity: A Multi-Center Study. *Frontiers in Endocrinology*, 12. Retrieved 18 February 2022 from <https://doi.org/10.3389/fendo.2021.713592>
- Liu, W., Zhuang, Z., Wang, W., Huang, T., & Liu, Z. (2021). An Improved Genome-Wide Polygenic Score Model for Predicting the Risk of Type 2 Diabetes. *Frontiers in Genetics*, 12, 63. Retrieved from <https://doi.org/10.3389/FGENE.2021.632385/XML/NLM>
- Lo, C. (2014). *Algorithms for Haplotype Phasing*. Retrieved 30 April 2021 from <https://cseweb.ucsd.edu/~chl107/pubs/re.pdf>
- Long, H. K., Prescott, S. L., & Wysocka, J. (2016). Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. *Cell*, 167(5), 1170–1187. Retrieved 18 February 2022 from <https://doi.org/10.1016/J.CELL.2016.09.018>
- Luft, J., Young, R. S., Meynert, A. M., & Taylor, M. S. (2020). Detecting oncogenic selection through biased allele retention in The Cancer Genome Atlas. *BioRxiv*, 2020.07.03.186593. Retrieved 18 February 2022 from <https://doi.org/10.1101/2020.07.03.186593>
- Machiela, M. J., & Chanock, S. J. (2015). LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics*, 31(21), 3555–3557. Retrieved 21 February 2022 from <https://doi.org/10.1093/BIOINFORMATICS/BTV402>
- Mackay, T. F. C. (2014). Epistasis and quantitative traits: using model organisms to study gene-gene interactions. *Nature Reviews. Genetics*, 15(1), 22–33. Retrieved 27 February 2019 from <https://doi.org/10.1038/nrg3627>
- Magkos, F., Hjorth, M. F., & Astrup, A. (2020). Diet and exercise in the prevention and treatment of type 2 diabetes mellitus. *Nature Reviews Endocrinology*, 16(10), 545–555. Retrieved 18 February 2022 from <https://doi.org/10.1038/s41574-020-0381-5>
- Mahajan, A., Taliun, D., Thurner, M., Robertson, N. R., Torres, J. M., Rayner, N. W., ... McCarthy, M. I. (2018). Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nature Genetics*, 50(11), 1505–1513. Retrieved 27 February 2019 from <https://doi.org/10.1038/s41588-018-0241-6>
- Mahajan, A., Wessel, J., Willems, S. M., Zhao, W., Robertson, N. R., Chu, A. Y., ... McCarthy, M. I. (2018). Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes. *Nature Genetics*, 50(4), 559–571. Retrieved 18 February 2022 from <https://doi.org/10.1038/S41588-018-0084-1>
- Manduchi, E., Chesi, A., Hall, M. A., Grant, S. F. A., & Moore, J. H. (2018). Leveraging putative enhancer-promoter interactions to investigate two-way epistasis in Type 2 Diabetes GWAS. In *Pacific Symposium on Biocomputing* (Vol. 0, pp. 548–558). World Scientific Publishing Co. Pte Ltd. Retrieved 10 March 2021 from https://doi.org/10.1142/9789813235533_0050
- Manolio, T. A. (2013). Bringing genome-wide association findings into clinical use. *Nature Reviews. Genetics*, 14(8), 549–558. Retrieved 18 February 2022 from <https://doi.org/10.1038/NRG3523>
- Manolio, T. A., Brooks, L. D., & Collins, F. S. (2008, May 1). A HapMap harvest of insights into the genetics of common disease. *Journal of Clinical Investigation*. American Society for Clinical Investigation. Retrieved 31 May 2021 from <https://doi.org/10.1172/JCI34772>
- Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., ... Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265), 747–753. Retrieved 27 February 2019 from <https://doi.org/10.1038/nature08494>
- Mansour Aly, D., Dwivedi, O. P., Prasad, R. B., Käräjämäki, A., Hjort, R., Thangam, M., ... Ahlqvist, E. (2021). Genome-wide association analyses highlight etiological differences underlying newly defined subtypes of diabetes. *Nature Genetics*, 53(11), 1534–1542. Retrieved 18 February 2022 from <https://doi.org/10.1038/s41588-021-00948-2>
- Marchini, J. (2019). Haplotype Estimation and Genotype Imputation. In *Handbook of Statistical Genomics* (Vol. 1, pp. 87–114). Wiley. Retrieved 30 April 2021 from <https://doi.org/10.1002/9781119487845.ch3>
- Marchini, J., Donnelly, P., & Cardon, L. R. (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genetics*, 37(4), 413–417. Retrieved 29 March 2019 from <https://doi.org/10.1038/ng1537>
- McAllister, K., Mechanic, L. E., Amos, C., Aschard, H., Blair, I. A., Chatterjee, N., ... Witte, J. S. (2017). Current Challenges and New Opportunities for Gene-Environment Interaction Studies of Complex Diseases. *American Journal of Epidemiology*, 186(7), 753–761. Retrieved 31 May 2021 from <https://doi.org/10.1093/aje/kwx227>
- McCarthy, M. I. (2010). Genomics, Type 2 Diabetes, and Obesity. *New England Journal of Medicine*, 363(24), 2339–2350. Retrieved 27 February 2019 from <https://doi.org/10.1056/NEJMra0906948>

- McCarthy, M. I. (2017). Painting a new picture of personalised medicine for diabetes. *Diabetologia*, 60(5), 793. Retrieved 18 February 2022 from <https://doi.org/10.1007/S00125-017-4210-X>
- McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A., & Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics* 2008 9:5, 9(5), 356–369. Retrieved 20 September 2021 from <https://doi.org/10.1038/NRG2344>
- McGuire, A. L., Gabriel, S., Tishkoff, S. A., Wonkam, A., Chakravarti, A., Furlong, E. E. M., ... Kim, J. S. (2020, October 1). The road ahead in genetics and genomics. *Nature Reviews Genetics*. Nature Research. Retrieved 10 March 2021 from <https://doi.org/10.1038/s41576-020-0272-6>
- McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., ... Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, 17(1), 122. Retrieved 21 January 2021 from <https://doi.org/10.1186/s13059-016-0974-4>
- Mercader, Josep M., Liao, R. G., Bell, A. D., Dymek, Z., Estrada, K., Tukiainen, T., ... Florez, J. C. (2017). A loss-of-function splice acceptor variant in *igf2* is protective for type 2 diabetes. *Diabetes*, 66(11), 2903–2914. Retrieved 14 September 2022 from <https://doi.org/10.2337/DB17-0187/-/DC1>
- Mercader, Josep Maria, & Florez, J. C. (2017). The Genetic Basis of Type 2 Diabetes in Hispanics and Latin Americans: Challenges and Opportunities. *Frontiers in Public Health*, 5. Retrieved 16 July 2021 from <https://doi.org/10.3389/FPUH.2017.00329>
- Mercader, Josep Maria, Saus, E., Agüera, Z., Bayés, M., Boni, C., Carreras, A., ... Estivill, X. (2008). Association of NTRK3 and its interaction with NGF suggest an altered cross-regulation of the neurotrophin signaling pathway in eating disorders. *Human Molecular Genetics*, 17(9), 1234–1244. Retrieved from <https://doi.org/10.1093/HMG/DDN013>
- Meyerovich, K., Ortis, F., & Cardozo, A. K. (2018). The non-canonical NF- κ B pathway and its contribution to β -cell failure in diabetes. *Journal of Molecular Endocrinology*, 61(2), F1–F6. Retrieved 12 August 2022 from <https://doi.org/10.1530/JME-16-0183>
- Miga, K. H., & Wang, T. (2021). The Need for a Human Pangenome Reference Sequence. *Annual Review of Genomics and Human Genetics*, 22, 81–102. Retrieved 28 October 2021 from <https://doi.org/10.1146/ANNUREV-GENOM-120120-081921>
- Miguel-Escalada, I., Bonàs-Guarch, S., Cebola, I., Ponsa-Cobas, J., Mendieta-Esteban, J., Atla, G., ... Ferrer, J. (2019). Human pancreatic islet three-dimensional chromatin architecture provides insights into the genetics of type 2 diabetes. *Nature Genetics*, 51(7), 1137–1148. Retrieved 30 July 2019 from <https://doi.org/10.1038/s41588-019-0457-0>
- Mitchell, K. J. (2012, January 23). What is complex about complex disorders? *Genome Biology*. BioMed Central. Retrieved 10 March 2021 from <https://doi.org/10.1186/gb-2012-13-1-237>
- Monaco, A., Pantaleo, E., Amoroso, N., Lacalamita, A., Lo Giudice, C., Fonzino, A., ... Bellotti, R. (2021). A primer on machine learning techniques for genomic applications. *Computational and Structural Biotechnology Journal*, 19, 4345–4359. Retrieved from <https://doi.org/10.1016/J.CSBJ.2021.07.021>
- Monir, M. M., & Zhu, J. (2017). Comparing GWAS Results of Complex Traits Using Full Genetic Model and Additive Models for Revealing Genetic Architecture. *Scientific Reports*, 7(1), 38600. Retrieved 27 February 2019 from <https://doi.org/10.1038/srep38600>
- Morán, I., Akerman, I., Van De Bunt, M., Xie, R., Benazra, M., Nammo, T., ... Ferrer, J. (2012). Human β cell transcriptome analysis uncovers lncRNAs that are tissue-specific, dynamically regulated, and abnormally expressed in type 2 diabetes. *Cell Metabolism*, 16(4), 435–448. Retrieved 18 February 2022 from <https://doi.org/10.1016/J.CMET.2012.08.010>
- Mularoni, L., Ramos-Rodríguez, M., & Pasquali, L. (2017). The Pancreatic Islet Regulome Browser. *Frontiers in Genetics*, 8(FEB). Retrieved 22 September 2022 from <https://doi.org/10.3389/FGENE.2017.00013>
- Nag, A., McCarthy, M. I., & Mahajan, A. (2020). Large-Scale Analyses Provide No Evidence for Gene-Gene Interactions Influencing Type 2 Diabetes Risk. *Diabetes*, 69(11), 2518–2522. Retrieved 19 February 2022 from <https://doi.org/10.2337/DB20-0224>
- Nathan, D. M., Buse, J. B., Davidson, M. B., Ferrannini, E., Holman, R. R., Sherwin, R., & Zinman, B. (2009). Medical Management of Hyperglycemia in Type 2 Diabetes: A Consensus Algorithm for the Initiation and Adjustment of Therapy: A consensus statement of the American Diabetes Association and the European Association for the Study of Diabetes. *Diabetes Care*, 32(1), 193–203. Retrieved 18 February 2022 from <https://doi.org/10.2337/DC08-9025>
- Newman, B., Selby, J. V., King, M. C., Slemenda, C., Fabsitz, R., & Friedman, G. D. (1987). Concordance for type 2 (non-insulin-dependent) diabetes mellitus in male twins. *Diabetologia*, 30(10), 763–768. Retrieved 18 February 2022 from <https://doi.org/10.1007/BF00275741>

- Nica, A. C., & Dermitzakis, E. T. (2013). Expression quantitative trait loci: present and future. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 368(1620), 20120362. Retrieved 23 July 2019 from <https://doi.org/10.1098/rstb.2012.0362>
- Nicholls, H. L., John, C. R., Watson, D. S., Munroe, P. B., Barnes, M. R., & Cabrera, C. P. (2020). Reaching the End-Game for GWAS: Machine Learning Approaches for the Prioritization of Complex Disease Loci. *Frontiers in Genetics*, 11, 350. Retrieved from <https://doi.org/10.3389/fgene.2020.00350>
- Niel, C., Sinoquet, C., Dina, C., & Rocheleau, G. (2015). A survey about methods dedicated to epistasis detection. *Frontiers in Genetics*, 6(SEP), 285. Retrieved from <https://doi.org/10.3389/fgene.2015.00285>
- Nowakowska, M., Zghebi, S. S., Ashcroft, D. M., Buchan, I., Chew-Graham, C., Holt, T., ... Kontopantelis, E. (2019). The comorbidity burden of type 2 diabetes mellitus: patterns, clusters and predictions from a large English primary care cohort. *BMC Medicine*, 17(1). Retrieved 18 February 2022 from <https://doi.org/10.1186/S12916-019-1373-Y>
- O'Connor, M. J., Schroeder, P., Huerta-Chagoya, A., Cortés-Sánchez, P., Bonàs-Guarch, S., Guindo-Martínez, M., ... Mercader, J. M. (2022). Recessive Genome-Wide Meta-analysis Illuminates Genetic Architecture of Type 2 Diabetes. *Diabetes*, 71(3), 554–565. Retrieved 18 August 2022 from <https://doi.org/10.2337/DB21-0545>
- O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufu, S., Haddad, D., McVeigh, R., ... Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44(D1), D733–D745. Retrieved 21 January 2021 from <https://doi.org/10.1093/nar/gkv1189>
- Orozco, L. D., Farrell, C., Hale, C., Rubbi, L., Rinaldi, A., Civelek, M., ... Pellegrini, M. (2018). Epigenome-wide association in adipose tissue from the METSIM cohort. *Human Molecular Genetics*, 27(10), 1830. Retrieved 12 August 2022 from <https://doi.org/10.1093/HMG/DDY093>
- Padilla-Martínez, F., Collin, F., Kwasniewski, M., & Kretowski, A. (2020). Systematic Review of Polygenic Risk Scores for Type 1 and Type 2 Diabetes. *International Journal of Molecular Sciences*, 21(5). Retrieved 22 September 2022 from <https://doi.org/10.3390/IJMS21051703>
- Panagiotou, O. A., Willer, C. J., Hirschhorn, J. N., & Ioannidis, J. P. A. (2013). The Power of Meta-Analysis in Genome-Wide Association Studies. *Annual Review of Genomics and Human Genetics*, 14(1), 441–465. Retrieved 30 April 2021 from <https://doi.org/10.1146/annurev-genom-091212-153520>
- Papatheodorou, I., Moreno, P., Manning, J., Fuentes, A. M. P., George, N., Fexova, S., ... Brazma, A. (2020). Expression Atlas update: From tissues to single cells. *Nucleic Acids Research*, 48(D1), D77–D83. Retrieved 10 March 2021 from <https://doi.org/10.1093/nar/gkz947>
- Pasquali, L., Gaulton, K. J., Rodríguez-Seguí, S. A., Mularoni, L., Miguel-Escalada, I., Akerman, I., ... Ferrer, J. (2014). Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nature Genetics*, 46(2), 136–143. Retrieved 27 February 2019 from <https://doi.org/10.1038/ng.2870>
- Pedregosa, F., Michel, V., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., ... Duchesnay, É. (2011). *Scikit-learn: Machine Learning in Python*. *Journal of Machine Learning Research* (Vol. 12). Retrieved 30 April 2021 from <http://scikit-learn.sourceforge.net>
- Pell, N., Garcia-Pras, E., Gallego, J., Naranjo-Suarez, S., Balvey, A., Suñer, C., ... Fernandez, M. (2021). Targeting the cytoplasmic polyadenylation element-binding protein CPEB4 protects against diet-induced obesity and microbiome dysbiosis. *Molecular Metabolism*, 54, 101388. Retrieved from <https://doi.org/10.1016/J.MOLMET.2021.101388>
- Piñero, J., Bravo, Á., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., ... Furlong, L. I. (2017). DisGeNET: A comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Research*, 45(D1), D833–D839. Retrieved 21 January 2021 from <https://doi.org/10.1093/nar/gkw943>
- Pope, S. D., & Medzhitov, R. (2018). Emerging Principles of Gene Expression Programs and Their Regulation. *Molecular Cell*, 71(3), 389–397. Retrieved 18 February 2022 from <https://doi.org/10.1016/J.MOLCEL.2018.07.017>
- Pozarickij, A., Williams, C., & Guggenheim, J. A. (2020). Non-additive (dominance) effects of genetic variants associated with refractive error and myopia. *Molecular Genetics and Genomics*, 295(4), 843–853. Retrieved 18 February 2022 from <https://doi.org/10.1007/s00438-020-01666-w>
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., ... Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, 81(3), 559–575. Retrieved 21 January 2021 from <https://doi.org/10.1086/519795>

- Quirke, V., & Gaudillière, J. P. (2008). The Era of Biomedicine: Science, Medicine, and Public Health in Britain and France after the Second World War. *Medical History*, 52(4), 441. Retrieved 5 August 2022 from <https://doi.org/10.1017/S002572730000017X>
- Reardon, S. (2021). A complete human genome sequence is close: how scientists filled in the gaps. *Nature*, 594(7862), 158–159. Retrieved from <https://doi.org/10.1038/D41586-021-01506-W>
- Rhodes, C. J. (2005). Type 2 Diabetes—a Matter of β -Cell Life and Death? *Science*, 307(5708), 380–384. Retrieved 18 February 2022 from <https://doi.org/10.1126/SCIENCE.1104345>
- Rizzo, H. E., Escaname, E. N., Alana, N. B., Lavender, E., Gelfond, J., Fernandez, R., ... Blanco, C. L. (2020). Maternal diabetes and obesity influence the fetal epigenome in a largely Hispanic population. *Clinical Epigenetics*, 12(1), 1–10. Retrieved 2 March 2022 from <https://doi.org/10.1186/s13148-020-0824-9>
- Rouillard, A. D., Gundersen, G. W., Fernandez, N. F., Wang, Z., Monteiro, C. D., McDermott, M. G., & Ma'ayan, A. (2016). The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database*, 2016. Retrieved 12 August 2022 from <https://doi.org/10.1093/DATABASE/BAW100>
- Saeedi, P., Petersohn, I., Salpea, P., Malanda, B., Karuranga, S., Unwin, N., ... Williams, R. (2019). Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th edition. *Diabetes Research and Clinical Practice*, 157, 107843. Retrieved 22 January 2021 from <https://doi.org/10.1016/j.diabres.2019.107843>
- Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12), 5463–5467. Retrieved 18 February 2022 from <https://doi.org/10.1073/PNAS.74.12.5463>
- Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, 2(3), 1–21. Retrieved 18 February 2022 from <https://doi.org/10.1007/S42979-021-00592-X>
- Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235), 467–470. Retrieved 18 February 2022 from <https://doi.org/10.1126/SCIENCE.270.5235.467>
- Scott, R. A., Scott, L. J., Mägi, R., Marullo, L., Gaulton, K. J., Kaakinen, M., ... Prokopenko, I. (2017). An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans. *Diabetes*, 66(11), 2888–2902. Retrieved 21 January 2021 from <https://doi.org/10.2337/db16-1253>
- Sebastian, A., Contreras-Moreira, B., Araid, F. N., Agustín, P. M., & Zaragoza, S. (2014). footprintDB: a database of transcription factors with annotated cis elements and binding interfaces. *Bioinformatics*, 30(2), 258–265. Retrieved 18 February 2022 from <https://doi.org/10.1093/BIOINFORMATICS/BTT663>
- Sender, R., Fuchs, S., & Milo, R. (2016). Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLOS Biology*, 14(8), e1002533. Retrieved 18 February 2022 from <https://doi.org/10.1371/JOURNAL.PBIO.1002533>
- Shendure, J., Balasubramanian, S., Church, G. M., Gilbert, W., Rogers, J., Schloss, J. A., & Waterston, R. H. (2017). DNA sequencing at 40: past, present and future. *Nature* 2017 550:7676, 550(7676), 345–353. Retrieved 24 August 2022 from <https://doi.org/10.1038/NATURE24286>
- Sheppard, B., Rappoport, N., Loh, P. R., Sanders, S. J., Zaitlen, N., & Dahl, A. (2021). A model and test for coordinated polygenic epistasis in complex traits. *Proceedings of the National Academy of Sciences of the United States of America*, 118(15). Retrieved 18 February 2022 from <https://doi.org/10.1073/pnas.1922305118>
- Siemiatycki, J., & Thomas, D. C. (1981). Biological models and statistical interactions: an example from multistage carcinogenesis. *International Journal of Epidemiology*, 10(4), 383–387. Retrieved 19 February 2022 from <https://doi.org/10.1093/IJE/10.4.383>
- Slatkin, M. (2008, June). Linkage disequilibrium - Understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*. Nature Publishing Group. Retrieved 30 April 2021 from <https://doi.org/10.1038/nrg2361>
- Smith, R. P., Lam, E. T., Markova, S., Yee, S. W., & Ahituv, N. (2012). Pharmacogene regulatory elements: From discovery to applications. *Genome Medicine*, 4(5), 1–13. Retrieved 18 February 2022 from <https://doi.org/10.1186/gm344>
- Solimena, M., Schulte, A. M., Marselli, L., Ehehalt, F., Richter, D., Kleeberg, M., ... Marchetti, P. (2018). Systems biology of the IMIDIA biobank from organ donors and pancreatectomised patients defines a novel transcriptomic signature of islets from individuals with type 2 diabetes. *Diabetologia*, 61(3), 641–657. Retrieved 27 February 2019 from <https://doi.org/10.1007/s00125->

017-4500-3

- Spracklen, C. N., Horikoshi, M., Kim, Y. J., Lin, K., Bragg, F., Moon, S., ... Sim, X. (2020). Identification of type 2 diabetes loci in 433,540 East Asian individuals. *Nature*, 582(7811), 240–245. Retrieved 21 January 2021 from <https://doi.org/10.1038/s41586-020-2263-3>
- Srinivasan, M., Choi, C. S., Ghoshal, P., Pliss, L., Pandya, J. D., Hill, D., ... Patel, M. S. (2010). β -Cell-specific pyruvate dehydrogenase deficiency impairs glucose-stimulated insulin secretion. *American Journal of Physiology - Endocrinology and Metabolism*, 299(6). Retrieved 11 August 2022 from <https://doi.org/10.1152/AJPENDO.00339.2010/ASSET/IMAGES/LARGE/ZH10121061020007.JPEG>
- Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., ... Lancet, D. (2016). The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Current Protocols in Bioinformatics*, 54(1), 1.30.1-1.30.33. Retrieved 18 February 2022 from <https://doi.org/10.1002/CPBI.5>
- Stranger, B. E., Montgomery, S. B., Dimas, A. S., Parts, L., Stegle, O., Ingle, C. E., ... Dermitzakis, E. T. (2012). Patterns of Cis Regulatory Variation in Diverse Human Populations. *PLOS Genetics*, 8(4), e1002639. Retrieved 18 February 2022 from <https://doi.org/10.1371/JOURNAL.PGEN.1002639>
- Sulaiman, N., Hachim, M. Y., Khalique, A., Mohammed, A. K., Al Heialy, S., & Taneera, J. (2022). EXOC6 (Exocyst Complex Component 6) Is Associated with the Risk of Type 2 Diabetes and Pancreatic β -Cell Dysfunction. *Biology 2022, Vol. 11, Page 388*, 11(3), 388. Retrieved 18 August 2022 from <https://doi.org/10.3390/BIOLOGY11030388>
- Swede, H., Stone, C. L., & Norwood, A. R. (2007). National population-based biobanks for genetic research. *Genetics in Medicine*, 9(3), 141–149. Retrieved 18 February 2022 from <https://doi.org/10.1097/gim.0b013e3180330039>
- Szymczak, S., Biernacka, J. M., Cordell, H. J., González-Recio, O., König, I. R., Zhang, H., & Sun, Y. V. (2009). Machine learning in genome-wide association studies. *Genetic Epidemiology*, 33 Suppl 1(SUPPL. 1). Retrieved 18 February 2022 from <https://doi.org/10.1002/GEPI.20473>
- Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., ... Abecasis, G. R. (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature 2021 590:7845*, 590(7845), 290–299. Retrieved 16 July 2021 from <https://doi.org/10.1038/s41586-021-03205-y>
- Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., & Meyre, D. (2019, August 1). Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*. Nature Publishing Group. Retrieved 28 July 2020 from <https://doi.org/10.1038/s41576-019-0127-1>
- Taylor, H. J., Hung, Y.-H., Narisu, N., Erdos, M. R., Kanke, M., Yan, T., ... Taylor, D. L. (2022). Human pancreatic islet microRNAs implicated in diabetes and related traits by large-scale genetic analysis. *BioRxiv*, 2022.04.21.489048. Retrieved 18 August 2022 from <https://doi.org/10.1101/2022.04.21.489048>
- The 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature*, 526(7571), 68–74. Retrieved 27 February 2019 from <https://doi.org/10.1038/nature15393>
- The Cost of Sequencing a Human Genome. (2021). Retrieved 18 February 2022, from <https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost>
- The DIAGRAM Consortium, The AGEN-T2D Consortium, The SAT2D Consortium, The MAT2D Consortium, & The T2D-GENES Consortium. (2014). Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nature Genetics*, 46(3), 234. Retrieved 18 February 2022 from <https://doi.org/10.1038/NG.2897>
- The ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414), 57–74. Retrieved 23 February 2021 from <https://doi.org/10.1038/nature11247>
- The Gene Ontology Consortium. (2000, May). Gene ontology: Tool for the unification of biology. *Nature Genetics*. Nat Genet. Retrieved 21 January 2021 from <https://doi.org/10.1038/75556>
- The Gene Ontology Consortium. (2021). The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Research*, 49(D1), D325–D334. Retrieved 21 January 2021 from <https://doi.org/10.1093/nar/gkaa1113>
- The Genome of the Netherlands Consortium. (2014). Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nature Genetics*, 46(8), 818–825. Retrieved 18 February 2022 from <https://doi.org/10.1038/ng.3021>
- The GTEx Consortium. (2017). Genetic effects on gene expression across human tissues. *Nature*,

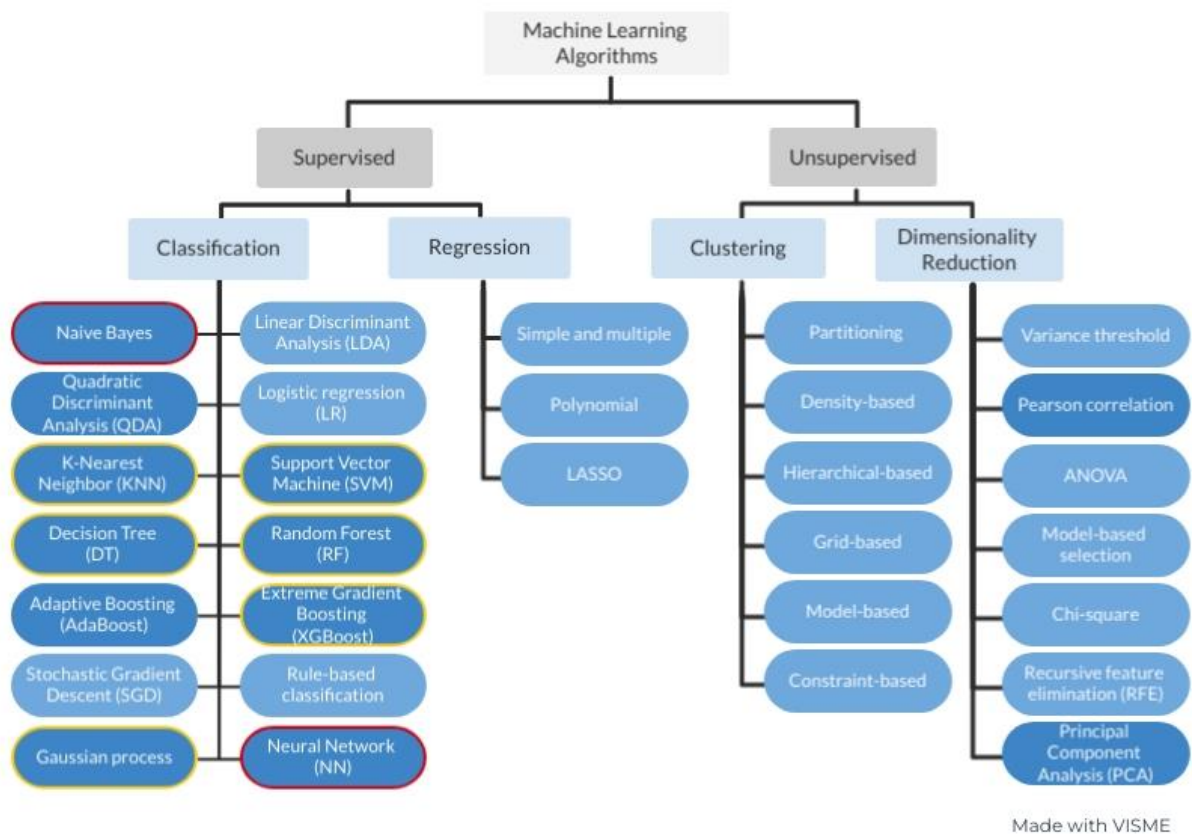
- 550(7675), 204–213. Retrieved 18 February 2022 from <https://doi.org/10.1038/nature24277>
- The GTEx Consortium. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*, 369(6509), 1318–1330. Retrieved 27 January 2021 from <https://doi.org/10.1126/SCIENCE.AAZ1776>
- The Haplotype Reference Consortium. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics*, 48(10), 1279–1283. Retrieved 21 January 2021 from <https://doi.org/10.1038/ng.3643>
- The International HapMap Consortium. (2003). The international HapMap project. *Nature*, 426(6968), 789–796. Retrieved 21 January 2021 from <https://doi.org/10.1038/nature02168>
- The International HapMap Consortium. (2005). A haplotype map of the human genome. *Nature*, 437(7063), 1299. Retrieved 18 February 2022 from <https://doi.org/10.1038/NATURE04226>
- The International HapMap Consortium. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164), 851–861. Retrieved 18 February 2022 from <https://doi.org/10.1038/nature06258>
- The Roadmap Epigenomics Consortium. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539), 317–329. Retrieved from <https://doi.org/10.1038/nature14248>
- The Telomere-to-Telomere Consortium. (2022). The complete sequence of a human genome. *Science*, 376(6588), 44–53. Retrieved 1 July 2022 from <https://doi.org/10.1126/science.abj6987>
- The UK10K Consortium. (2015). The UK10K project identifies rare variants in health and disease. *Nature*, 526(7571), 82–90. Retrieved 27 February 2019 from <https://doi.org/10.1038/nature14962>
- Thomas, D. (2010). Gene-Environment-Wide Association Studies: Emerging Approaches. *Nature Reviews Genetics*, 11(4), 259. Retrieved 16 July 2021 from <https://doi.org/10.1038/NGR2764>
- Turner, M., van de Bunt, M., Torres, J. M., Mahajan, A., Nylander, V., Bennett, A. J., ... McCarthy, M. I. (2018). Integration of human pancreatic islet genomic data refines regulatory mechanisms at Type 2 Diabetes susceptibility loci. *ELife*, 7. Retrieved 27 February 2019 from <https://doi.org/10.7554/eLife.31977>
- Timpson, N. J., Greenwood, C. M. T., Soranzo, N., Lawson, D. J., & Richards, J. B. (2018, February 1). Genetic architecture: The shape of the genetic contribution to human traits and disease. *Nature Reviews Genetics*. Nature Publishing Group. Retrieved 10 March 2021 from <https://doi.org/10.1038/nrg.2017.101>
- Tse-Wen Chang. (1983). Binding of cells to matrixes of distinct antibodies coated on solid surface. *Journal of Immunological Methods*, 65(1–2), 217–223. Retrieved from [https://doi.org/10.1016/0022-1759\(83\)90318-6](https://doi.org/10.1016/0022-1759(83)90318-6)
- Udler, M. S., Kim, J., Grotthuss, M. von, Bonàs-Guarch, S., Cole, J. B., Chiou, J., ... Florez, J. C. (2018). Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: A soft clustering analysis. *PLoS Medicine*, 15(9). Retrieved 21 September 2021 from <https://doi.org/10.1371/JOURNAL.PMED.1002654>
- Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., ... Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers* 2021 1:1, 1(1), 1–21. Retrieved 9 September 2021 from <https://doi.org/10.1038/s43586-021-00056-9>
- Umans, B. D., Battle, A., & Gilad, Y. (2021). Where Are the Disease-Associated eQTLs? *Trends in Genetics*, 37(2), 109–124. Retrieved 18 February 2022 from <https://doi.org/10.1016/J.TIG.2020.08.009>
- van de Bunt, M., Manning Fox, J. E., Dai, X., Barrett, A., Grey, C., Li, L., ... Gloyn, A. L. (2015). Transcript Expression Data from Human Islets Links Regulatory Signals from Genome-Wide Association Studies for Type 2 Diabetes and Glycemic Traits to Their Downstream Effectors. *PLOS Genetics*, 11(12), e1005694. Retrieved 27 February 2019 from <https://doi.org/10.1371/journal.pgen.1005694>
- Van De Geijn, B., Mcvicker, G., Gilad, Y., & Pritchard, J. K. (2015). WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nature Methods*, 12(11), 1061–1063. Retrieved 18 February 2022 from <https://doi.org/10.1038/nmeth.3582>
- Venter, J. C., Adams, M. D., Sutton, G. G., Kerlavage, A. R., Smith, H. O., & Hunkapiller, M. (1998). Shotgun sequencing of the human genome. *Science*, 280(5369), 1540–1542. Retrieved 18 February 2022 from <https://doi.org/10.1126/science.280.5369.1540>
- Verma, S. S., Lucas, A., Zhang, X., Veturi, Y., Dudek, S., Li, B., ... Ritchie, M. D. (2018). Collective feature selection to identify crucial epistatic variants. *BioData Mining*, 11(1). Retrieved 18 February 2022 from <https://doi.org/10.1186/S13040-018-0168-6>
- Vickaryous, M. K., & Hall, B. K. (2006). Human cell type diversity, evolution, development, and classification with special reference to cells derived from the neural crest. *Biological Reviews of*

- the Cambridge Philosophical Society*, 81(3), 425–455. Retrieved 18 February 2022 from <https://doi.org/10.1017/S1464793106007068>
- Vignal, A., Milan, D., SanCristobal, M., & Eggen, A. (2002). A review on SNP and other types of molecular markers and their use in animal genetics. *Genetics, Selection, Evolution: GSE*, 34(3). Retrieved 18 February 2022 from <https://doi.org/10.1186/1297-9686-34-3-275>
- Viñuela, A., Varshney, A., van de Bunt, M., Prasad, R. B., Asplund, O., Bennett, A., ... McCarthy, M. I. (2020). Genetic variant effects on gene expression in human pancreatic islets and their implications for T2D. *Nature Communications*, 11(1), 1–14. Retrieved 21 January 2021 from <https://doi.org/10.1038/s41467-020-18581-8>
- Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J. (2017, July 6). 10 Years of GWAS Discovery: Biology, Function, and Translation. *American Journal of Human Genetics*. Cell Press. Retrieved 10 March 2021 from <https://doi.org/10.1016/j.ajhg.2017.06.005>
- Vujkovic, M., Keaton, J. M., Lynch, J. A., Miller, D. R., Zhou, J., Tcheandjieu, C., ... Saleheen, D. (2020). Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis. *Nature Genetics*, 52(7), 680–691. Retrieved 21 January 2021 from <https://doi.org/10.1038/s41588-020-0637-y>
- Wainschtein, P., Jain, D., Zheng, Z., Cupples, L. A., Shadyab, A. H., McKnight, B., ... Visscher, P. M. (2022). Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. *Nature Genetics* 2022 54:3, 54(3), 263–273. Retrieved 1 July 2022 from <https://doi.org/10.1038/s41588-021-00997-7>
- Wang, D. G., Fan, J. B., Siao, C. J., Berno, A., Young, P., Sapolsky, R., ... Lander, E. S. (1998). Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*, 280(5366), 1077–1082. Retrieved 18 February 2022 from <https://doi.org/10.1126/SCIENCE.280.5366.1077>
- Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1), 57–63. Retrieved 18 February 2022 from <https://doi.org/10.1038/NRG2484>
- Watanabe, K., Stringer, S., Frei, O., Umičević Mirkov, M., de Leeuw, C., Polderman, T. J. C., ... Posthuma, D. (2019). A global overview of pleiotropy and genetic architecture in complex traits. *Nature Genetics*, 51(9), 1339–1348. Retrieved 30 April 2021 from <https://doi.org/10.1038/s41588-019-0481-0>
- Watanabe, R. M., Valle, T., Hauser, E. R., Ghosh, S., Eriksson, J., Kohtamäki, K., ... Boehnke, M. (1999). Familiality of quantitative metabolic traits in Finnish families with non-insulin-dependent diabetes mellitus. Finland-United States Investigation of NIDDM Genetics (FUSION) Study investigators. *Human Heredity*, 49(3), 159–168. Retrieved 18 February 2022 from <https://doi.org/10.1159/000022865>
- Watson, J. D., & Crick, F. H. C. (1953). Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, 171(4356), 737–738. Retrieved 30 April 2021 from <https://doi.org/10.1038/171737a0>
- Wei, W. H., Hemani, G., & Haley, C. S. (2014, November 25). Detecting epistasis in human complex traits. *Nature Reviews Genetics*. Nature Publishing Group. Retrieved 11 March 2021 from <https://doi.org/10.1038/nrg3747>
- Willemsen, G., Ward, K. J., Bell, C. G., Christensen, K., Bowden, J., Dalgård, C., ... Spector, T. (2015). The Concordance and Heritability of Type 2 Diabetes in 34,166 Twin Pairs From International Twin Registers: The Discordant Twin (DISCOTWIN) Consortium. *Twin Research and Human Genetics: The Official Journal of the International Society for Twin Studies*, 18(6), 762–771. Retrieved 18 February 2022 from <https://doi.org/10.1017/THG.2015.83>
- Wray, N. R., Yang, J., Hayes, B. J., Price, A. L., Goddard, M. E., & Visscher, P. M. (2013, July). Pitfalls of predicting complex traits from SNPs. *Nature Reviews Genetics*. NIH Public Access. Retrieved 11 March 2021 from <https://doi.org/10.1038/nrg3457>
- Zheng, Y., Chen, Z., Pearson, T., Zhao, J., Hu, H., & Prospero, M. (2020, April 1). Design and methodology challenges of environment-wide association studies: A systematic review. *Environmental Research*. Academic Press Inc. Retrieved 31 May 2021 from <https://doi.org/10.1016/j.envres.2020.109275>
- Zhu, S., Hou, S., Lu, Y., Sheng, W., Cui, Z., Dong, T., ... Wan, Q. (2021). USP36-Mediated Deubiquitination of DOCK4 Contributes to the Diabetic Renal Tubular Epithelial Cell Injury via Wnt/ β -Catenin Signaling Pathway. *Frontiers in Cell and Developmental Biology*, 9. Retrieved 2 March 2022 from <https://doi.org/10.3389/FCCELL.2021.638477>
- Zughaier, S. M., Stauffer, B. B., & McCarty, N. A. (2014). Inflammation and ER Stress Downregulate

BDH2 Expression and Dysregulate Intracellular Iron in Macrophages. *Journal of Immunology Research*, 2014. Retrieved 12 August 2022 from <https://doi.org/10.1155/2014/140728>

SUPPLEMENTAL MATERIAL

11. Supplemental Material



Supplemental Figure 1. Machine Learning algorithms based on the type of problem to be solved. Each machine learning (ML) algorithm is specialised in a different type of analysis. Thus, the selection of a ML method, although challenging, represents one of the most important steps in a study. For this reason, in this figure are represented the most common ML models based on the learning type and the specific type of problem that can be usually solved with them. Consequently, this list presents the most suitable and broadly used ML approaches in Biomedicine for classification, regression, clustering, or dimensionality reduction. The models represented in dark blue in the diagram have been used in this thesis and will be explained in this section. The models with a yellow border can be used for classification and regression. The models with a red border can be used for classification and clustering.

Supplemental Table 1. Population haplotype reference panels.

PROJECT	YEAR	DATA TYPE	N.INDIVIDUALS	POPULATIONS (ANCESTRY)	N.SNPs	OTHER VARIANTS
HapMap	2007	Genotyping array	270	4 populations (Africa, Asia and Europe)	3.1 million	
GoNI	2014	WGS	769 (250 parent-offspring)	Netherland (Europe - Dutch)	20.4 million	1.2 million insertions and deletions
1000G	2015	Sequencing and genotyping array data	2,504	26 populations (Africa, East Asia, Europe, South Asia and the Americas)	84.7 million	3.6 million Indels and 60,000 SVs
UK10K	2015	WGS and WES	~10,000 (3,781 healthy and 6,000 with rare disease, severe obesity, and neurodevelopmental disorders)	United Kingdom (Europe - British)	42 million	~3.5 million Indels and 18,739 large deletions
HRC	2016	WGS	64,976	20 studies (mostly Europe, but also Africa, East Asia, South Asia and the Americas)	39,235,157	
TopMed	2021	WGS	53,831 (130,000 individuals projected)	> 80 studies	381,343,078	28,980,753 Indels

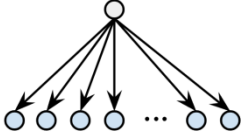
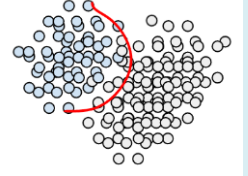
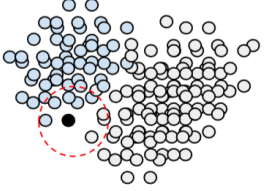
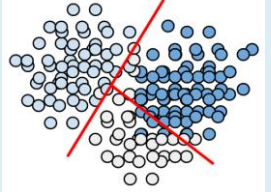
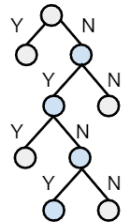
Supplemental Table 2. Data types.

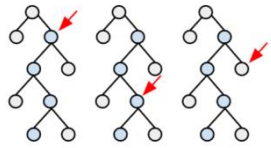
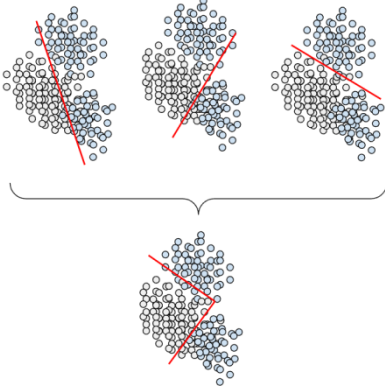
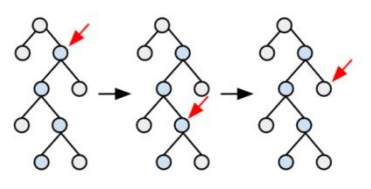
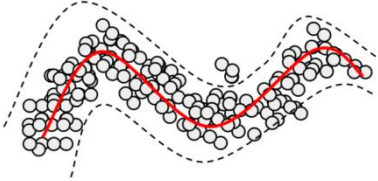
DATA TYPE	DESCRIPTION	EXAMPLES
Structured	Well formatted, ordered, organised and easily accessible data	SQL databases, csv files
Unstructured	No formatted data which usually complicates the analysis	Text, multimedia
Semi-structured	Data presenting some organisation facilitating the analysis	Non-SQL databases, HTML, JSON, XML
Metadata	Data describing the input dataset which can include some related relevant information. It can be used to improve the performance of the ML method	Information related to the origin of the data

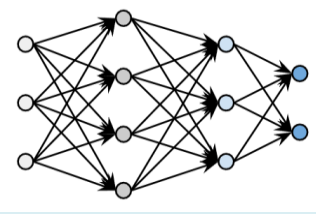
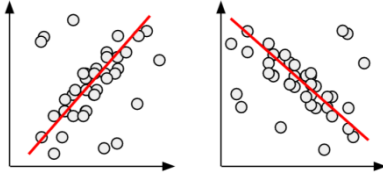
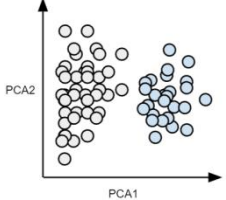
Supplemental Table 3. Machine Learning models based on the learning type.

LEARNING TYPE	DESCRIPTION	EXAMPLES	ALGORITHM
Supervised	Labelled	Groups of patients that can be classified in cases and controls	Decision Tree, Naïve Bayes, Support Vector Machine
Unsupervised	Unlabelled	Groups of patients from which a classification will be obtained	K-means clustering, Principal Component Analysis
Semi-supervised	A combination of supervised and unsupervised learning	Find groups of patients and then find a way of classifying them in these groups	A combination of supervised and unsupervised learners
Reinforced	Learning only based on obtaining a better outcome	Automation or optimization problems	
Multitask	Help other learners with simultaneous multiple tasks outcomes	-	
Ensemble	Combination of learners in a unique learner	-	Boosting, Bagging
Instance-based	An already learned pattern which will be applied only to test new data by comparing it with the already known training instances	-	K-Nearest Neighbour

Supplemental Table 4. Supervised Machine Learning classifiers and dimensionality reduction techniques applied in this thesis.

MODEL	PROBLEM	BASIS	DESCRIPTION	ADVANTAGES	DISADVANTAGES	SCHEMA
Naive Bayes	Classification and clustering	Conditional probability	Creates trees based on their probability of happening	Binary and multi-class classification Small amount of training data Easy to interpret	Strong assumptions of features independence	
Quadratic Discriminant Analysis (QDA)	Classification	Conditional densities and Bayes rule	Creates a decision boundary based on features quadratic combinations	Easily computed No hyperparameter tuning		
K-Nearest Neighbour (KNN)	Classification and regression	Similarity measures	Creates a decision based on the majority vote to a nearest neighbour	Robust to noise	Needs to be adjusted for the optimal number of neighbours to be considered	
Support Vector Machine (SVM)	Classification, regression and other tasks	Principle of margin calculation	Creates hyperplanes to separate classes maximising the distance between the margin and the classes	Effective in high-dimensional spaces	Problems with noisy data and overlapping target classes	
Decision Tree (DT)	Classification and regression	Sorting by value	Creates trees where each node represents an attribute of a group and each branch represents the value that the node can take	Easy to interpret Accepts numerical and categorical features	Tends to overfit Noisy Weak classifier	

MODEL	PROBLEM	BASIS	DESCRIPTION	ADVANTAGES	DISADVANTAGES	SCHEMA
Random Forest (RF)	Ensemble classification and regression	Sorting by value	Creates multiple decision trees and uses the majority voting or averages to obtain the result	Accepts numerical and categorical features Minimises the overfitting Increases the prediction accuracy	Reduces the interpretability	
Adaptive Boosting (AdaBoost)	Ensemble classification	Iteration	Creates a classifier based on the combination of many poor classifiers. It improves by learning from their errors	Improves the efficiency of the classifier	Can trigger overfits Sensitive to noisy data and outliers	
Extreme Gradient Boosting (XGBoost)	Ensemble classification and regression	Sorting by value	Creates multiple decision trees, minimising the loss function and performing regularisation	Accepts numerical and categorical features Minimises the overfitting Increases the prediction accuracy Scalable Fast Handles sparse data Handles missing data	Can struggle to learn in cases where a lot of noise is present	
Gaussian Process	Classification and regression	Probability distribution	Creates a probability distribution over functions	The prediction interpolates the observations Gives an estimate of its uncertainty Can be adjusted for different kernels	Computationally expensive Not sparse Lose efficiency in high dimensional spaces (more than few dozens of features)	

MODEL	PROBLEM	BASIS	DESCRIPTION	ADVANTAGES	DISADVANTAGES	SCHEMA
Neural Networks (NN)	Classification and clustering	Linear regression	Creates different node layers with all the nodes connected to another. If the output of any node is above a threshold value it gets activated and sends data to the next layer	Flexible Accepts variable input size Accepts non-linear data Handles missing data	Long training times High computing memory requirements The output contains uncertainty	
Pearson correlation	Dimensionality reduction	Linear correlation	Finds correlation between features to find variables with no linear correlation	Reduces overfitting	Problems with missing data	
Principal Component Analysis (PCA)	Dimensionality reduction	Covariance matrix eigenvalues	Identifies the highest eigenvalues of a covariance matrix to project in a subspace of equal or fewer dimensions	Reduces overfitting	Problems with missing data Independent variables become less interpretable Information loss	

Supplemental Table 5. ML binary classifiers effectiveness and reliability measures.

MEASURE	FORMULA	DESCRIPTION	IMBALANCE BEHAVIOUR
Precision	$\frac{TP}{TP + FP}$	Measures the goodness of the classification among the predicted diseased individuals. This magnitude explains the proportion of truly predicted diseased individuals among the whole group of diseased predictions made. Therefore, it is related to the statistical type error I but presenting a clear dependency on the prior distribution of the data.	Not recommended
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$	Measures the goodness of the predictions among the total group of individuals. This quantity is explained by the ratio of true predictions among the total number of predictions made.	Not recommended
Recall	$\frac{TP}{TP + FN}$	Measures the goodness of the classification among the diseased individuals group. It is also named true positive rate (TPR) or sensitivity. This parameter is calculated by assessing the proportion of truly diseased classified patients among the whole group of diseased individuals. It is the complement of the type II error rate (1 - type II error rate).	Not recommended
Specificity	$\frac{TN}{TN + FP}$	Measures the goodness of the classification among the non-diseased individuals group. It is also known as false positive rate (FPR). It is estimated by measuring the proportion of the predicted individuals that truly do not have the disease over the group of non-diseased individuals.	Not recommended
F1-score	$\frac{2 \times Precision \times Recall}{Precision + Recall}$	Harmonic mean of precision and recall. It is also known as balanced accuracy.	Balanced measure but not recommended
Matthews correlation coefficient	$\frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$	Balanced measure with its results ranging between -1 and 1. Consequently, a perfect prediction is obtained when the coefficient is 1, and a completely bad prediction in case of a -1 coefficient.	Balanced measure
ROC curve and AUROC	It is calculated based on recall (y axis) and fallout fallout=1-specificity (x axis).	Represents the sensitivity versus specificity. The evaluation of the ROC curve is done by computing the Area Under the ROC curve (AUROC). This curve tends to maximise the correctly classified positive values. It is useful to compare the results that can be obtained from different models and to discard suboptimal models.	Not recommended
Precision-Recall curve	It is calculated based on precision (y axis) and recall (x axis).	Is a measure of the success of the prediction. The evaluation of this curve is commonly done by calculating the area under the curve. Thus, a high area represents both high recall and high precision, therefore, accurate results.	Particularly useful in the presence of data imbalance

Supplemental Table 6. Some popular publicly available databases with functional information.

TYPE	CONTENT	EXAMPLES
Genes and isoforms categorization and description	Annotated genomics, transcript, protein sequence records, protein-coding genes, pseudogenes, long non-coding RNAs (lncRNAs), and small non-coding RNAs (sncRNAs)	RefSeq (O'Leary et al., 2016), GENCODE (Frankish et al., 2019; Harrow et al., 2012)
Gene and gene products functional descriptions	Gene functional information at different levels: biological process, molecular function and cellular component	GeneOntology (The Gene Ontology Consortium, 2000, 2021)
Protein and macromolecular complexes roles	Human pathways and processes including signal transduction, transport, DNA replication, metabolism and other cellular processes	Reactome Pathway database (Croft et al., 2011; Jassal et al., 2020)
TF with annotated elements and binding interfaces	Curated DNA binding sites and annotations of binding interfaces with their corresponding TFs transcription binding	FootprintDB (Sebastian, Contreras-Moreira, Araid, Agustín, & Zaragoza, 2014)
TF regulatory elements and regulatory interactions	<i>cis</i> - and <i>trans</i> - regulatory elements, and TF-target interactions	TRED (Jiang et al., 2007), TRRUST (Han et al., 2015)
Global and tissue-specific gene expression regulators	Genomic variants association with gene and transcript expression	GTEx (The GTEx Consortium, 2020), Gene Expression Atlas (Papatheodorou et al., 2020)
Epigenomic features profiles	DNA methylation, histone modifications, chromatin accessibility and small RNA transcripts	Epigenomic Roadmap Project (The Roadmap Epigenomics Consortium, 2015)

Supplemental Table 7. Main organs dysfunction that can derive T2D.

ORGAN	FUNCTION	PROBLEM	CAUSES	CONSEQUENCE	OTHER ORGANS
Adipose tissue	Use insulin to do triglyceride synthesis and induce the uptake of free fatty acid (FFA)	IR		Impaired glucose uptake, causing elevated glucose levels in plasma (glucotoxicity), consequent impaired insulin secretion (IIS), and promoting an enhanced FFA release (lipotoxicity)	The elevated levels of FFA induce hepatic and muscle IR
Pancreatic beta-cells	Secrete insulin	IS	IR, lipotoxicity, and glucotoxicity increase the demand on beta-cell IS, thus fasting the beta-cell failure progress and apoptosis	Beta-cells deterioration or death can lead to insulin secretory dysfunctions resulting in elevated glucose levels in blood (hyperglycemia)	
Skeletal muscle	One of the major receptors of glucose in the glucose uptake process	IR	Obesity and low levels of physical activity contribute to muscle IR	Bad insulin signalling and IR can lead to hyperglycemia	Progressive beta-cell failure
Liver	Main organ in the glucose production process under insulin regulation	IR		Overproduction of glucose	Progressive beta-cell failure
Gut	After glucose ingestion, it releases hormones that stimulate IS, promoting satiety, slowing gastric emptying, and inhibiting glucagon secretion				
Pancreatic alpha-cells	The major source of glucagon in response to low levels of glucose in blood (hypoglycemia)	Impaired glucagon secretion		Hyperglycemia	Progressive beta-cell failure
Kidneys	Small producers of glucose and filters of glucose to the urine in case of excess			Hyperglycemia	Progressive beta-cell failure
Brain	Main organ involved in food intake, appetite regulation, and a major responsible for glucose utilisation	IR		Suppress the inhibition of appetite and reduce satiety, promoting an imbalanced feeding and usually inducing obesity	Progressive beta-cell failure

POLYMORPHIC INVERSIONS

PUBLICATION

Polymorphic Inversions Underlie the Shared Genetic Susceptibility of Obesity-Related Diseases

Juan R. González,^{1,2,3,4,*} Carlos Ruiz-Arenas,^{1,2} Alejandro Cáceres,^{1,2,3} Ignasi Morán,⁵ Marcos López-Sánchez,^{6,7} Lorena Alonso,⁵ Ignacio Tolosana,¹ Marta Guindo-Martínez,⁵ Josep M. Mercader,^{5,8,9,10} Tonu Esko,^{11,12} David Torrents,^{5,13} Josefa González,¹⁴ and Luis A. Pérez-Jurado^{2,6,7,15}

The burden of several common diseases including obesity, diabetes, hypertension, asthma, and depression is increasing in most world populations. However, the mechanisms underlying the numerous epidemiological and genetic correlations among these disorders remain largely unknown. We investigated whether common polymorphic inversions underlie the shared genetic influence of these disorders. We performed an inversion association analysis including 21 inversions and 25 obesity-related traits on a total of 408,898 Europeans and validated the results in 67,299 independent individuals. Seven inversions were associated with multiple diseases while inversions at 8p23.1, 16p11.2, and 11q13.2 were strongly associated with the co-occurrence of obesity with other common diseases. Transcriptome analysis across numerous tissues revealed strong candidate genes for obesity-related traits. Analyses in human pancreatic islets indicated the potential mechanism of inversions in the susceptibility of diabetes by disrupting the *cis*-regulatory effect of SNPs from their target genes. Our data underscore the role of inversions as major genetic contributors to the joint susceptibility to common complex diseases.

Introduction

Obesity is a disorder with increasing but non-uniform prevalence in the world population and one of the major public health burdens.¹ Obesity (MIM: 615812)-derived morbidity and years of life lost strongly associate to a broad range of highly prevalent diseases, including type 2 diabetes (MIM: 125853), cardiovascular disease (MIM: 608901), asthma (MIM: 600807), and (neuro)psychological disturbance such as depression (MIM: 608516) or intellectual disability, among others.² While the causes underlying the multiple co-occurrences of obesity are likely complex and diverse, common mechanisms underlying these comorbidities, which are potential targets for preventive or therapeutic intervention, are largely unknown.

One of the possible genetic mechanisms of comorbidity can be through rare copy number variants (CNVs), which are more prevalent in people with some severe forms of obesity^{3,4} and might confer at least part of the increased risk for obesity via developmental delay.⁵ Most of these findings have been described in pediatric obesity.^{6,7}

Genomic inversions, copy-neutral changes in the orientation of chromosomal segments with respect to the refer-

ence, are (also) excellent candidates for being important contributors to the genetic architecture of common diseases. Inversion polymorphisms can alter the function of the including and neighboring genes by multiple mechanisms, disrupting genes, separating their regulatory elements, affecting chromatin structure, and maintaining a strong linkage of functional variants within an interval that escape recombination. Therefore, by putatively affecting multiple genes in numerous ways, inversions are important sources of shared genomic variation underlying different human diseases and traits. Consequently, human inversions show genetic influences in multiple phenotypes. For instance, the common inversion at 8p23.1 has been independently linked to obesity,⁸ autism (MIM: 209850),⁹ neuroticism (MIM: 607834),¹⁰ and several risk behavior traits,¹¹ while inversion at 17q21.31 has been associated with Alzheimer (MIM: 607822)¹² and Parkinson (MIM: 168600)¹³ diseases, heart failure,¹⁴ and intracranial volume.¹⁵ We previously reported a ~40% of population attributable risk for the co-occurrence of asthma and obesity given by a common inversion polymorphism at 16p11.2.¹⁶ In addition, transcriptional effects have been documented in several tissues for inversions at 17q21.31^{13,17} and 16p11.2.¹⁶

¹Barcelona Institute for Global Health (ISGlobal), Barcelona 08003, Spain; ²MIM (Hospital del Mar Research Institute), Barcelona 08003, Spain; ³Centro de Investigación Biomédica en Red en Epidemiología y Salud Pública (CIBERESP), Barcelona 08003, Spain; ⁴Department of Mathematics, Universitat Autònoma de Barcelona (UAB), Barcelona 08193, Spain; ⁵Joint BSC-CRG-IRB Research Program in Computational Biology, Barcelona Supercomputing Center (BSC-CNS), Barcelona 08034, Spain; ⁶Department of Experimental and Health Sciences (CEXS), Universitat Pompeu Fabra, Barcelona 08003, Spain; ⁷Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), Barcelona 08003, Spain; ⁸Programs in Metabolism and Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA; ⁹Diabetes Unit and Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA 02114, USA; ¹⁰Department of Medicine, Harvard Medical School, Boston, MA 02115, USA; ¹¹Estonian Genome Center, University of Tartu, Tartu 51010, Estonia; ¹²Institute of Molecular and Cell Biology, University of Tartu, 51010 Tartu, Estonia; ¹³Institut de Recerca i Estudis Avançats (ICREA), Barcelona 08003, Spain; ¹⁴Institute of Evolutionary Biology (CSIC-Universitat Pompeu Fabra), Barcelona 08003, Spain; ¹⁵Women's and Children's Hospital, South Australian Health and Medical Research Institute & University of Adelaide, Adelaide, SA 5005, Australia

*Correspondence: juanr.gonzalez@isglobal.org

<https://doi.org/10.1016/j.ajhg.2020.04.017>

© 2020 American Society of Human Genetics.



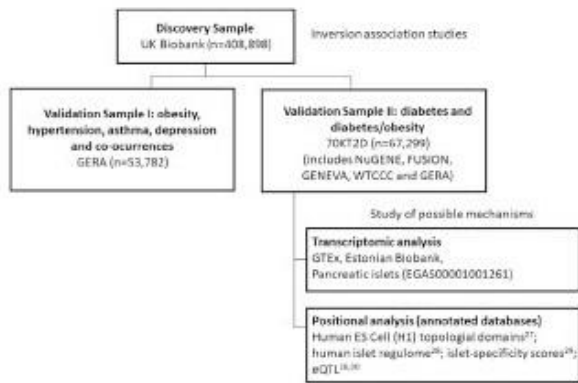


Figure 1. Discovery and Validation Datasets

The flow chart shows the discovery sample and the validation datasets as well as the datasets used for post-genomic data analyses. Sample size (n) used from each dataset after performing quality control are also shown.

It is estimated that each human genome contains about 156 inversions.¹⁸ Therefore, inversions constitute a substantial source of genetic variability. Many of those polymorphic inversions show signatures of positive or balancing selection associated with functional effects.¹⁹ However, the overall impact of polymorphic inversions on human health remains largely unknown because they are difficult to genotype in large cohorts. We overcame this limitation by recently reporting a subset of 20 inversions that can be genotyped with SNP array data as they are old in origin, low or not recurrent, and frequent in the population.²⁰ We have also included an additional inversion in our catalog, 16p11.2, previously validated and genotyped in diverse populations.¹⁶ Three of the inversions are submicroscopic (0.45–4 Mb), flanked by large segmental duplications, and contain multiple genes. Five are small (0.7–5 kb) and intragenic, and 13 are intergenic of variable size (0.7–90 kb) but highly enriched in pleiotropic genomic regions.²¹ While this is clearly not a comprehensive set of inversions, it is probably the largest set that can be genotyped in publicly available datasets.

In this manuscript, we aimed to study the association of 21 common polymorphic inversions in Europeans with highly prevalent co-morbid disorders and related traits. We particularly aimed to decipher the role of inversions in known epidemiological co-occurrences with obesity such as diabetes, hypertension (MIM: 145500), asthma, and mental diseases like depression, bipolar disorder (MIM: 125480), or neuroticism. For significant associations, we investigated whether causal pathways could be established and the most likely underlying mechanisms.

Material and Methods

Discovery Dataset

The UK Biobank (UKB) is a population-based cohort involving 500,000 individuals aged between 37 and 73 years, recruited

across UK in the period 2006–2010. Further details on the quality control and genotyping are described in the study design.²² Phenotypic information is recorded via questionnaires and interviews (e.g., demographics and health status) and SNP genotypes were generated from the Affymetrix Axiom UK Biobank and UKB-LEVE arrays. We based our study on 408,898 individuals from European descent and from whom inversion genotypes were called using SNP array data. Principal components computed by the UK Biobank (data-field 22009) were used in the analyses to control for population stratification.

Replication Datasets

Different public datasets with access grant to the co-authors were used to attempt to replicate our positive findings in the association studies (Figure 1). The next sections describe these resources.

Genetic Epidemiology Research on Aging (GERA)

The GERA cohort (dbGaP: phs000674.v1.p1) consists of more than 100,000 adults from the Northern California Region (USA). Only individuals with reported race (variable phv00196837.v2.p2) equal to white were selected for the analyses (n = 56,638). The resulting studied cohort is 40% male, 60% female, and ranges in age from 18 to more than 100 years old, with an average age of 64 years at the time of the survey (2007). Individuals were genotyped with Affymetrix Axiom_KP_UCSF_EUR. After quality control of the inversion genotyping calling process, a total of 53,782 individuals with information about sex, age, principal components for genetic ancestry, and several diseases including obesity (9,439 cases), diabetes (6,529 cases), hypertension (27,009 cases), asthma (8,716 cases), and depression (6,924 cases) were used in the replication studies.

70KforT2D: Diabetes and Obesity

The 70KforT2D study (70KT2D)²³ includes five datasets, two publicly available in EGA (NuGENE and GENEVA) and three available in dbGAP (FUSION, WTCCC, and GERA). Notice that 70KforT2D includes case subjects diagnosed with diabetes and obesity from the GERA cohort. We used information about being diabetic or not as described elsewhere.²³ The five datasets were used to attempt to replicate the significant findings in the UK Biobank data on diabetes. The WTCCC dataset was removed from the obesity and obesity/diabetes analysis since we did not have access to body mass index (BMI) information for that study. The GERA dataset was split in two (GERA1 and GERA2) to speed up the imputation and inversion calling procedure since it is a large dataset. After performing QC on inversion genotypes, a total of 67,299 individuals were used in the replication step (54,801 control subjects and 12,498 diabetic subject). Data were accessed from the portal cg.bsc.es/70kfort2d.

The obesity variable was created using the body mass index (BMI) variable. We considered control individuals those having BMI between 18.5 and 24.9 and obese people those having BMI > 30.0. For obesity associations, we excluded individuals with diabetes. As a result, a total of 34,316 individuals (23,818 control subjects and 10,498 obese subjects) were used for that purpose. The co-occurrence of obesity and diabetes was studied by comparing individuals with no obesity and no diabetes as the reference category with individuals being obese and diabetic simultaneously. This ended up with a total of 23,818 control and 5,715 obese/diabetic individuals. Next, we further describe the studies included in the 70KT2D dataset along with their accession numbers.

Northwestern NuGene Project: Type 2 Diabetes (NUGENE) (dbGaP: phs000237.v1.p1) contains data from individuals from

the Northwestern University Medical Center (USA). For this study, T2D case subjects were included if they had been diagnosed with type 2 diabetes, they took drugs to treat type 2 diabetes, or they presented abnormal diabetes-related blood measures. Control subjects were included if they had not been diagnosed with type 2 diabetes, they did not take drugs to treat type 2 diabetes, they presented normal diabetes-related blood measures, and they did not have any family history of diabetes (either type 1 or type 2). In both groups, subjects with type 1 diabetes were excluded. These individuals were genotyped with Illumina Human1M-Duov3_B.

The Finland-United States Investigation of NIDDM Genetics - G WAS Study (FUSION) (dbGaP: phs000100.v4.p1) aims to investigate the association between genetics and type 2 diabetes in Finnish families. For this study, case subjects were included if they had been diagnosed with type 2 diabetes, they took drugs to treat type 2 diabetes, or they presented abnormal diabetes-related blood measures. Control subjects were included if they presented normal diabetes-related blood measures and were frequency matched to the case subjects by age, sex, and birth province. In both groups, individuals with family history of type 1 diabetes were excluded. These individuals were genotyped with Illumina HumanHap300v1.1.

GENEVA Genes and Environment Initiatives in Type 2 Diabetes (Nurses' Health Study/Health Professionals Follow-up Study) (dbGaP: phs000091.v2.p1) is a nested case-control (2,720 case subjects and 3,180 control subjects) study from two USA female cohorts: the Nurses' Health Study (NHS) and the Health Professionals Follow-up Study (HPFS) with a mean age of 57 ranging from 40 to 78. These individuals were genotyped with Affymetrix AFFY_6.0.

Geographical Variation in Europe

POPRES project (dbGaP: phs000145.v4.p2, access granted to the authors) was used to estimate inversion frequencies in European countries and regions. This project aimed to facilitate exploratory genetic research by assembling a DNA resource from a large number of subjects participating in multiple studies throughout the world. We selected European individuals (variable phv00173964.v2.p2) leading a total of 3,071 samples. A geographic label (North, Center, South) was assigned to each individual using information of variable phv00066613.v2.p2.

Transcriptomic Analyses

GTEX Analysis

We associated the 21 chromosomal inversions to changes in gene expression in GTEX project. We determined inversion genotypes on the GTEX v7 genotype calls from dbGAP (dbGaP: phs000424.v7.p2, accession granted to the authors). We included only samples classified as European with a confidence higher than 90% by peddy.²⁴ Inv3_003 was discarded as the calling was not confident. Gene expression counts from RNA-seq data were downloaded using recount2.²⁵ We computed the association between gene expression and inversions using voom²⁶ and limma.²⁷ The linear model included the inversion coded as additive (0: NN, 1: NI, 2: II) and the same covariates than GTEX (first three genome-wide PCA components, sex, and covariates from PEER). In each tissue, we selected those features having more than 10 counts in at least 10% of the samples. We corrected the association results per tissue for multiple comparisons by using a false discovery rate (FDR) adjusted p value per tissue.

EGCUT Biobank

Estonian Gene Expression Cohort was used to attempt to replicate positive transcriptomic results found in GTEX. The cohort is

composed of 1,048 randomly selected samples (mean age 37 ± 16.6 years; 50% females) from the 53,000 samples in the Estonian Genome Center Biobank, University of Tartu. Whole-genome gene-expression levels from whole blood RNA were obtained by Illumina HT12v3 arrays according to manufacturer's protocols. Low-quality samples were excluded. All probes with primer polymorphisms were discarded, leaving 34,282 probes. Raw gene expression data were Log-Quantile normalized using MixupMapper software. DNA was genotyped with Human370CNV array.

Pancreatic Islets

We analyzed the transcriptomic effect of inversions 8p23.1 and 16q11.2 on 118 pancreatic human islet samples using RNA-sequencing counts and high-density genotyping data.²⁸ DNA genotype data (EGA: EGAS00001001261) was used to call inversion genotypes using *scoreInvHap*, then the association between gene expression and inversions was assessed using voom²⁶ and limma.²⁷ Only genes in the inversion regions were analyzed and un-corrected p values were reported as a measure of association.

Positional Analyses

For the positional analyses, several annotations were gathered from the following sources: TAD boundaries from the Human ES Cell (H1) topological domains;²⁹ promoters, enhancers, CTCF-peaks, and ATAC-seq open-chromatin regions from the human islet regulome annotation;³⁰ islet-specificity scores were calculated using the gene expression data from Miguel-Escalada et al.,³¹ and eQTL SNP-gene associations from van de Bunt et al.²⁸ and Fadista et al.³² The chromatin landscape coverage percentage was calculated using a sliding window of 500 kb and 1 Mb for inversions 8p23.1 and 16p11.2, respectively, using steps of 1% of the window size, and calculating the percentage of covered nucleotides by significant signal in each of the categories. For the islet-specific expression analysis, we calculated the non-islet median expression level and difference between the 75 and 50 quartiles, and we considered as islet specific any gene that was expressed in islet >3 quartiles over the median of non-islet expression. Visualization was done in python3 using the matplotlib graphics library.

Statistical Methods

SNP Imputation and Inversion Calling

SNP microarray data were imputed with *imputeInversion* pipeline prior to inversion calling (see Web Resources). This pipeline was designed to impute only those SNPs inside the inversion region or closer than 500 kb to the inversion breakpoints. This step is recommended before performing inversion calling. *imputeInversion* uses shapeITv2.r904 to phase,³³ Minimac3³⁴ to impute, and 1000 Genomes as reference haplotypes. Variants with an imputation R2 < 0.3 were discarded. Genotype probabilities were used to call inversions using *scoreInvHap*²⁰ which is available at Bioconductor. *scoreInvHap* computes a similarity score between an individual's alleles and the reference alleles in each chromosomal status. We used the development version of *scoreInvHap*, which includes references for 21 inversions. These methods were used to perform inversion calling in the discovery and replication studies as well as in individuals from POPRES.

Inversion Frequencies

Inversion frequencies were estimated in UKB and POPRES studies using SNPassoc package.³⁵ A trend test implemented in the R function *prop.trend.test* was used to assess whether inversion frequencies in European regions from POPRES (North, Center, South) showed a significant cline. Principal component analysis was used

to visualize inversion frequencies across European regions of POPRES dataset.

Obesity and Obesity Co-occurrence Traits

Obesity trait was created using body mass index (BMI) information. First, BMI was categorized in five categories using World Health Organization (WHO) classification which considers the following categories: underweight (BMI below 18.5), normal weight (BMI between 18.5 and 25), pre-obesity (BMI between 25 and 29.9), obesity class I (BMI between 30 and 34.9), and obesity classes II and III (BMI above 35). Obesity was considered as obesity classes I, II, and III and was compared with normal weight category. The analysis of obesity co-occurrence with diabetes, hypertension, asthma, depression, and neuroticism was performed by comparing individuals with normal weight and no presence of the disease with individuals being obese and having the disease of interest.

Inversion Association Analyses

Each inversion was independently associated with all the traits by using generalized linear models implemented in SNPAssoc package.³⁵ The models were adjusted for gender, age, and the first four principal components obtained from GWAS data in order to control for population genetic differences. The inversions were analyzed using an additive model. Multiple comparison problem was addressed by correcting for the total number of inversions and the phenotypes analyzed by considering the effective number of tests (18 independent tests) using Li and Ji method³⁶ that accounts for correlation among traits. This ended up with a corrected *p* value equal to 0.00128.

Causal Inference

Mediation analysis using *mediation* R package³⁷ was used to evaluate whether inversion 8p23.1 mediates the association between obesity and diabetes. Additive Bayesian network models using *abn* R package³⁸ were used to determine optimal Bayesian network models to identify statistical dependencies between inversions 8p23.1, 16p11.2, and 11q13.2 and obesity, diabetes, and hypertension in the UKB dataset, and validated in the GERA cohort. The most probable network structure was estimated using exact order-based approach as implemented in the *mostprobable* function of *abn* package.

Data Availability

The data used in this work were obtained from publicly available datasets that are accessible through public repositories: UKB study, dbGaP, EGA, GTeX, and GEO. The inversion calling of UKB samples will be available through their platform. The inversion calling for the other samples and the complete transcriptomic summary statistics of the 21 inversions are available in our GitHub repository (see [Web Resources](#)).

Results

Frequency and Stratification of Inversions in European Populations

Using *scoreInvHap*, we first called the inversion status of individuals from the UK Biobank (UKB) with European ancestry ($n = 408,898$). We confirmed the previously reported frequency in the 1000 Genomes project of the 21 inversions analyzed in this work (Table 1). As inversion frequencies have a strong demographic effect, we also analyzed 12 European countries from the POPRES study (Figure S1). We observed significant clines along north-

south latitude for several inversions (Table 1 and Figure S2A) as well as subtle ancestral differences (Figure S2B). Thus, population stratification was considered when performing association analyses as explained in the methods section.

Inversions at 8p23.1, 16p11.2 Robustly Associate with Obesity and Obesity-Related Traits

The discovery phase of the study used data from UKB. We performed association analyses between the 21 inversions with obesity and co-morbid diseases and traits (see [Material and Methods](#)). These include obesity, diabetes, stroke, hypertension, asthma, chronic obstructive pulmonary disease (COPD), depression, and bipolar disorder, along with related traits or phenotypes classified as morphometric (4 traits), metabolic (5 traits), lipidic (2 traits), respiratory (3 traits), and behavioral (3 traits) (Figure 2). Table S1 shows the total number of case and control subjects used to perform the association analyses on each trait. The significant associations were further validated in the GERA independent dataset that contains information about several diseases. Positive results found in diabetes were validated in the 70KT2D dataset, which includes GERA among others (NUGENE, FUSION, GENEVA, and WTCCC) (see [Material and Methods](#) and Figure 1 which describes the comprehensive data analysis performed in the different datasets).

The analyses on the UKB revealed several genetic influences of inversions on obesity and related common diseases (Figure 2). We observed a total of 74 significant associations after correcting for the number of inversions analyzed and the effective number of tests to consider the multiple analyzed traits (see [Material and Methods](#)). In general, we observed higher numbers of associations and stronger effects for the largest inversions at 8p23.1, 16p11.2, and 17q21.31, consistent with the fact that they encapsulate more genes. Some smaller inversions such as the ones at 11q13.2 and Xq13.2 also showed notable effects such as shared susceptibility and strength, respectively. We found a prominent inflation of association suggesting common genetic influences of the inversions across multiple phenotypes (Figure S3A). Some of the associations found have already been reported, such as those at inversion 8p23.1 with obesity⁸ and neuroticism¹⁰ and the one with inversion 16p11.2 with obesity.¹⁶

As a summary of the relevant findings, we observed that inversions at 8p23.1, 16p11.2, and 11q13.2 are all strongly associated with several obesity-related diseases (Figure 2). Remarkably, the non-inverted (N) allele of inversion 8p23.1 (i.e., the risk allele) is independently associated with diabetes (OR = 1.04, $p = 1.1 \times 10^{-3}$), hypertension (OR = 1.04, $p = 7.0 \times 10^{-16}$) and asthma (OR = 1.03, $p = 7.0 \times 10^{-5}$) (Table 2). The association with diabetes was replicated in the 70KT2D study (Figure 3A) (OR = 1.08, $p = 1.1 \times 10^{-8}$) as well as the association with obesity (OR = 1.08, $p = 5.6 \times 10^{-6}$) and the association with

Table 1. Characteristics of the 21 Genomic Inversions

Chr. Band	Coordinates	Num. SNPs	Length (Kb)	Inv. Freq. ²⁰	UKB	European Populations (POPRES)			Trend p Value
						North	Center	South	
1p22.1	chr1:92,131,841-92,132,615	6	0.77	11.23	10.1	8.9*	9.1*	14.4*	0.0057*
1q31.3	chr1:197,756,784-197,757,982	5	1.2	19.68	20.2	19.4	21.7	19.1	0.8781
2p22.3	chr2:33,764,554-33,765,272	6	0.72	15.11	15.5	13.8	13.5	11.7	0.3199
2q22.1	chr2:139,004,949-139,009,203	13	4.25	71.47	75.3	76.6*	71.9*	66.4*	0.0003*
3q26.1	chr3:162,545,362-162,547,641	6	2.28	56.16	51.1	53.4*	55.2*	61.1*	0.0140*
6p21.33	chr6:31,009,222-31,010,095	5	0.87	63.12	62	61.3*	65.0*	72.8*	0.0001*
6q23.1	chr6:130,848,198-130,852,318	12	4.12	6.56	7.6	7.3	8.7	8.1	0.6070
7p14.3	chr7:31,586,765-31,592,019	11	5.25	23.56	23.5	22.6	23.3	26.5	0.1605
7p11.2	chr7:54,302,450-54,376,389	180	73.9	50.39	51	52.1	51.2	54.4	0.4715
7q11.22	chr7:70,426,185-70,438,879	10	12.7	63.52	61.8	61.0	61.8	62.4	0.6196
7q36.1	chr7:151,010,030-151,012,107	5	2.08	19.88	20.7	20.1	24.0	24.7	0.0775
8p23.1	chr8:8,055,789-11,980,649	13,411	3,925	57.95	55.6	56.5	54.9	53.6	0.3424
11p12	chr11:41,162,296-41,167,044	7	4.75	15.81	15.4	14.3	13.9	14.6	0.8479
11q13.2	chr11:66,018,563-66,019,946	5	1.38	34.39	28.5	32.4	31.3	30.5	0.5287
12q13.11	chr12:47,290,470-47,309,756	43	19.3	7.46	6.6	6.2*	7.9*	10.9*	0.0085*
12q21.2	chr12:71,532,784-71,533,816	4	1.03	36.98	38.8	37.4	36.5	33.3	0.1647
14q23.3	chr14:65,842,304-65,843,165	4	0.86	29.42	25.5	26.5	26.9	26.4	0.9823
16p11.2	chr16:28,424,774-28,788,943	361	364.17	ND	40.5	39.3*	32.0*	29.1*	0.0007*
17q21.31	chr17:43,661,775-44,372,665	3,637	711	23.96	22.6	15.1*	19.4*	22.1*	0.0035*
21q21.3	chr21:28,020,653-28,021,711	11	1.06	51.29	49.2	51.6	52.1	57.4	0.0651
Xq13.2	chrX:72,215,927-72,306,774	135	90.8	13.3	13.9	12.4*	12.1*	8.5*	0.0400*

The table shows the coordinates, SNP content, size, and inversion frequency obtained from 1000 Genomes as described in Ruiz-Arenas et al.,²⁰ the UKB and European regions (north, center and south) using the regions described in the POPRES dataset (see [Material and Methods](#)). The p value corresponds to a trend test to assess north-south linear association (asterisk indicates those significant at 5% level).

hypertension, which was validated in the GERA study (OR = 1.03, $p = 0.0183$) (Table 2). We also found a significant association between the non-inverted (N) allele of inversion 16p11.2 and obesity (OR = 1.05, $p = 3.9 \times 10^{-24}$) that was replicated in the GERA study (OR = 1.07, $p = 1.4 \times 10^{-4}$). The significant association found in the UKB for the inversion 11q13.2 was not validated in the GERA study (OR = 1.03, $p = 0.0712$). Consistently, the analysis of UKB study also revealed association of inversions at 8p23.1 and 16p11.2 with different obesity-related traits such as body mass index (BMI), waist circumference, high density lipoprotein (HDL), or systolic and diastolic blood pressure, among others (Figure 2).

Some interesting associations in the discovery sample included those of inversion 17q21.31 with HDL, waist circumference, waist-hip ratio, and systolic and diastolic blood pressure (Figure 2). Interestingly, this inversion also showed a significant role in behavioral traits such as mood swing, depression, and bipolar disorder, which would need further validation. While we also found significant association of the inversion 6p21.33 with asthma

(OR = 1.02, $p = 0.0215$) and different respiratory capacity traits (FEV1, $p = 3.4 \times 10^{-9}$ and FVC, $p = 3.2 \times 10^{-9}$) the association with asthma was not replicated in the GERA study. The inversion 7q11.2 was associated with different morphometric traits (BMI, waist circumference, and waist-to-hip ratio) and will require further validation studies.

Inversions at 8p23.1, 16p11.2, and 11q13.2 Are More Strongly Associated with the Co-occurrence of Diseases than with Single Diseases

Remarkably, the N-allele of the inversion 8p23.1 was significantly associated with the co-occurrence of obesity with diabetes (OR = 1.08, $p = 3.1 \times 10^{-7}$), hypertension (OR = 1.07, $p = 1.7 \times 10^{-16}$), or asthma (OR = 1.08, $p = 3.0 \times 10^{-11}$). These results were validated in the GERA and 70KT2D (Table 2). For obesity/diabetes, we observed an OR = 1.17 ($p = 1.4 \times 10^{-13}$) (Table 2 and Figure 3C) and none of the SNPs located within the inverted region were significantly associated at a genome-wide level (minimum $p = 3.8 \times 10^{-5}$) (Figure S4A). Finally, we also found a

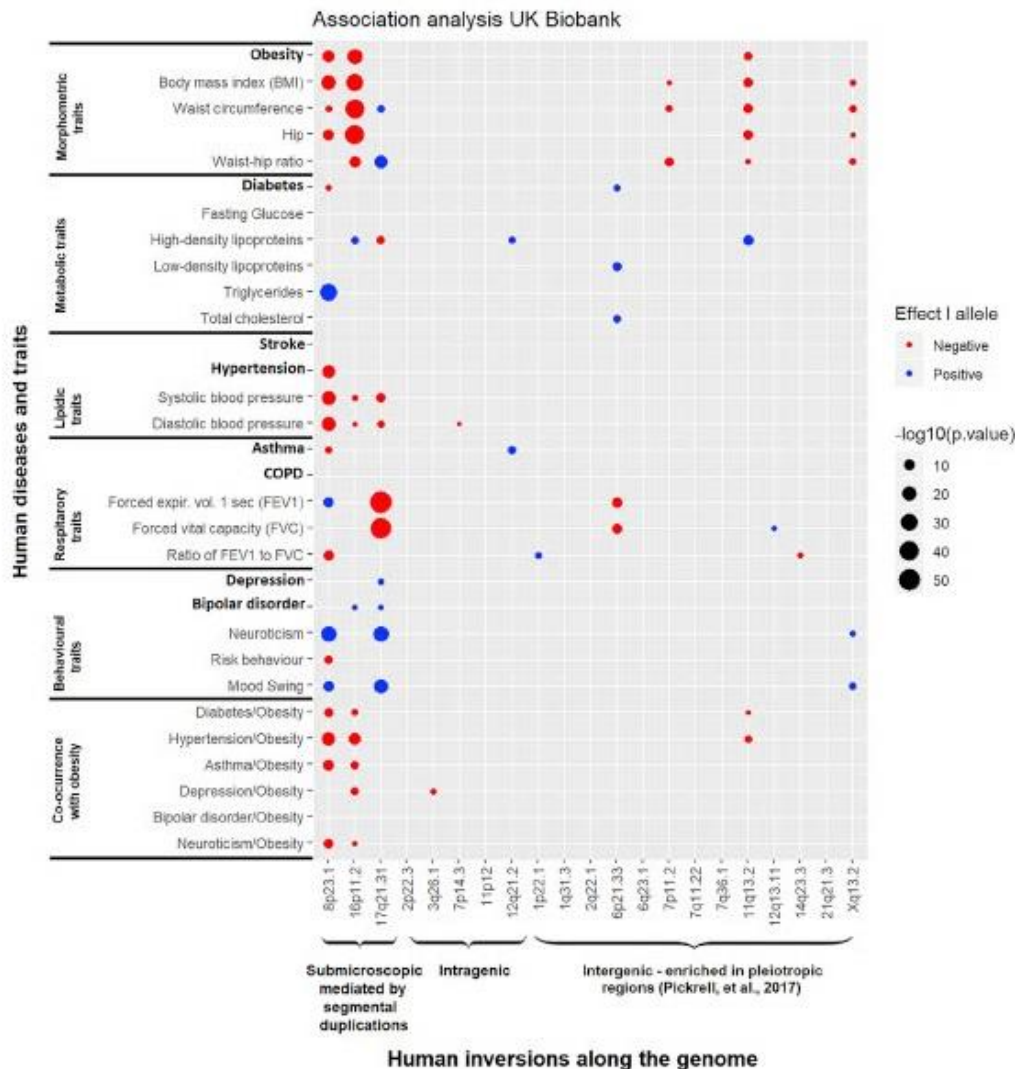


Figure 2. Association Analyses between 21 Inversions and 8 Diseases (in Bold) and 17 Traits and the Co-occurrence of Obesity with 6 Other Complex Diseases
Circles represent the direction (color) and the two-tailed $-\log_{10}$ p value (size) of the association for different groups of traits (morphometric, metabolic, lipidic, respiratory, and behavioral) and the epidemiological well-established co-occurrence of obesity-related diseases. Inversions are grouped by size and features: (1) submicroscopic are large (0.4–4 Mb) encompassing multiple genes and flanked by segmental duplications; (2) intragenic are located within a gene, either intronic or containing one exon; and (3) intergenic are enriched in pleiotropic regions.

significant association of the N-allele of inversion 11q13.2 with the co-occurrence of obesity with diabetes (OR = 1.05, $p = 0.0011$) and hypertension (OR = 1.03, $p = 2.9 \times 10^{-5}$) (Figure 2), which was not validated in the GERA study.

The study of inversion 16p11.2 also revealed some new significant associations between the inversion and the co-occurrence of obesity with several diseases (Figure 2). The co-occurrence with diabetes at UKB (OR = 1.06, $p = 7.5 \times 10^{-5}$) was independently replicated in the 70KT2D study (OR = 1.13, $p = 1.2 \times 10^{-8}$), where none of the SNPs located within the inverted region were significantly

associated at a genome-wide level (minimum p : 0.0214) (Figure S4B). In addition, the significant co-occurrence with hypertension observed in the UKB study (OR = 1.06, 2.7×10^{-14}) was validated in the GERA study (OR = 1.05, $p = 0.0357$) further confirming the robustness of these findings (see Table 2 reporting the effect of the risk allele N).

In order to further illustrate that the association of the inversion is not driven by single variants, we downloaded data from the GWAS catalog and checked whether the GWAS signals for the analyzed traits are associated (i.e., tags) with the inversions. No tag-SNPs for any of these

Table 2. Association between Inversions 8p23.1 and 16p11.2 and Different Obesity-Related Traits in UKB and Replication Datasets

Disease	Inversion 8p23.1 (Effect of Risk-Haplotype: N-Allele)				Inversion 16p11.2 (Effect of Risk-Haplotype: N-Allele)			
	UKB		Replication		UKB		Replication	
	OR CI95%	p Value	OR CI95%	p Value	OR CI95%	p Value	OR CI95%	p Value
Obesity	1.04 (1.03–1.05)	2.4×10^{-13}	1.08 (1.04–1.11)	5.6×10^{-6}	1.05 (1.04–1.06)	3.9×10^{-24}	1.07 (1.03–1.10)	1.4×10^{-4}
Diabetes	1.04 (1.01–1.06)	1.1×10^{-3}	1.08 (1.05–1.11)	1.1×10^{-8}	1.02 (0.99–1.04)	0.1450	1.07 (1.04–1.11)	1.2×10^{-6}
Hypertension	1.04 (1.03–1.05)	7.0×10^{-16}	1.03 (1.00–1.05)	0.0183	1.01 (1.00–1.02)	0.0184	1.02 (0.99–1.05)	0.2127
Asthma	1.03 (1.01–1.04)	7.0×10^{-5}	1.02 (0.90–1.05)	0.2225	1.00 (0.99–1.01)	0.9529	1.00 (0.97–1.04)	0.8074
Depression	0.98 (0.97–0.99)	0.0119	1.01 (0.97–1.05)	0.6630	0.98 (0.96–1.00)	0.0184	1.01 (0.98–1.05)	0.5384
Joint Occurrence of Obesity with:								
Diabetes	1.08 (1.05–1.11)	3.1×10^{-7}	1.17 (1.12–1.22)	1.4×10^{-33}	1.06 (1.03–1.08)	7.5×10^{-5}	1.13 (1.08–1.17)	1.2×10^{-8}
Hypertension	1.07 (1.05–1.08)	1.7×10^{-16}	1.06 (1.02–1.11)	6.9×10^{-3}	1.06 (1.05–1.07)	2.7×10^{-14}	1.05 (1.00–1.10)	0.0357
Asthma	1.08 (1.06–1.10)	3.0×10^{-11}	1.09 (1.02–1.16)	9.7×10^{-3}	1.05 (1.03–1.07)	7.4×10^{-6}	1.08 (1.01–1.15)	0.0287
Depression	1.04 (1.02–1.07)	1.4×10^{-3}	1.12 (1.04–1.20)	3.8×10^{-3}	1.06 (1.03–1.08)	1.4×10^{-6}	1.03 (0.95–1.11)	0.5241

The table shows the odds ratios (OR) and their confidence intervals at 95% (CI95%) for the non-inverted allele and different diseases and the joint co-occurrence with obesity at UKB and replication datasets. The p corresponds to the best genetic model depict in the first column of each inversion.

traits were found. In particular, the results for the three inversions associated with the co-occurrence of obesity with other traits showed the following results: the median R^2 between SNPs in the 8p23.1 region and the inversion was 0.36 (IQR: 0.17–0.46), 0.71 (IQR: 0.62–0.89) for the inversion 16p11.2, and all the SNPs are not associated (i.e., linkage equilibrium) ($R^2 < 0.06$) for the inversion 11q13.2.

Regulatory Region and Gene Disruption Are the Mechanisms Underlying the Effect of Inversions on Obesity and Diabetes

To investigate the possible mechanisms underlying the shared genetic influences of the inversions with obesity and its co-morbidities, we analyzed the transcriptional effects of the 21 inversions on different tissues from the GTEx project (see [Material and Methods](#)). As a result of these analyses, we found that inversion 8p23.1 modulated the transcription in brain, pancreas, and adipose tissue of the pseudogene *FAM86B3P* (HGNC: 44371), as well as the genes *MFHAS1* (MIM: 605352), *IL19* (MIM: 605687), *HAND2* (MIM: 602407), *FDFT1* (MIM: 184420), *FAM167A* (MIM: 610085), *ER11* (MIM: 608739), *CHAC1* (MIM: 614587), *CCL22* (MIM: 602957), *CCL19* (MIM: CCL19), and *BLK* (MIM: 191305) in other tissues ([Figure 3D](#)). Genes *FDFT1* (MIM: 184420), *C8orf13* (MIM: 610085), *CLDN23* (MIM: 609203), *NEIL2* (MIM: 608933), *MTMR9* (MIM: 606260), *MSRA* (MIM: 606260), and *BLK* (MIM: 191305) and were also differentially expressed in blood samples from the validation study we performed in the independent general population cohort belonging to EGCUT Biobank ([Figure 3E](#)). For the inversion 16p11.2, we found a total of 30 genes differentially expressed at 5% FDR level in blood, brain, pancreas, or adipose tissue including *TUFM* (MIM: 602389), *SULT1A2* (MIM:

601292), *SPNS1* (MIM: 612583), *EIF3CL* (MIM: 603916), and *FOXO1* (MIM: 136533) among many others ([Table S2](#)). These results were also observed in the blood samples of the validation cohort from EGCUT Biobank ([Figure S5](#)). The genes affected by the other inversions and the different tissues can be found in [Table S2](#).

Inversions 8p23.1 and 16p11.2 Affect Key Genes Associated with Diabetes in Pancreatic Islets

We conducted a more detailed analysis of gene expression on a relevant tissue to support the association on diabetic/obese individuals. We first genotyped the inversions and analyzed RNA sequencing in human pancreatic islets from 89 deceased donors (see [Material and Methods](#)). This revealed a significant association between inversion 8p23.1 and the expression levels of *CLDN23* ($p = 1.3 \times 10^{-3}$) and *ER11* ($p = 0.0356$). We observed a nominally significant interaction of inversion 8p23.1 with obese/diabetic status associated with the expression of lncRNA *FAM66A* (HGNC: 30444) ($p = 0.0254$), where individuals carrying the risk allele for obesity and diabetes also present *FAM66A* downregulation ([Figure 3F](#)). In addition, results with inversion at 16p11.2 also revealed a significant interaction between the inversion and obese/diabetic status for the expression of *NUPRI* (MIM: 614812) ($p = 0.0116$) and *ATXN2L* (MIM: 607931) ($p = 0.0167$) ([Figure S6](#)).

cis-Regulation is Disrupted by Breakpoints of Inversions 8p23.1 and 16p11.2

We also investigated whether the positional effects of the inversions could be associated with diabetes (see [Material and Methods](#)). [Figure 4A](#) shows the chromatin landscape of the region of the inversion 8p23.1 as well as the location of all genes having a significant alteration of expression, including those that are islet-specifically expressed. A cluster of islet-specific genes is located outside the rightmost boundary of the inversion but inside the inversion's

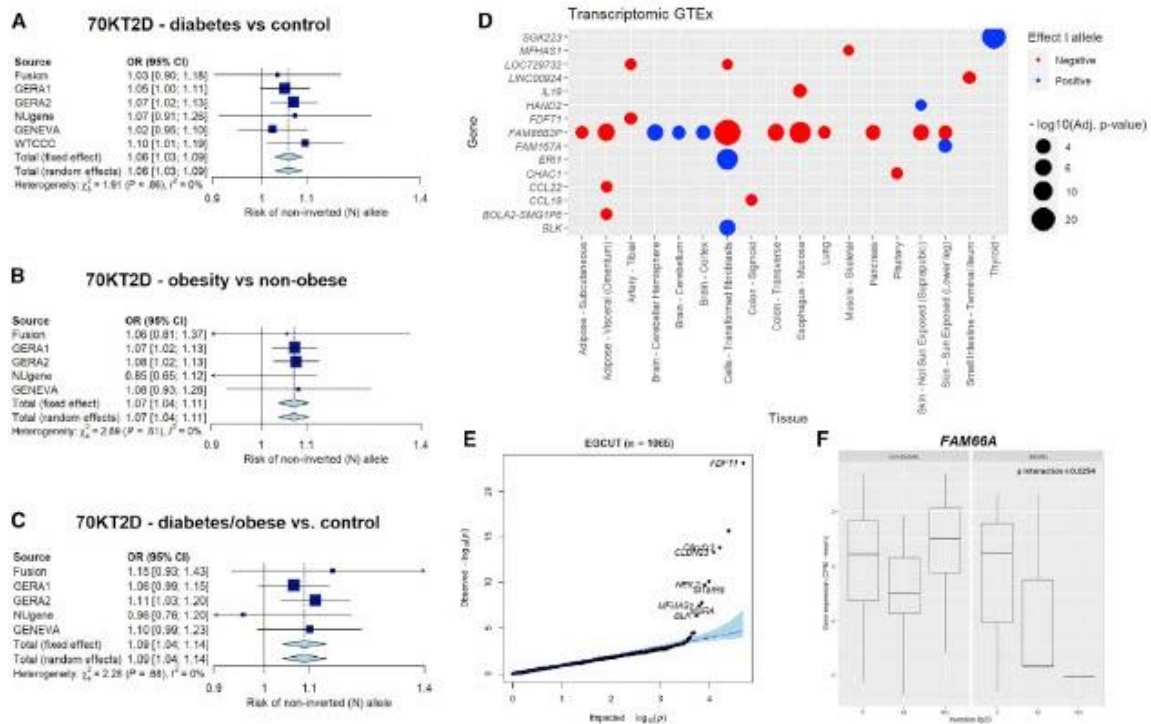


Figure 3. Validation of Positive Associations between the Inversion 8p23.1 with Diabetes, Obesity, and Their Co-occurrence in the 70KT2D Dataset and Transcriptional Allelic Effects in Samples from EGCT Biobank and GTEx Tissues (A–C) 95% confidence intervals and meta-analysis of datasets belonging to 70KT2D for the association of inversion 8p23.1 with diabetes (A), obesity (B), and obese and diabetic individuals (C). (D) Differential expressed genes at inversion genotypes (at 5% FDR) in different tissues from GTEx, showing effect of the I allele (color) and the two-tailed $-\log_{10}$ p value (size) of the association. (E) Differentially expressed genes at inversion genotypes (at 5% FDR) in blood samples from EGCT Biobank. (F) *FAM66A* gene expression interaction between diabetic status and inversion 8p23.1 in pancreatic islets samples ($p = 0.0254$). The box-plots indicate the interquartile range and median of gene expression levels.

topologically associated domains (TADs). Therefore, it is likely that the regulatory regions of these genes lie across the inversion's boundary, and thus their *cis*-regulatory SNPs being separated from their target genes by the right breakpoint of the inversion 8p23.1 in the case of genes *FAM66A* and *FAM66D* (HGNC: 24159) (Figure 4A). Similarly, the analysis of the inversion 16p11.2 also revealed four eQTLs in which the *cis*-regulatory SNPs were separated from their target genes by the inversion breakpoints: *TUFM*, *SULT1A1* (MIM: 171150), *EIF3C* (MIM: 603916), and *EIF3CL* (Figure 4B). *EIF3CL* is disrupted by the inversion breakpoint providing a different mechanism of action for this gene (Figure 4B).

Obesity Mediates the Association of Inversions with Diabetes and Hypertension

We first aimed to disentangle the shared genetic influence of the inversion 8p23.1 in obesity and diabetes. To this end, a Bayesian network analysis was performed on the discovery study (see Material and Methods). Based on the BIC, the most likely model was for the sequence $\text{inv}8p23.1 \rightarrow \text{obesity} \rightarrow \text{diabetes}$, suggesting a mediatory effect of

obesity on the association between the inversion and diabetes (Figure 5A). The same network was obtained in the GERA cohort. This was consistent with mediation analyses showing that 38.7% (CI95%: 25.2%–59.0%) of the diabetes risk variance explained by the inversion 8p23.1 was mediated by obesity ($p < 10^{-16}$). Then, we also investigated whether inversion 8p23.1, 16p11.2, and 11q13.2 act jointly or not on obesity, diabetes, and hypertension. The Bayesian network analysis including the three inversions in the model revealed that the inversions 8p23.1 and 16p11.2 independently associated with diabetes and hypertension being mediated by obesity (Figure 5B).

Discussion

Epidemiological studies largely support the co-occurrence of obesity with numerous traits and diseases such as diabetes, hypertension, asthma, and psychiatric disorders among others.^{40,41} The extent to which obesity is a cause, a consequence, or shares common causes with these traits is subject of intense research.^{42–44} Here, we show that at

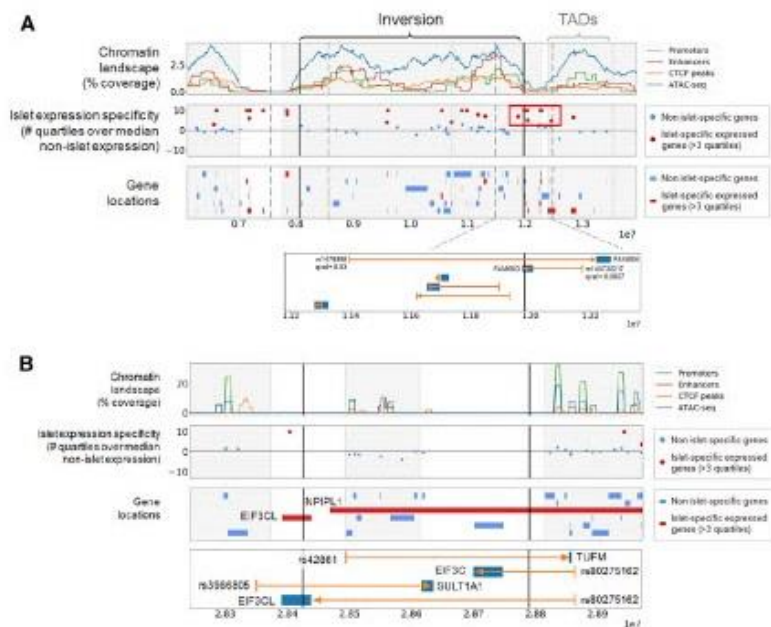


Figure 4. Mechanisms Underlying the Inversion Association with Diabetes

(A) Islet-specific expression of inversion 8p23.1 genes. We observed a cluster of islet-specific genes, mainly lncRNAs, next to the distal inversion breakpoint that could be separated from regulatory elements located inside the inverted region. The bottom panel depicts an eQTLs (rs1478898) of *FAM66A* disrupted by the inversion distal breakpoint.²⁸ *FAM66D* has its gene body split in two by the inversion, and would also have its promoter separated from its eQTL SNP (rs140730217) by the inversion. This could be the most likely causal candidate. (B) Same information for the inversion 16p11.2. *TUFM* and *EIF3C* have their lead eQTL SNP separated by the inversion breakpoint. There is no evidence in the centiSNP database³⁹ for SNP rs42861 to be causal, suggesting that it should be in LD with the causal variant. This promoter region SNP is located in a segmental duplication block that is closer to *TUFM* in the inverted haplotypes. Therefore, positional changes made by the inversion can affect *TUFM* expression by separating the gene from regulatory sequences and subsequently increasing obesity risk.

least two common polymorphic inversions at 8p23.1 and 16q11.2 offer a genetic substrate to some widely observed co-morbidities of obesity, such as those with diabetes, hypertension, asthma, and depression.

The analysis of UKB dataset validated the estimated inversion allele frequencies in European populations reported in our recent analyses.²⁰ The observed differences of some inversion allele frequencies among major populations could explain part of the existing geographic variability in disease incidence.⁴⁵ In particular, the reported cline of the inversion at 8p23.1 and 16p11.2 could capture a proportion of the observed North-South European differences in obesity,⁴⁶ diabetes, and hypertension⁴⁷ incidence.

The analysis of our discovery sample also confirmed previous reported associations of inversions with phenotypes, such as neuroticisms for the inversions 17q21.31 and 8p23.1,¹⁰ obesity for inversion 8p23.1,⁸ and the co-occurrence of asthma and obesity with the inversion 16p11.2.¹⁶ In addition, we discovered and robustly validated new associations of the inversion 8p23.1 with diabetes and hypertension as well as the co-occurrences of obesity with diabetes, hypertension, and asthma. These results suggest a relevant role of the inversion 8p23.1 in this metabolic syndrome.⁴⁸

Our data suggest a causal path in which obesity mediates the observed association between inversions and several complex diseases. In particular, obesity mediates the independent effect of inversions at 8p23.1 and 16p11.2 on diabetes. Transcriptome analyses have revealed candidate genes to mediate this effect, such as *BLK*, involved in pancreatic β -cell insulin metabolism whose rare mutations

are associated with young age of onset diabetes,⁴⁹ or *FDFT1*, linked to C-reactive protein (CRP) and lipids levels⁵⁰ and one of the strong candidates for obesity in gene expression networks derived from mouse intercrosses.⁵¹ A more specific analysis of transcriptome and eQTLs on pancreatic islets leads to another interesting gene: *FAM66A*. *FAM66* is a multiple copy non-coding gene located in the flanking segmental duplications of the 8p23.1 inversion breakpoint highly expressed in brain and with low-level expression in pancreas. Diabetic individuals carrying the N-allele have lower gene expression, while no differential expression across inversion genotypes is observed in control individuals. Consistently, allele-specific expression analysis of this gene shows clear differences in expression in pancreatic cells of already symptomatic diabetic subjects. Remarkably, a copy-number gain variant including *FAM66* gene has been associated with increased risk of diabetes.⁵² Our positional analyses also pointed out at *FAM66D* (8p23.1) as a candidate since the gene body was split in two by the inversion breakpoint.

We have also shown that inversion at 16p11.2 affects the joint effect of obesity with diabetes and hypertension and that this effect is independent of the effect found for inversion 8p23.1. Moreover, the odds ratios found for these associations are stronger than those observed when analyzing those diseases independently. The functional consequences of this inversion were previously reported to be mediated by deregulation of *TUFM*, *SULT1A1*, *SULT1A2*, *SH2B1* (MIM: 608937), *APOB48R* (MIM: 605220), and *EIF3C* in blood.¹⁶ Position transcriptional analysis in pancreatic islets revealed that *TUFM* and

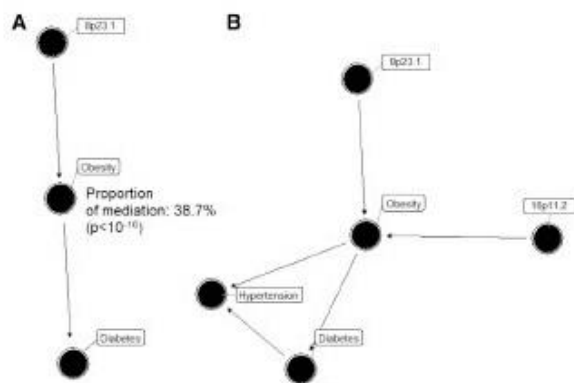


Figure 5. Mediation Effect of Obesity in the Causal Link between Inversions and Diabetes and Hypertension

(A) Mediation analysis of obesity in the association between inversion 8p23.1 and diabetes, showing a proportion of the mediation of 38% (p value < 10⁻⁶), which is the Best Bayesian Network when analyzing these three variables. Significant test for the proportion of the median showed a p value < 10⁻¹⁶. (B) Best Bayesian Network based on AIC obtained after including obesity, hypertension, diabetes, and inversions 8p23.1, 16p11.2, and 11q13.2. Results are obtained from UKB data.

EIF3C have their lead eQTL SNPs separated in the inverted allele. Remarkably, the eQTL SNP rs42861 of *TUFM* does not seem to be causal in the centiSNP database,³⁹ suggesting that it is in linkage disequilibrium with the causal variant. This SNP is located in the promoter region that is closer to *TUFM* in the inverted haplotypes. This supports the hypothesis that the positional changes made by the inversion can affect *TUFM* gene expression and subsequently have an effect increasing the risk for obesity/diabetes. Positional analyses also pointed out *EIF3CL*, a gene also split in two by the inversion breakpoint, and with some isoforms preferentially expressed in human pancreatic islets.³¹

The inversions at 8p23.1 and 16p11.2 were also associated with the joint occurrence of obesity with behavioral traits, in particular with depression. These data further support our hypothesis that polymorphic inversions are strong candidates for the joint genetic susceptibility to co-occurring diseases by simultaneously affecting multiple genes. The observation that some SNPs located in both inversion regions are not or weakly associated with the analyzed traits, while inversion haplotypes are associated even at genome-wide significant level for GWAS with the strongest association found in people having more than one disease, also indicate that inversions are main contributors to the shared genetic susceptibility of co-occurring diseases. The fact that inverted alleles do not recombine preserving haplotypes in strong linkage disequilibrium highly suggest that the underlying evolutionary genetic event that has maintained or selected functional eQTLs in *cis* in these haplotypes is the inversion. Functional analyses in the appropriate tissue in case and control subjects, as the one we performed for obesity and diabetes, will shed

light into the genes and mechanisms involved in behavioral or psychiatric traits.

Our hypothesis that inversions underlie the shared genetic susceptibility to common diseases is particularly supported by our findings in large inversions. These inversions encapsulate multiple genes and their associations with phenotypes were highly significant and could be replicated. Smaller inversions showed significant effects for numerous traits in the discovery study but only one result could be confirmed, namely the correlation of inversion at 11q13.2 with obesity and related traits and also with the co-occurrence of obesity, hypertension, and diabetes. Similarly, this study opens the door to further association studies of these and other inversions with traits and disorders not studied in this work. Additionally, the large number of significant genes associated with different tissues as well as the significant associations found for some traits also provides good candidate genes for some human diseases that are likely under the influence of inversions. These include, among others, autism, Alzheimer disease, and Parkinson disease.

In conclusion, we report the largest association study of genomic inversions and human traits that represents a breakthrough for genomic association of comorbid disorders, in which polymorphic inversions were often previously disregarded. Our results underscore the role of some inversions as major genetic contribution to the joint susceptibility to common diseases. The results in obesity and diabetes reveal a mechanism in which *cis*-regulatory SNPs are separated from their target genes by inversion breakpoints. Our findings set a new framework for future studies which are now accessible to the research community thanks to inversion genotyping tools such as our scoreInvHap method.²⁰

Supplemental Data

Supplemental Data can be found online at <https://doi.org/10.1016/j.ajhg.2020.04.017>.

Acknowledgments

This research has received funding from Ministerio de Ciencia, Innovación y Universidades (MICIU), Agencia Estatal de Investigación (AEI), and Fondo Europeo de Desarrollo Regional, UE (RTI2018-100789-B-I00) also through the "Centro de Excelencia Severo Ochoa 2019-2023" Program (CEX2018-000806-S); and the Catalan Government (SGR2017/801 and #016FLB 00272 to C.R.-A.) through the CERCA Program. J.G. is funded by the European Commission (H2020-ERC-2014-CoG-647900) and the MINECO/AEI/FEDER, EU (BFU2017-82937-P). The L.A.P.-J. lab was funded by the Spanish Ministry of Science and Innovation (ISCIII-FEDER P13/02481), the Catalan Department of Economy and Knowledge (SGR2014/1468, SGR2017/1974, and ICREA Acadèmia), and also acknowledges support from the Spanish Ministry of Economy and Competitiveness "Programa de Excelencia María de Maeztu" (MDM-2014-0370). This research was conducted using the UK Biobank Resource under Application Number

43983. The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS.

Declaration of Interests

L.A.P.-J. is a founding partner and scientific advisor of qGenomics Laboratory. All other authors declare no conflict of interest.

Received: January 7, 2020

Accepted: April 28, 2020

Published: May 28, 2020

Web Resources

dbGaP, <https://www.ncbi.nlm.nih.gov/gap>

European Genome-phenome Archive (EGA), <https://www.ebi.ac.uk/ega>

GWAS Catalog, <https://www.ebi.ac.uk/gwas/>

HUGO Gene Nomenclature Committee (HGNC), <https://www.genenames.org/>

imputeInversion, <https://github.com/isglobal-brge/imputeinversion>
Inversion Associations, https://github.com/isglobal-brge/inversion_analyses

OMIM, <https://www.omim.org/>

WHO, <http://www.euro.who.int/en/health-topics/disease-prevention/nutrition/a-healthy-lifestyle/body-mass-index-bmi>

References

1. GBD 2015 Obesity Collaborators, Afshin, A., Forouzanfar, M.H., Reitsma, M.B., Sur, P., Estep, K., Lee, A., Marczak, L., Mokdad, A.H., Moradi-Lakeh, M., et al. (2017). Health Effects of Overweight and Obesity in 195 Countries over 25 Years. *N. Engl. J. Med.* 377, 13–27.
2. Dixon, J.B. (2010). The effect of obesity on health outcomes. *Mol. Cell. Endocrinol.* 316, 104–108.
3. Locke, A.E., Kahali, B., Berndt, S.I., Justice, A.E., Pers, T.H., Day, F.R., Powell, C., Vedantam, S., Buchkovich, M.L., Yang, J., et al.; Lifelines Cohort Study; ADIPOGen Consortium; AGEN-BMI Working Group; CARDIOGRAMplusC4D Consortium; CKDGen Consortium; GLGC; ICBP; MAGIC Investigators; MuTHER Consortium; MiGen Consortium; PAGE Consortium; ReproGen Consortium; GENIE Consortium; and International Endogene Consortium (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature* 518, 197–206.
4. Serra-Juhé, C., Martos-Moreno, G.Á., Bou de Pieri, F., Flores, R., González, J.R., Rodríguez-Santiago, B., Argente, J., and Pérez-Jurado, L.A. (2017). Novel genes involved in severe early-onset obesity revealed by rare copy number and sequence variants. *PLoS Genet.* 13, e1006657.
5. Kaminsky, E.B., Kaul, V., Paschall, J., Church, D.M., Bunke, B., Kunig, D., Moreno-De-Luca, D., Moreno-De-Luca, A., Mulle, J.G., Warren, S.T., et al. (2011). An evidence-based approach to establish the functional and clinical significance of copy number variants in intellectual and developmental disabilities. *Genet. Med.* 13, 777–784.
6. Selvanayagam, T., Walker, S., Gazzellone, M.J., Kellam, B., Cytynbaum, C., Stavropoulos, D.J., Li, P., Birken, C.S., Hamilton, J., Weksberg, R., and Scherer, S.W. (2018). Genome-wide copy number variation analysis identifies novel candidate loci associated with pediatric obesity. *Eur. J. Hum. Genet.* 26, 1588–1596.
7. Vuillaume, M.L., Naudion, S., Banneau, G., Diene, G., Cartault, A., Cailley, D., Bouron, J., Toutain, J., Bourrouillou, G., Vigouroux, A., et al. (2014). New candidate loci identified by array-CGH in a cohort of 100 children presenting with syndromic obesity. *Am. J. Med. Genet. A.* 164A, 1965–1975.
8. Cáceres, A., and González, J.R. (2015). Following the footprints of polymorphic inversions on SNP data: from detection to association tests. *Nucleic Acids Res.* 43, e53.
9. Gutiérrez Arumi, A. (2015). Ancestral genomic submicroscopic inversions of human genome and their relation with multifactorial human diseases (Univ. Pompeu Fabra).
10. Okbay, A., Baselmans, B.M., De Neve, J.-E., Turley, P., Nivard, M.G., Fontana, M.A., Meddens, S.F., Linnér, R.K., Rietveld, C.A., Derringer, J., et al.; Lifelines Cohort Study (2016). Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nat. Genet.* 48, 624–633.
11. Karlsson Linnér, R., Biroli, P., Kong, E., Meddens, S.F.W., Wedow, R., Fontana, M.A., Lebreton, M., Tino, S.P., Abdellaoui, A., Hammerschlag, A.R., et al.; 23and Me Research Team; eQTLgen Consortium; International Cannabis Consortium; and Social Science Genetic Association Consortium (2019). Genome-wide association analyses of risk tolerance and risky behaviors in over 1 million individuals identify hundreds of loci and shared genetic influences. *Nat. Genet.* 51, 245–257.
12. Laws, S.M., Friedrich, P., Diehl-Schmid, J., Müller, J., Eisele, T., Büml, J., Förstl, H., Kurz, A., and Riemenschneider, M. (2007). Fine mapping of the MAPT locus using quantitative trait analysis identifies possible causal variants in Alzheimer's disease. *Mol. Psychiatry* 12, 510–517.
13. Zabetian, C.P., Hutter, C.M., Factor, S.A., Nutt, J.G., Higgins, D.S., Griffith, A., Roberts, J.W., Leis, B.C., Kay, D.M., Yearout, D., et al. (2007). Association analysis of MAPT H1 haplotype and subhaplotypes in Parkinson's disease. *Ann. Neurol.* 62, 137–144.
14. Pilbrow, A.P., Lewis, K.A., Perrin, M.H., Sweet, W.E., Moravec, C.S., Tang, W.H.W., Huising, M.O., Troughton, R.W., and Cameron, V.A. (2016). Cardiac CRFR1 Expression Is Elevated in Human Heart Failure and Modulated by Genetic Variation and Alternative Splicing. *Endocrinology* 157, 4865–4874.
15. Ikram, M.A., Fornage, M., Smith, A.V., Seshadri, S., Schmidt, R., Debette, S., Vrooman, H.A., Sigurdsson, S., Ropele, S., Taal, H.R., et al.; Early Growth Genetics Consortium; and Cohorts for Heart and Aging Research in Genomic Epidemiology Consortium (2012). Common variants at 6q22 and 17q21 are associated with intracranial volume. *Nat. Genet.* 44, 539–544.
16. González, J.R., Cáceres, A., Esko, T., Cusco, I., Puig, M., Esnaola, M., Reina, J., Siroux, V., Bouzigon, E., Nadif, R., et al. (2014). A common 16p11.2 inversion underlies the joint susceptibility to asthma and obesity. *Am. J. Hum. Genet.* 94, 361–372.
17. de Jong, S., Chepelev, I., Janson, E., Strengman, E., van den Berg, L.H., Veldink, J.H., and Ophoff, R.A. (2012). Common inversion polymorphism at 17q21.31 affects expression of multiple genes in tissue-specific manner. *BMC Genomics* 13, 458.
18. Chaisson, M.J.P., Sanders, A.D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., Gardner, E.J., Rodriguez, O.L., Guo, L.,

- Collins, R.L., et al. (2019). Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* *10*, 1784.
19. Giner-Delgado, C., Villatoro, S., Lerga-Jaso, J., Gayà-Vidal, M., Oliva, M., Castellano, D., Pantano, L., Bitarello, B.D., Izquierdo, D., Noguera, I., et al. (2019). Evolutionary and functional impact of common polymorphic inversions in the human genome. *Nat. Commun.* *10*, 4222.
 20. Ruiz-Arenas, C., Cáceres, A., López-Sánchez, M., Tolosana, I., Pérez-Jurado, L., and González, J.R. (2019). scoreInvHap: Inversion genotyping for genome-wide association studies. *PLoS Genet.* *15*, e1008203.
 21. Pickrell, J.K., Berisa, T., Liu, J.Z., Séguire, L., Tung, J.Y., and Hinds, D.A. (2016). Detection and interpretation of shared genetic influences on 42 human traits. *Nat. Genet.* *48*, 709–717.
 22. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* *12*, e1001779.
 23. Bonás-Guarch, S., Guindo-Martínez, M., Miguel-Escalada, I., Grarup, N., Sebastian, D., Rodríguez-Fos, E., Sánchez, F., Planas-Félix, M., Cortes-Sánchez, P., González, S., et al. (2018). Re-analysis of public genetic data reveals a rare X-chromosomal variant associated with type 2 diabetes. *Nat. Commun.* *9*, 321.
 24. Pedersen, B.S., and Quinlan, A.R. (2017). Who's Who? Detecting and Resolving Sample Anomalies in Human DNA Sequencing Studies with Peddy. *Am. J. Hum. Genet.* *100*, 406–413.
 25. Collado-Torres, L., Nellore, A., Kammers, K., Ellis, S.E., Taub, M.A., Hansen, K.D., Jaffe, A.E., Langmead, B., and Leek, J.T. (2017). Reproducible RNA-seq analysis using recount2. *Nat. Biotechnol.* *35*, 319–321.
 26. Law, C.W., Chen, Y., Shi, W., and Smyth, G.K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* *15*, R29.
 27. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* *43*, e47.
 28. van de Bunt, M., Manning Fox, J.E., Dai, X., Barrett, A., Grey, C., Li, L., Bennett, A.J., Johnson, P.R., Rajotte, R.V., Gaulton, K.J., et al. (2015). Transcript Expression Data from Human Islets Links Regulatory Signals from Genome-Wide Association Studies for Type 2 Diabetes and Glycemic Traits to Their Downstream Effectors. *PLoS Genet.* *11*, e1005694.
 29. Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* *485*, 376–380.
 30. Pasquali, L., Gaulton, K.J., Rodríguez-Seguí, S.A., Mularoni, L., Miguel-Escalada, I., Akerman, I., Tena, J.J., Morán, I., Gómez-Maín, C., van de Bunt, M., et al. (2014). Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nat. Genet.* *46*, 136–143.
 31. Miguel-Escalada, I., Bonás-Guarch, S., Cebola, I., Ponsa-Cobas, J., Mendieta-Esteban, J., Atla, G., Javierre, B.M., Rolando, D.M.Y., Farabella, I., Morgan, C.C., et al. (2019). Human pancreatic islet three-dimensional chromatin architecture provides insights into the genetics of type 2 diabetes. *Nat. Genet.* *51*, 1137–1148.
 32. Fadista, J., Vikman, P., Laakso, E.O., Mollet, I.G., Esguerra, J.L., Taneera, J., Storm, P., Osmark, P., Ladenvall, C., Prasad, R.B., et al. (2014). Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism. *Proc. Natl. Acad. Sci. USA* *111*, 13924–13929.
 33. Delaneau, O., Zagury, J.-F., and Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* *10*, 5–6.
 34. Das, S., Forer, L., Schönerr, S., Sidore, C., Locke, A.E., Kwong, A., Vrieze, S.I., Chew, E.Y., Levy, S., McGue, M., et al. (2016). Next-generation genotype imputation service and methods. *Nat. Genet.* *48*, 1284–1287.
 35. González, J.R., Armengol, L., Solé, X., Guinó, E., Mercader, J.M., Estivill, X., and Moreno, V. (2007). SNPassoc: an R package to perform whole genome association studies. *Bioinformatics* *23*, 644–645.
 36. Li, J., and Ji, L. (2005). Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Hereditas* *95*, 221–227.
 37. Tingley, D., Yamamoto, T., Hirose, K., Keele, L., and Imai, K. (2014). mediation : R Package for Causal Mediation Analysis. *J. Stat. Softw.* *59*, 1–38.
 38. Lewis, F.I., and Ward, M.P. (2013). Improving epidemiologic data analyses through multivariate regression modelling. *Emerg. Themes Epidemiol.* *10*, 4.
 39. Moyerbrailean, G.A., Kalita, C.A., Harvey, C.T., Wen, X., Luca, F., and Pique-Regi, R. (2016). Which Genetics Variants in DNase-Seq Footprints Are More Likely to Alter Binding? *PLoS Genet.* *12*, e1005875.
 40. Banks, J., Marmot, M., Oldfield, Z., and Smith, J.P. (2006). Disease and disadvantage in the United States and in England. *JAMA* *295*, 2037–2045.
 41. Stunkard, A.J., Faith, M.S., and Allison, K.C. (2003). Depression and obesity. *Biol. Psychiatry* *54*, 330–337.
 42. Martins-Silva, T., Vaz, J.D.S., Hutz, M.H., Salatino-Oliveira, A., Genro, J.P., Hartwig, F.P., Moreira-Maia, C.R., Rohde, L.A., Borges, M.C., and Tovo-Rodrigues, L. (2019). Assessing causality in the association between attention-deficit/hyperactivity disorder and obesity: a Mendelian randomization study. *Int. J. Obes.* *43*, 2500–2508.
 43. Xu, S., Gilliland, F.D., and Conti, D.V. (2019). Elucidation of causal direction between asthma and obesity: a bi-directional Mendelian randomization study. *Int. J. Epidemiol.* *48*, 899–907.
 44. Millard, L.A.C., Davies, N.M., Tilling, K., Gaunt, T.R., and Davey Smith, G. (2019). Searching for the causal effects of body mass index in over 300 000 participants in UK Biobank, using Mendelian randomization. *PLoS Genet.* *15*, e1007951.
 45. Puig, M., Casillas, S., Villatoro, S., and Cáceres, M. (2015). Human inversions and their functional consequences. *Brief. Funct. Genomics* *14*, 369–379.
 46. Berghöfer, A., Pischon, T., Reinhold, T., Apovian, C.M., Sharma, A.M., and Willich, S.N. (2008). Obesity prevalence from a European perspective: a systematic review. *BMC Public Health* *8*, 200.
 47. Wolf-Maier, K., Cooper, R.S., Banegas, J.R., Giampaoli, S., Hense, H.-W., Joffres, M., Kasterinen, M., Poulter, N., Primatesta, P., Rodríguez-Artalejo, F., et al. (2003). Hypertension prevalence and blood pressure levels in 6 European countries, Canada, and the United States. *JAMA* *289*, 2363–2369.

48. Povel, C.M., Boer, J.M.A., Reiling, E., and Feskens, E.J.M. (2011). Genetic variants and the metabolic syndrome: a systematic review. *Obes. Rev.* 12, 952–967.
49. Borowiec, M., Liew, C.W., Thompson, R., Boonyasrisawat, W., Hu, J., Mlynarski, W.M., El Khattabi, I., Kim, S.-H., Marselli, L., Rich, S.S., et al. (2009). Mutations at the BLK locus linked to maturity onset diabetes of the young and beta-cell dysfunction. *Proc. Natl. Acad. Sci. USA* 106, 14460–14465.
50. Ligthart, S., Vaez, A., Hsu, Y.-H., Stolk, R., Uitterlinden, A.G., Hofman, A., Alizadeh, B.Z., Franco, O.H., Dehghan, A.; Inflammation Working Group of the CHARGE Consortium; PMI-WG-XCP; and Lifelines Cohort Study (2016). Bivariate genome-wide association study identifies novel pleiotropic loci for lipids and inflammation. *BMC Genomics* 17, 443.
51. Logsdon, B.A., Hoffman, G.E., and Mezey, J.G. (2012). Mouse obesity network reconstruction with a variational Bayes algorithm to employ aggressive false positive control. *BMC Bioinformatics* 13, 53.
52. Bailey, J.N.C., Lu, L., Chou, J.W., Xu, J., McWilliams, D.R., Howard, T.D., Freedman, B.L., Bowden, D.W., Langefeld, C.D., and Palmer, N.D. (2013). The Role of Copy Number Variation in African Americans with Type 2 Diabetes-Associated End Stage Renal Disease. *J. Mol. Genet. Med.* 7, 61.

GENOME WIDE ASSOCIATION STUDIES

REVIEW PUBLICATION

Review

In Search of Complex Disease Risk through Genome Wide Association Studies

Lorena Alonso ^{1,*} , Ignasi Morán ^{1,*} , Cecilia Salvoro ^{1,*}  and David Torrents ^{1,2}

¹ Life Sciences Department, Barcelona Supercomputing Center (BSC), 08034 Barcelona, Spain; david.torrents@bsc.es

² Institució Catalana de Recerca i Estudis Avançats (ICREA), 08010 Barcelona, Spain

* Correspondence: lorena.alonso@bsc.es (L.A.); ignasi.moran@bsc.es (I.M.); cecilia.salvoro@bsc.es (C.S.)

Abstract: The identification and characterisation of genomic changes (variants) that can lead to human diseases is one of the central aims of biomedical research. The generation of catalogues of genetic variants that have an impact on specific diseases is the basis of Personalised Medicine, where diagnoses and treatment protocols are selected according to each patient's profile. In this context, the study of complex diseases, such as Type 2 diabetes or cardiovascular alterations, is fundamental. However, these diseases result from the combination of multiple genetic and environmental factors, which makes the discovery of causal variants particularly challenging at a statistical and computational level. Genome-Wide Association Studies (GWAS), which are based on the statistical analysis of genetic variant frequencies across non-diseased and diseased individuals, have been successful in finding genetic variants that are associated to specific diseases or phenotypic traits. But GWAS methodology is limited when considering important genetic aspects of the disease and has not yet resulted in meaningful translation to clinical practice. This review presents an outlook on the study of the link between genetics and complex phenotypes. We first present an overview of the past and current statistical methods used in the field. Next, we discuss current practices and their main limitations. Finally, we describe the open challenges that remain and that might benefit greatly from further mathematical developments.

Keywords: bioinformatics; genomics; GWAS; chi-square; logistic regression; generalized linear models; Markov models; imputation; machine learning; polygenic risk scores



Citation: Alonso, L.; Morán, I.; Salvoro, C.; Torrents, D. In Search of Complex Disease Risk through Genome Wide Association Studies. *Mathematics* 2021, 9, 3083. <https://doi.org/10.3390/math9233083>

Academic Editors: Manuel Franco, Juana María Vivo and Xiaoping Liu

Received: 4 October 2021

Accepted: 25 November 2021

Published: 30 November 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Complex traits, such as height, blood pressure, or some types of diseases, arise from the combination of multiple environmental and genetic factors (see Box 1 for definitions of fundamental concepts). In these, each of the involved genetic variants is expected to only make a marginal contribution to the whole phenotype, each explaining <1% and often <0.5%, of phenotypic variability [1–3]. Consequently, hundreds or even thousands of loci are likely to be involved for each trait [4–6]. Complex diseases, such as diabetes [7], asthma [8], cardiovascular diseases [9], or Alzheimer's disease [10], tend to appear late in life and strongly affect the quality of life of millions of individuals around the world, exerting a large economic and social pressure on developed global healthcare systems. For instance, diabetes incurred in an estimated cost of USD 327 billion in 2017 in the United States alone, a value that increased 26% with respect to 2012 [11]. To help alleviate this burden, a long-standing goal of biomedicine has been to gain a better understanding of the molecular mechanisms and the genetic architecture behind these diseases, enabling better prognosis, prevention, and treatment protocols.

In addition to the multifactorial architecture of complex traits, covariate effects, population substructure, or disease heterogeneity [12] make the identification of the underlying causal genomic variants a statistical, mathematical, and computational challenge. The recent increase in sample sizes and the improvement of statistical frames have helped

increase sensibility but have also imposed computational and methodological burdens that are becoming the bottleneck of these types of analyses. This increasing complexity has forced many studies to reduce their overall scope, which they may accomplish by excluding the analysis of the X chromosome or by restricting the analysis of the additive model, disregarding all other inheritance models that should be considered. This substantially limits the chances of identifying novel genetic markers that are associated with disease, as we recently demonstrated [13,14].

Despite these challenges, Genome-Wide Association Studies (GWAS) represent one of the most successful approaches for identifying genetic variants that are associated with the risk of developing particular complex diseases. In this review, we will provide an overview of the statistical models and approaches that are currently applied to the identification of association between genetic variants and complex diseases in biomedical research.

Box 1. Fundamental concepts

- Complex trait or disease: A multifactorial phenotype resulting from the combination of numerous environmental and genetic factors.
- Genome-Wide Association Study (GWAS): A statistical method to discover the genomic variability that is associated with a complex trait or disease.
- Genomic or genetic variant: A genomic location known to present variability within a population.
- Personalised medicine: The application of preventive and treatment protocols adjusted to the patient's genomic profile.
- Phenotype: A measurable characteristic in the individuals of a population, such as height, eye colour, blood pressure, or disease state.

2. Preliminary Genome Biology Concepts

The human genome is considerably variable. Two human beings differ in 4.1–5 million genomic sites on average, for a total of around 20 million bases (~0.6% of the total genome) [15]. This genetic variability determines not only the differences in physical appearance, such as height or eye colour, but also the predisposition of an individual to develop diseases.

Distinguishing the genetic variants that are responsible of normal human variability from those affecting disease risk is thus fundamental to predict, diagnose, and possibly treat diseases, contributing to personalised medicine efforts. In this scenario, GWAS represents a resourceful strategy that can be used to identify variants that are associated with complex diseases. Despite substantial advancements, this remains a challenging task: in complex diseases, the contribution of each of the genetic variants to the final phenotype has been proven to be low and to come later in life, which is in contrast to rare diseases, where variants usually have a much stronger effect in the individual and may already be present during early developmental stages [1,14].

In general terms, each individual inherits this variability through parental germ cells. For example, when the genomic variation consists of a change at a single nucleotide position, it is called a Single Nucleotide Variant (SNV), but larger, structural variants (e.g., duplications, deletions) that have the potential of affecting up to millions of nucleotides also exist (see Box 2 for definitions of genomic concepts). As a result of the meiosis process, any genomic position (loci) is thus present in two copies (alleles). The set of alleles in a single homologous chromosome is defined as a haplotype, and the combination of all alleles identifies the individual's genotype. The study of these genotypes in regard to their relationship with diseases is one of the central aims of biomedicine. It allows us to generate comprehensive genetic maps for each disease and to use them to easily screen, for example, newborns and to be able to predict the disease risk for that newborn and to plan preventive protocols.

Most genomic variants are biallelic, meaning that only two different alleles (generally named *A* and *B*) exist in the population. In this scenario and considering that all individuals have two copies of the genome, at any given variable locus (position), an individual

displays one of three possible genotypes: AA , AB , or BB . When compared to the human reference genome [16], the allele matching the reference (e.g., A) is termed the reference allele, while the other (e.g., B) is termed the alternate allele. Consequently, the three possible genotypes are labelled as the homozygous reference (*hom. ref.* or AA), the homozygous alternate (*hom. alt.* or BB), or heterozygous (*het.* or AB).

Each of these genetic variants, which likely arose from single different individuals, are spread and fixed within the population over long periods of time and follow evolutionary rules based on the harm or benefit that each variation provides to the individual. As a consequence of this process, variants have different frequencies within each population, as they are carried by different proportions of individuals. Variants with frequencies $> 5\%$ are defined as common, while variants with frequencies $1 - 5\%$ or $< 1\%$ are defined as low-frequency and rare, respectively. SNVs with a frequency of $> 1\%$ in the population are typically called Single Nucleotide Polymorphisms (SNPs). Since complex diseases are common, originally, only common variants were considered to be implicated (common disease-common variant hypothesis); the possibility of extending GWAS even to low-frequency and rare variants has shown, however, that variants across the entire frequency spectrum are likely to be involved [3]. The effect size, which is the contribution of these variants to the phenotype, is generally measured by an odds ratio (the odds of having the disease with the variant divided by the odds of having the disease without it) for a binary trait. Typically, an inverse relationship exists between the frequency of a variant and its effect on diseases: high-impact variants are normally found at lower frequencies because of a stronger negative selection pressure (Figure 1) [17].

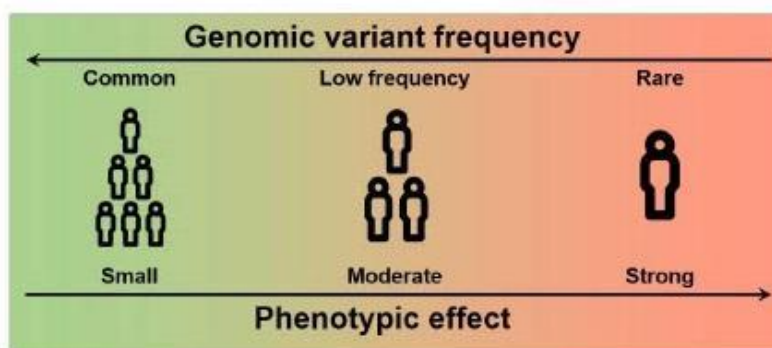


Figure 1. Relationship between allele frequency and effect size. High effect variants tend to have a lower frequency in the population and vice versa.

Finally, it is worth noting that even though $\sim 50\%$ of the genome is inherited from each parent, the nucleotides in a chromosome are not inherited independently. Instead, the genomic material is exchanged in large, linked fragments, that are delimited by recombination hotspots, which are genomic regions that are more prone to recombination. As a result, these large genomic fragments contain multiple alleles that are inherited as a whole from the same parent; these alleles are said to be in linkage disequilibrium (LD).

Given this biological framework, we can now better appreciate the challenges of studying the genomic causes of complex traits and diseases. The main aim is to identify the genomic variability that leads to a higher risk of disease. However, it is likely that there are thousands of genomic loci with different levels of implications and with different frequencies in different populations. Therefore, the identification of unique causal variants is typically obscured by multiple variants in linkage disequilibrium, and the biological consequences of these variants are not immediately apparent. Thus, the study of complex traits and diseases remains an open prospect.

Box 2. Genomic concepts.

- **Allele:** One of the possible genomic sequences that exist in a population for a given locus.
- **Allelic Frequency:** The frequency in which a certain allele is found within a population.
- **Genomic locus:** A region of the genome.
- **Genomic marker:** A specific variant that is used as a proxy for nearby variants in high linkage disequilibrium.
- **Genotype:** The specific combination of alleles of an individual. When compared to a reference genome, the genotype of a variant may be reference homozygous, heterozygous, or alternate homozygous.
- **Haplotype:** The list of alleles that are present in the same homologous chromosome.
- **Inheritance model:** A quantitative model for how the genotype of a variant might contribute to the phenotype. The most frequently used is the additive model, but the dominant, recessive, and heterodominant models are also utilized.
- **Linkage Disequilibrium (LD):** When alleles are inherited together in an individual more often than expected by chance. This is a consequence of the inheritance of these alleles in haplotype blocks instead of them being independent of each other.
- **Single nucleotide variant/polymorphism (SNV/SNP):** The most frequent type of genomic variant, in which the alleles differ in a single nucleotide position. SNPs are SNVs with a frequency of >1%.

3. Genome Wide Association Studies (GWAS)**3.1. Definition**

In order to take on this challenging task, GWAS was proposed as a statistical method that could be used to identify the genomic variants that are associated with complex traits or diseases. Specifically, GWAS are statistical analyses that aim to find the associations between genomic variability and a particular trait or disease [17]. Previous studies have required each functional hypothesis to be specifically tested in the context of a disease. In contrast, GWAS allow for the exploration of the genetic architecture of diseases at the genome-wide level, without the need of prior hypotheses beyond the existence of a genetic component behind the disease.

These studies collect genotypes and phenotypes of a large number of participants, generally in the order of tens of thousands, or even millions. To study a complex disease (binary trait), participants are separated into cases (affected) and controls (non-affected) (Figure 2). Then, a prior characterisation of the variation landscape is needed for each of the participating individuals, i.e., the genotypes and haplotypes, which are inferred from the lists of variants that have been identified within each participant. Whereas whole-genome sequencing currently provides the most complete map of genomic variation for an individual, it is still a very expensive and time-consuming assay, especially when considering the large number of participants within these types of studies. Instead, GWAS typically use DNA hybridisation microarray technologies, a more affordable alternative (see Box 3 for definitions of technical concepts). DNA microarrays, however, are designed to interrogate only a limited set of pre-selected genomic variants (generally between 500 k and 2 M) [18]. These variants are chosen to be common across the population, so that many of the individuals can carry them, and are also chosen considering LD blocks, so that only a single variant in each block is typically probed. In this manner, these subsets of variants are greatly informative and can be used to infer almost the full genotype variability landscape of each individual, as we will discuss in detail later (Section 4.2).

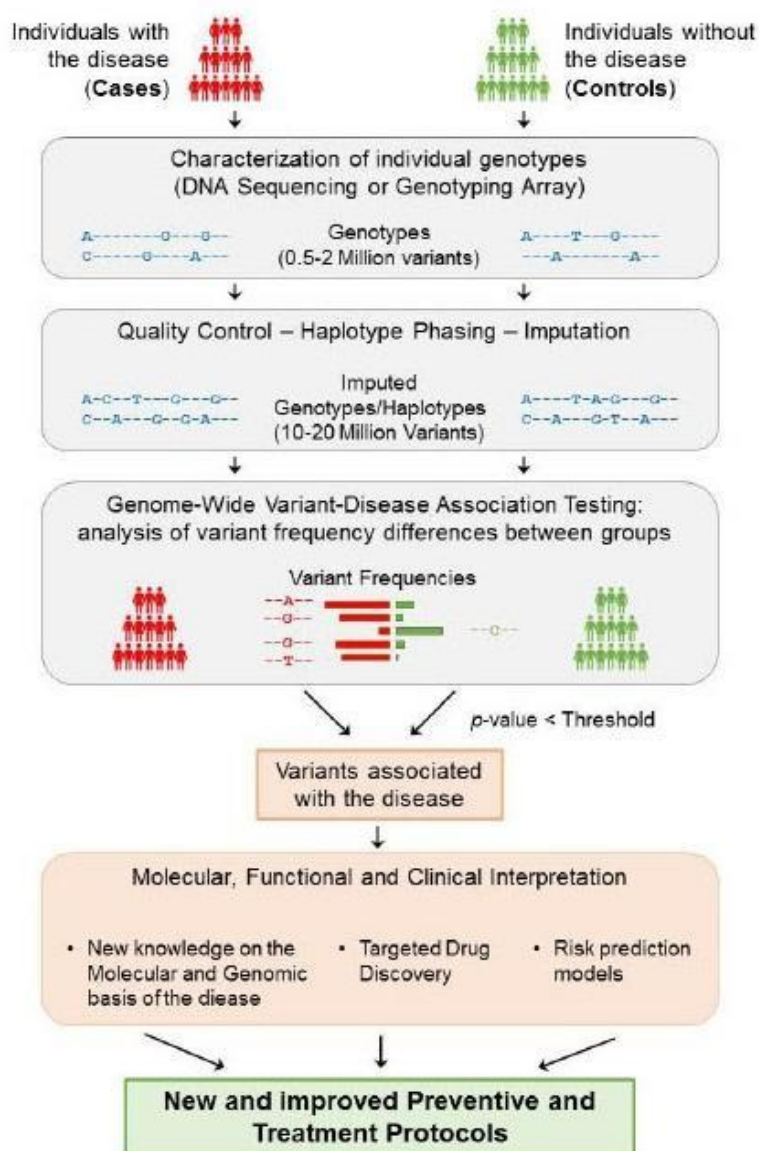


Figure 2 General strategy underlying GWAS. The study of a complex disease through GWAS starts with the selection of a large group of individuals that can be segregated into cases (affected) and controls (non-affected). Then, each individual genotype is characterized using DNA sequencing techniques or genotyping arrays, obtaining the genotyping information of 0.5–2 million variants from each individual. After ensuring the quality of these data, phasing and imputation techniques are usually applied to increase the number of variants that can be tested to 10–20 million. Each resulting genomic variant is then independently tested to find significant differences in the genotype frequencies between the two groups. Consequently, if a variant is significantly predominant in a group based on an adjusted *p*-value threshold, then the variant is said to be associated with the disease. Disease-associated variants can then be further analysed to gain insight into their molecular, functional, and clinical implications. As a result of this process, the knowledge obtained from GWAS can help generate and improve the protocols for the better detection, prevention, and treatment of complex diseases.

Then, each genomic variant is independently tested for significant differences in the genotype frequency between the two groups. Thus, if a variant is found to be present significantly more frequently in cases than they are in controls (or vice versa), then that variant is said to be associated with the disease (Figure 2). If the study is sufficiently powered, then a few genomic loci (containing a small number of variants, typically in high LD) will be identified as being significantly associated with the phenotype. For quantitative traits, the individual phenotypes are usually expressed as a continuous variable, and the association is evaluated based on the correlation between the trait and each variant genotype.

Finally, the genomic variants that are significantly associated with a trait or disease (termed “GWAS variants”) provide a list of candidates for further functional analyses to determine in which way they affect the function of the cell and, in the case of disease, ultimately help provide better prevention and treatment protocols.

3.2. Analytical Frameworks for GWAS

With the increasing interest in the study of complex traits, several statistical frameworks and tools have been developed in recent years in order to perform GWAS analyses [19]. In the following subsections, we will explain how these statistical models test for associations between genomic variability and phenotypes. We will mainly discuss methods to perform GWAS on binary traits (i.e., diseases). However, the analysis of quantitative traits is also presented. Moreover, given that the additive model is the most common in GWAS, the methodology will be formulated under this model. However, in Section 3.2.1, we will showcase how to work with the non-additive inheritance models. Hence, we will start with a simple model for binary traits by first detailing the use of contingency tables (Section 3.2.1) and will move towards more complete models, such as logistic regression (Section 3.2.2), regression model extensions (Section 3.2.3), and Bayesian regression analyses (Section 3.2.4).

In all of these analyses, to statistically model a GWAS, it is first necessary to define:

- The number of individuals included in the sample of the study N . In binary traits, these individuals are divided according to their phenotype, i.e., into N_a cases (diseased) and N_o controls (non-diseased), where $N = N_a + N_o$.
- A set of genomic variants $\{V_1, \dots, V_m\}$, $m \in \{1, \dots, M | M < \infty\}$ that are analysed for each individual present in the population.
- The genotype G_i for each variant, which can take a genotype value from $\{AA, AB, BB\} = \{hom.ref, het, hom.alt\}$. This genotype can be encoded differently depending on the hypothesised inheritance model by defining a function $f: G_{ij} \rightarrow \{0, 1, 2\}$, where $\{0, 1, 2\}$ encodes for additive ($f(AA) = 0, f(AB) = 1, f(BB) = 2$), $\{0, 1, 1\}$ for dominant ($f(AA) = 0, f(AB) = 1, f(BB) = 1$), $\{0, 0, 1\}$ for recessive ($f(AA) = 0, f(AB) = 0, f(BB) = 1$), or $\{0, 1, 0\}$ for heterodominant ($f(AA) = 0, f(AB) = 1, f(BB) = 0$). For the purpose of statistical testing, one of the alleles, typically the alternate, is defined as the effect allele.
- Based on the space defined by the genotype, each genomic variant V_i can be considered as a simple random variable $V_i: \Omega \rightarrow G_i$, so that $\forall g \in G_i \exists \omega \in \Omega$ for which $V_i(\omega) = g$, with Ω as the space of events.
- The phenotype P_j for each individual in the population is given a trait of study, which, in the case of binary traits, is assigned as $\{0, 1\} = \{control, case\} = \{diseased, non-diseased\}$. The phenotype can be modelled by a Bernoulli distribution $P_j \sim B(p_j)$, where p_j is the unknown probability of an individual having the disease.

Then, for each tested genomic variant, two outputs are expected:

- A measure of the statistical confidence on the association with the phenotype in the form of a p -value.
- A measure of the effect size of having one of the alleles, which is typically expressed by beta (β) for quantitative traits and an odds ratio (OR) for binary traits.

3.2.1. Contingency Tables

The classical approach for finding associations between genotypes and a binary phenotype consists of constructing a 2×2 contingency table of the allelic counts in each group. Once the contingency table is prepared, the allele frequencies can be measured and tested to find any possible relation with the disease [20].

First, given a specific variant V_i in a population with N individuals, where N_a are cases (diseased) and N_o are controls (non-diseased) and where $N = N_a + N_o$, for each individual j from the population of study, the space of the genotypes of each variant $G_{ij} = \{AA, AB, BB\} = \{hom.ref, het, hom.alt\}$ can be defined. Thus, the contingency table of the observed genotype counts in the population of study (Table 1) is constructed as:

Table 1. Contingency table of observed genotypes.

	AA	AB	BB	Total
Cases	$n_{hom.ref.a}$	$n_{het.a}$	$n_{hom.alt.a}$	N_a
Controls	$n_{hom.ref.o}$	$n_{het.o}$	$n_{hom.alt.o}$	N_o
Total	$n_{hom.ref}$	n_{het}	$n_{hom.alt}$	N

Moreover, given that the genotype is defined by two alleles, a function f can be defined relating the space of genotypes G_i to the space of alleles $A_i = \{A, B\}$ as $f : G_i \rightarrow A_i$. In this case, the contingency table of the observed allelic counts in the population of study is obtained (Table 2):

Table 2. Contingency table of observed allelic counts.

	A	B	Total
Cases	$2n_{hom.ref.a} + n_{het.a}$	$n_{het.a} + 2n_{hom.alt.a}$	$2N_a$
Controls	$2n_{hom.ref.o} + n_{het.o}$	$n_{het.o} + 2n_{hom.alt.o}$	$2N_o$
Total	$2n_{hom.ref} + n_{het}$	$n_{het} + 2n_{hom.alt}$	$2N$

Particularly, each variant V_i from the population can be defined as a simple random variable $V_i : \Omega \rightarrow A_i$, so that $\forall a \in A_i \exists \omega \in \Omega$, which means that $V_i(\omega) = a$, with Ω as the space of events. Therefore, a probability function can be defined by $p_i : \{a_i \in A_i\} \rightarrow [0, 1]$, where $p_i = P(V_i = a_i)$. Thus, the expected allele counts $E(V_i = a_i) = \sum a_i p_i$ are expressed as (Table 3):

Table 3. Contingency table of expected allelic counts.

	A	B
Cases	$\frac{2N_a(2n_{hom.ref} + n_{het})}{2N}$	$\frac{2N_a(n_{het} + 2n_{hom.alt})}{2N}$
Controls	$\frac{2N_o(2n_{hom.ref} + n_{het})}{2N}$	$\frac{2N_o(n_{het} + 2n_{hom.alt})}{2N}$

Under the assumption of independence of observing allele A or allele B in the study population, a Fisher’s exact test can be applied to these contingency tables to test for differences between the allelic frequencies in each group.

Moreover, if the sample size is large enough ($N > 20$) and under the assumption of independence, a chi-squared test can be performed instead to check for differences between the observed frequencies ($Observed = \frac{N \cdot observations}{N \cdot total}$) and expected frequencies (which derived from Table 3, $Expected = \frac{N \cdot expected \ counts}{N \cdot total}$):

$$\sum \frac{(Observed - Expected)^2}{Expected} \sim \chi^2_1.$$

To calculate the odds ratio *OR*, Table 3 can be simplified and annotated as (Table 4):

Table 4. Simplified contingency table of expected allelic counts.

	A	B
Cases	n_{Aa}	n_{Ba}
Controls	n_{Ao}	n_{Bo}

As a result, from Table 4, the odds ratio can be expressed as $OR = \frac{n_{Ba}/n_{Bo}}{n_{Aa}/n_{Ao}} = \frac{n_{Ba}n_{Ao}}{n_{Aa}n_{Bo}}$.

Given that the additive model is the most common in GWAS, the methodology described above, which is based on the contingency tables, has been formulated under this model. For each individual *j* in the population, the space for the genotypes of each variant *V_{ij}* was defined as $G_{ij} = \{AA, AB, BB\}$. For the additive model, this space is encoded by defining a function $f : G_{ij} \rightarrow \{0, 1, 2\}$, where $f(AA) = 0$, $f(AB) = 1$, $f(BB) = 2$. Nonetheless, depending on the encoding of the different inheritance models, this function *f* takes different values: $\{0, 1, 1\}$ for the dominant ($f(AA) = 0$, $f(AB) = 1$, $f(BB) = 1$), $\{0, 0, 1\}$ for recessive ($f(AA) = 0$, $f(AB) = 0$, $f(BB) = 1$), or $\{0, 1, 0\}$ for heterodominant ($f(AA) = 0$, $f(AB) = 1$, $f(BB) = 0$). As a result, Table 1 can be reconstructed for the non-additive models, as shown in Table 5:

Table 5. Contingency table of observed genotypes for the different genetic models.

	Dominant Model (0,1,1)		Recessive Model (1,0,0)		Heterodominant Model (0,1,0)		Total
	AA	AB + BB	AA + AB	BB	AA + BB	AB	
Cases	$n_{hom.ref.a}$	$n_{het.o} + n_{hom.alt.o}$	$n_{hom.ref.a} + n_{het.o}$	$n_{hom.alt.o}$	$n_{hom.ref.o} + n_{hom.alt.o}$	$n_{het.o}$	N_a
Controls	$n_{hom.ref.o}$	$n_{het.o} + n_{hom.alt.o}$	$n_{hom.ref.o} + n_{het.o}$	$n_{hom.alt.o}$	$n_{hom.ref.o} + n_{hom.alt.o}$	$n_{het.o}$	N_o
Total	$n_{hom.ref}$	$n_{het} + n_{hom.alt}$	$n_{hom.ref} + n_{het}$	$n_{hom.alt}$	$n_{hom.ref} + n_{hom.alt}$	n_{het}	N

Moreover, this encoding can be applied to further study the different genetic models in each of the approaches that will be detailed in the following subsections.

Contingency tables were particularly successful in the first GWAS, leading to the identification of novel associations to complex disease [21,22]. Therefore, some common bioinformatic tools still include options to perform the chi-squared test for association [23]. However, one important issue that is not covered by the contingency table analyses is the fact that the thousands or millions of individuals in a GWAS can share some potentially confounding qualities, apart from the trait of interest, such as age or sex. The effects of these known covariates need to be corrected in order to avoid the concealment of the genomic associations to disease risk or the emergence of spurious associations.

3.2.2. Logistic Regression

Logistic regression models are broadly used for the study of GWAS to analyse the explainability of the phenotype in terms of the genotype. Particularly, the study of association under this model facilitates the simultaneous analysis of multiple variables, thus allowing the study of covariates in addition to genomic variants.

Therefore, a logistic regression model can be formulated based on the analysis of a population with *N* individuals, where *N_a* are cases (diseased) and *N_o* are controls (non-diseased) and where $N = N_a + N_o$. For each individual *j* in the population of study, the phenotype takes the values $P_j \in \{0, 1\} = \{control, case\} = \{diseased, non - diseased\}$. Thus, the study of an individual *j* being diseased can be modelled by a Bernoulli distribution $P_j \sim B(p_j)$, where *p_j* is the unknown probability of an individual having the disease. As a result, the phenotype of the *N* individuals of the population can be modelled by a binomial distribution $P \sim Bin(p_j, n_j)$. Particularly, based on the observation of $m \in \{1, \dots, M | M < \infty\}$ genomic variants *V_i*, $i = 1, \dots, m$, where their genotype can take a value from the space $G_{ij} = \{AA, AB, BB\} = \{hom.ref, het, hom.alt\}$, the probability

of an individual being diseased can be explained by the genotype as $p_j = E\left(\frac{P}{n_j} \mid G_{ij}\right)$. Consequently, the ratio of the probability of individual j having the disease or not, given a particular genotype, is expressed as $\frac{p_j}{1-p_j}$.

Therefore, a *logit* function transformation can be applied to this ratio

$$\text{logit}(p_j) = \ln\left(\frac{p_j}{1-p_j}\right), \tag{1}$$

thus fitting the logistic regression model for each variant

$$\text{logit}(p_j) \sim \beta_0 + \beta_1 G_{ij}. \tag{2}$$

From this logistic regression model, beta coefficients $\beta_i, i = \{0, 1\}$ are estimated, for example, by applying the maximum likelihood or least squares approaches.

The genotype effect on disease risk is then measured by the odds ratio, which can be calculated as

$$OR = \exp(\beta_1). \tag{3}$$

Finally, the association of the genotype with the disease is determined by testing the hypothesis of $\beta_1 \neq 0$.

One of the advantages of the logistic regression model in GWAS analysis is the possibility of including covariate effects. To this end, the model can be extended so that the expected phenotype for individual j with genotype G_{ij} can be conditioned on t additional covariates X_{kj} with $k = 1, \dots, t, t < \infty$, so that

$$p_j = E\left(\frac{P}{n_j} \mid G_{ij}, X_{1j}, X_{2j}, \dots, X_{tj}\right).$$

Correspondingly, the logistic regression model

$$\text{logit}(p_j) \sim \beta_0 + \beta_1 G_{ij} + \beta_2 X_{1j} + \dots + \beta_{t+1} X_{tj}$$

can be used to estimate the betas, which can then be tested for associations individually ($\beta_1, \beta_2, \dots, \beta_m \neq 0, m = 1, \dots, t + 1$). In this case, the significant β_k coefficients can be considered as measures of the genotype and covariate effects, and the *OR* for each of them can be calculated as previously detailed in Equation (3). By including possible confounding effects as covariates in the logistic regression model, a more precise estimate of the genotype effect on disease and thus a more robust association result can be obtained.

Due to their power and flexibility, logistic regression models have been the most used approach in GWAS for complex diseases, leading to the discovery of novel loci and broadening the genetic and biological understanding of a variety of diseases [24,25]. In line with this success, many bioinformatic tools for logistic regression modelling and association have been developed [23,26–28].

3.2.3. Further Extensions and Developments of Regression Models in GWAS

All of the strategies presented in the previous sections were designed to work with binary phenotypes such as diseases. However, regression models can also be easily applied to the study of quantitative traits [29]. In this case, in a study of a population with N individuals, for each individual j , the phenotype takes the values $P_j \in \sigma(\mathbb{R})$ with $\sigma(\mathbb{R})$ the Borel set. Thus, the study of the individual's phenotype P_j can be performed using a linear regression model based on the genotype of $m \in \{1, \dots, M \mid M < \infty\}$ genomic variants $V_i, i = 1, \dots, m$, where each variant genotype can take a value from the space $G_{ij} = \{AA, AB, BB\} = \{hom.ref, het, hom.alt\}$. Therefore, the linear regression model is expressed as

$$P_j \sim \beta_0 + \beta_1 G_{ij} \tag{4}$$

and the betas β_i are the parameters of the model. Particularly, the genotype effect on the risk of disease is measured by the beta $\beta = \beta_1$. Then, a hypothesis test for association is used to check whether the genotype is associated with the trait $\beta_1 \neq 0$.

Overall, the regression methods for GWAS can be extended with a generalized linear model (GLM) [30]. If the trait is quantitative and if the assumptions of genotype independence, homoscedasticity, and normality of residuals hold, then a simple linear model can be fitted. If the trait is binary, under the same assumptions, a logit transformation can be applied, and a logistic regression model can then be fitted. When the assumptions are violated, different types of models can be derived, such as Poisson regression or ANOVA methods.

As a further extension of regression methods, mixed models have recently started to be applied in GWAS. Mixed models take their name from the regression of both fixed and random effects on the outcome variable. In GWAS, genotypes and non-genetic covariates are fitted as fixed effects, together with a genetic relationship matrix (GRM), which are fitted as a random effect. The GRM carries information on the genetic relatedness between the individuals of the study; mixed models therefore correct for genetic correlations between individuals, which are a major source of confounding in association. This way, the need for excluding related individuals from a GWAS is overcome, thus increasing the discovery power [31]. Similar to GLMs, mixed models can also be applied to quantitative or binary phenotypes, and tools for linear or logistic mixed models have been developed accordingly [32–34]. Mixed models have proven to be particularly suitable for GWAS in large biobanks [31,34–36].

In conclusion, regression models showed a considerable ability to accommodate different hypotheses in terms of covariates and genetic models, producing powerful and robust results. For these reasons, regression approaches are currently the method of choice in GWAS.

3.2.4. Bayesian Statistics

GWAS Bayesian approaches were developed in parallel to GWAS regression models as an attempt to refine and improve their results, increasing their discovery power.

Thus, based on the study of a population with N individuals, where N_a are cases (diseased) and N_o are controls (non-diseased) and where $N = N_a + N_o$, for each individual j in the population of study, the phenotype takes values $P_j \in \{0, 1\} = \{control, case\} = \{diseased, non - diseased\}$ for binary traits, or $P_j \in \sigma(\mathbb{R})$, with $\sigma(\mathbb{R})$ the Borel set, for qualitative traits.

Under these scenarios, the logistic and linear regression models can be constructed as they are in Equations (2) and (4), respectively. Then, Bayesian results are provided in the form of the posterior probabilities of regression estimates:

$$P(\beta_{1j} | G_{ij}) \propto P(G_{ij} | \beta_{1j})P(\beta_{1j}) \tag{5}$$

where $P(G_{ij} | \beta_{1j})$ is obtained from the regression model (e.g., the likelihood of observing a particular phenotype $L(Y_j | \beta_0, \beta_1)$) and where the prior $P(\beta_{1j})$ can be estimated based on β_{1j} inference approaches, such as the Jacobian transformation, normal approximation or uniform distributions. These calculated posterior probabilities can be used as priors to fit a regression model again. Therefore, the β_{ij} coefficients (thus the genotype effect on disease) will be better estimated, reducing the proportion of false-positive results [37,38].

Moreover, Bayesian methods can also be applied to reduce the dimensionality of a GWAS. Dimensionality reductions are based on the assumption that the number of variants with a non-zero effect p tends to be far smaller than the total number of analysed variants k ($k \gg p$). With Bayesian approaches, the initial set of variants (V_1, \dots, V_m) , $m \in \{1, \dots, M | M < \infty\}$ is reduced to those with a higher probability of escaping the zero

effect, relying on the posterior probability (5). A vector γ is constructed by applying the indicator of the non-zero effect to each variant:

$$\gamma = (V_1, \dots, V_m) 1_{P(\beta_{ij}|G_{ij}) \neq 0} \text{ where } 1_{P(\beta_{ij}|G_{ij}) \neq 0} = \begin{cases} 1, & P(\beta_{ij}|G_{ij}) \neq 0 \\ 0, & P(\beta_{ij}|G_{ij}) = 0 \end{cases}$$

Therefore, under the binary trait scenario, which corresponds to the logistic regression model, the probability of an individual being diseased can be explained by the genotype as $p_j = E\left(\frac{p}{n_j} \mid G_{ij}(\gamma)\right)$. Thus, the ratio between the probability of individual j having the disease or not given a particular genotype will be expressed under the model $\text{logit}(p_j) \sim \beta_0 + \beta_1 G_{ij}(\gamma)$. Similarly, under the quantitative trait scenario, which corresponds to the linear regression model, the explanation of the individual phenotype based on its genotype is expressed by the model $P_j \sim \beta_0 + \beta_1 G_{ij}(\gamma)$. Last, a regression model is fitted to obtain the betas, which are tested to check whether the genotype is associated with the disease [39–42]. As a result of reducing the number of simultaneously performed tests, the multiple-testing correction burden is also reduced, thus increasing the detection power (Section 3.3).

Bayesian statistical methods have proven the relevance of reducing the number of tests to improve the results that can be obtained from GWAS [43,44]. Therefore, many bioinformatic tools have been developed and have been updated to facilitate the association analysis based on Bayesian models [28,32].

3.3. Statistical Interpretation of GWAS Results

As it is common in statistical analyses, a significance threshold is required to decide on the significance of the obtained results. This level of significance is measured with a p -value threshold, typically 0.05 or 0.01 for a 5% and 1% probability of rejecting the null hypothesis when it is true (false positive), respectively. However, in a GWAS, huge numbers of tests are performed (one for each genomic variant, usually in the order of millions). Therefore, multiple testing correction with an adjusted p -value threshold is needed to determine statistical significance.

For this purpose, the use of standard Bonferroni’s multiple-testing correction, which consists in dividing the p -value threshold by the total number of tests, could be suggested. However, this would assume full statistical independence between all of the performed tests. Given that genomic variants are not independent of each other, due to linkage disequilibrium (LD) as previously described, the resulting threshold would then be exceedingly stringent. Instead, GWAS typically assume that there are a million truly independent genomic loci, as was estimated in the European population [45]. With this assumption, the Bonferroni correction results in a p -value threshold [46] of

$$p = \frac{0.05}{1,000,000} = 5 \times 10^{-8}$$

which is the most commonly used threshold to accept or reject a GWAS association. This threshold is referred to as the genome-wide significance threshold.

The unconditional (absolute) validity of this estimation has however been questioned, and thus, the search for an adequate p -value threshold to use in GWAS has grown into a parallel subject of study. For instance, multiple additional statistical procedures have been proposed, such as the Sidak correction, False Discovery Rate (FDR), permutation test, Bayesian approaches, and dimensionality reduction-based methods.

The representation of the GWAS results presents a different challenge. In order to represent the millions of statistical results in a visual manner, the association p -values are typically displayed in a Manhattan plot (Figure 3). In this type of scatter plot, each genomic variant that has been tested for association is represented as a point, the X axis comprises all of the genomic positions, and the Y axis measures the obtained p -values, which are typically scaled in $-\log_{10}$. The significance threshold (e.g., 5×10^{-8}) is marked with a horizontal line so that the results that are significant after multiple testing correction can be

easily spotted. The name of these plots derives from the expectation that the results would look similar to the skyline of Manhattan, with significant loci rising as skyscrapers from the ground. In the reality of GWAS, however, these rich skylines are seldom obtained, as it is more common to observe only a handful of loci that reach such levels of significance.

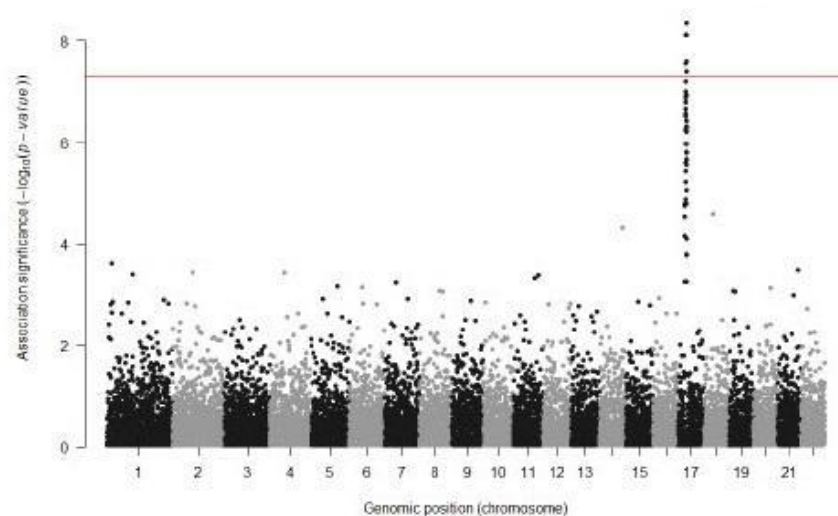


Figure 3. Example of a Manhattan plot. The X axis shows all of the tested variants by their genomic location, and the Y axis shows the strength of the statistical association. The significance threshold (red line) has been increased to correct for multiple GWAS analyses in the study.

In addition to identifying significant associations between genomic variants and phenotypes, GWAS also estimate the odds ratio (OR) for each genomic locus, an effect size estimate of the increased odds of having the disease per risk allele count [47]. An $OR = 1$ thus implies no association with the disease, an $OR > 1$ implies that the effect allele is a risk allele, increasing the risk of developing the disease, and an $OR < 1$ implies a protective allele, decreasing the risk of disease. In the case of quantitative traits, which require no logarithm transformation, the magnitude of the effect can be directly measured using the β coefficient of the regression. Thus, $\beta = 0$ implies no association with the trait, but $\beta > 0$ and $\beta < 0$ imply a positive or negative association with the allele, respectively.

Unfortunately, effect sizes tend to be overestimated, which is mainly due to the bias caused by an effect named the winner's curse. The quantification, correction, and bias-reduction on the effect size estimator has been a GWAS-parallel subject of study [48] given its relevance to the heritability contribution.

Box 3. Technical concepts.

- Beta: An estimation of the effect size of a variant for a quantitative phenotype: the coefficients obtained from fitting a regression of the genotypes to the phenotype.
- Cohort: A group of individuals.
- DNA hybridisation array: A technology to identify the genotypes of a specific subset of variants of an individual.
- Effect size: A measure of the contribution of a genomic variant to a specific phenotype.
- Imputation: A statistical method to infer missing genotypes given a reduced set of known genotypes and a reference panel.
- Odds Ratio: An estimation of the effect size of a variant for a binary phenotype: the odds of having the disease with a variant divided by the odds of having the disease without it.
- Phasing: A statistical method to infer the haplotypes of an individual to determine which alleles belong to the same chromosomal sequence.
- Reference panel: A set of well characterised haplotypes of a group of individuals, used as a reference to infer non-genotyped variants in other individuals.
- Whole genome sequencing: A technology that provides the complete nucleotide sequence of an individual genome.

4. Current Practice and GWAS Limitations

GWAS have had a history of success in the study of complex traits, enabling the identification of the genomic loci involved in these phenotypes for the first time. Indeed, GWAS have so far discovered more than 276 thousand genomic associations for more than 4 thousand traits and diseases [49–51]. However, almost 20 years of analyses have also highlighted their limitations, which preclude more genomic associations from being identified [21,22]. Here, we discuss the main critical points of GWAS in detail, and we explain how the methodology can be extended to mitigate some of these. Next, we describe the most common complementary approaches and the existing alternatives that are attempting to solve these limitations.

4.1. Power and Sample Size

One of the main concerns in a GWAS is whether the study is powered enough to detect any association with a trait. The statistical power of association for a given variant strongly depends on the magnitude of its effect size and on its frequency in the population. Strong effect sizes are easier to capture, and common variants generally provide higher power. However, due to evolutionary selective pressures, effect sizes and frequencies are generally inversely correlated, with rarer alleles showing stronger ORs. In practical terms, current GWAS have mostly revealed associations for common variants with ORs of around 1.05–1.3 [52].

A natural way to increase power in GWAS is to increase the size of the sample under study (N). Increasing sample size would allow the identification of smaller effects for common variants as well as open the possibility to study rare variants. Motivated by this need, large-scale initiatives have been established in the form of international consortia to pool multiple resources and thus generate larger cohorts for subsequent analyses. These efforts have pushed the discovery of new loci and our understanding of complex disease genetics [53–56]. Further, biobanks have been established to make these large collections of genotypic and phenotypic data available for future studies [57–59]. However, given the sensible nature of these genomic and medical data, accessibility restrictions have been put in place, which often hinder or discourage their reutilisation by further scientific efforts.

Another commonly used strategy to increase sample size in GWAS is meta-analysis based on the statistical combination of previous GWAS results from different studies on the same phenotype. Requiring only GWAS summary statistics (e.g., sample size, effect sizes and p -values), meta-analyses are far more cost-effective than the generation of new genotype–phenotype datasets and thus have been used extensively [13,60,61].

Meta-analysis approaches are based on a weighted sum of the effects obtained in each of the studies, thus providing an estimate of the association of each genetic marker over

all of them. For example, in a meta-analysis for M studies where each variant V_i has been assigned an effect β_{ij} for the j -th study, a Stouffer's Z-score can be calculated by assigning a weight for the estimated allelic effect on each study w_{ij} so that the allelic effect across all the studies will be

$$Z_i = \frac{\sum_{j=1}^M \beta_{ij} w_{ij}}{\sqrt{\sum_{j=1}^M w_{ij}^2}} \sim \chi_1^2$$

which estimates the association to disease over all tests.

In addition, the genetic heterogeneity between the different studies is measured, which is based on Cochran's Q -test, by the statistic

$$Q_i = \sum_{j=1}^M w_{ij} (Z_i - \beta_{ij})^2 \sim \chi_{M-1}^2$$

for each SNV i . This measure helps to detect associations that are not consistent across the studies, which might then be filtered out if necessary.

Despite the proven value in increasing power, large sample sizes in GWAS present many challenges, nonetheless. The recruitment and genotyping of individuals might be extremely expensive in terms of time and resources. Despite having received more attention in recent years, data sharing is still limited and difficult, even in the form of summary statistics. Further, recent studies have estimated that unprecedented sample sizes, in the order of millions, might be needed to capture the entire spectrum of the variants associated with a trait [62]. Different strategies other than simply increasing the number of analysed samples might be thus more feasible to increase discovery power and will be briefly discussed in the following sections.

4.2. Increasing the Number of Genomic Variants

Another important factor in determining the discovery power is the correlation (LD) existing between the interrogated variants and the real, underlying causal variant [47]. Higher discovery power can be achieved by increasing the number of tested variants, thus obtaining a higher density coverage of the genome and increasing the probability of directly testing variants that are strongly correlated with the causal ones. However, as described in Section 2, GWAS typically use DNA microarray technologies, which only provide the genotypes for a limited subset (0.5 to 2 M) of all of the SNVs in a genome [63].

A technique that is commonly used to increase the number of variants that can be tested in a GWAS is genomic imputation. Starting from genotyping array data, genotypes of over 10 million variants can be inferred for an entire group of individuals (also named cohort) [64], with a reduced number of missing values [65,66].

Imputation is usually preceded by a phasing step, in which haplotypes for each individual are inferred starting from genotypes, typically from array data. Then, the studied haplotypes are statistically compared with those in reference panels, which are panels of thousands of individuals with a deeply characterised haplotype [15,67–71]. Through this comparison, the genotype probabilities for variants in the reference panels are imputed into the cohort haplotypes [72]. Several methods and tools have been developed to phase and impute [65,73–76]. Most of them are essentially based on Markov Chains (MC), Hidden Markov Models (HMM), Markov Chain Monte Carlo (MCMC), and the expectation-maximisation algorithm [28,77]. Other tools have also been developed to combine the imputation results from different panels [14].

As a result, given a population with N individuals, where, $m \in \{1, \dots, M | M < \infty\}$ variants $V_i, i = 1, \dots, m$, are inspected for each individual j , each variant genotype can take a value from the space of genotypes $G_{ij} = \{AA, AB, BB\} = \{hom.ref, het, hom.alt\}$. Based on the space defined by the genotype, each genomic variant V_i can be considered as a simple random variable $V_i: \Omega \rightarrow G_{ij}$, so that $\forall g \in G_{ij} \exists \omega \in \Omega$ for which $V_i(\omega) = g$, with Ω as the space of events. Under this scenario, the imputation model can be formalized

by first stating that each variant genotype G_{ij} for the individual j has a corresponding haplotype $H_{ij} = \{(0,0), (0,1), (1,0), (1,1)\}$, which is defined by a function $f : G_{ij} \rightarrow H_{ij}$, where $f(AA) = (0,0)$, $f(AB) = \{(0,1), (1,0)\}$, $f(BB) = (1,1)$. Thus, the haplotype space H_{ij} is a partition of the genotype space G_{ij} . For simplicity, each haplotype H can be written as a pair set $H = (H_{ij}^{(1)}, H_{ij}^{(2)})$, $H_{ij}^{(k)} \in \{0,1\}$, $k \in \{1,2\}$. The aim of imputation is to infer the missing genotypes based on the posterior probability $P(G_{ij}|H)$ for each individual in a LD region by comparing the individual haplotypes in that region with the N haplotypes $H = \{H_1, H_2, \dots, H_N\}$ present in a reference panel (Figure 4).

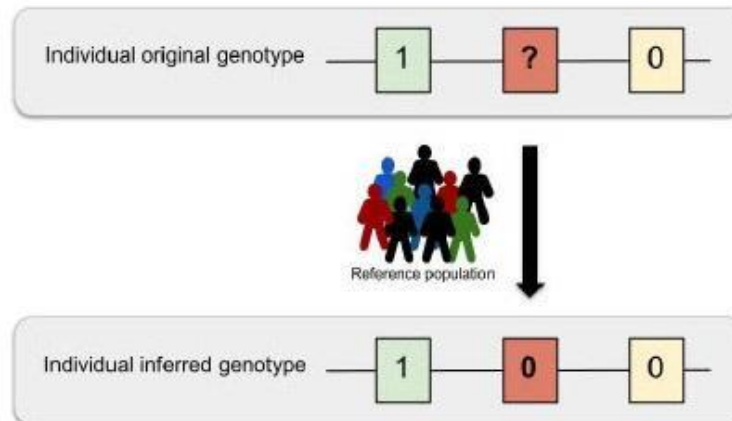


Figure 4. Imputation schema. The genotypes originating from DNA hybridisation arrays only provide information on a limited set of genomic variants (0.5 to 2 million sites). These missing variant genotypes can be statistically inferred by using one or multiple reference haplotype panels in a process named genomic imputation.

For example, in Hidden Markov Model (HMM) approaches, the posterior probability of each genotype, given the haplotype, can be calculated as

$$P(G_{ij}|H) = \sum_{H_{ij}^{(1)}, H_{ij}^{(2)}} P(G_{ij}|H_{ij}^{(1)}, H_{ij}^{(2)}, H) P(H_{ij}^{(1)}, H_{ij}^{(2)}|H) \tag{6}$$

where the term $P(H_{ij}^{(1)}, H_{ij}^{(2)}|H)$ is the prior probability for each hidden state change along the sequence, and $P(G_{ij}|H_{ij}^{(1)}, H_{ij}^{(2)}, H)$ models the probability that the genotype will be similar to the haplotypes that are copied from the reference. By estimating the genomic recombination rate across the region ρ based on the effective population size and the mutation rate θ , Equation (6) can be simplified to

$$P(G_{ij}|H, \theta, \rho) = P(G_{ij}|H_{ij}^{(1)}, H_{ij}^{(2)}, \theta) P(H_{ij}^{(1)}, H_{ij}^{(2)}|H, \rho) \tag{64}$$

Given that both θ and ρ can be estimated from the population of study and that the haplotypes can be inferred from the HMM, this model can be used to infer missing genotypes in the study population.

The accuracy of the different imputation methods can be assessed by masking known genotypes and imputing them using surrounding variants. The correlation between the estimations and the true values can be used to measure the imputation accuracy. Based on this method, current error rates range between 5.10 to 6.33% [28].

Genotype imputation offered the possibility of comprehensively investigating variants throughout the genome, including rare variants, at a large scale for the first time. However,

the imputation of rare variants still presents difficulties. Although rare variants are present in reference panels, those are usually in low LD with the common variants from the genotyping array; therefore, they are imputed with less accuracy. Further, rare variants tend to be more private, and only a fraction of these can be possibly present in reference panels; thus, only a few can be imputed. In the future, when whole genome sequencing is affordable for large studies, the imputation process will cease to be necessary since all of the genomic variants will be obtained from the DNA of the participants. However, until then, genotype imputation provides the most valid alternative for comprehensive GWAS.

4.3. Genetic and Population Heterogeneity

Genetic heterogeneity between individuals of shared ancestry or between those of different ancestries is a factor that further complicates the study of polygenic traits. The same apparent phenotype (especially diseases) might be the result of different combinations of genomic variants in different individuals. Genetic heterogeneity is typically overlooked in GWAS, as individuals with the same broad disease are considered as a homogeneous group of cases. In this scenario, GWAS can only capture the most shared signals, and less prevalent genomic associations might be masked.

An attempt to reduce this issue has been made by classifying cases into sub-groups by using multiple clinical variables or by defining sub- or endo-phenotypes. For example, a disease such as Type 2 diabetes is broadly defined by a high content of glucose in the blood, but different clinical sub-types have recently been identified using measures such as age of disease onset or body-mass index [78]. The rationale is that these phenotypic sub-groups might reflect more genetically homogeneous groups and may thus help us to identify the underlying genomic loci that differentiate them. Even though this strategy entails a decrease in the dimensional reduction of the sample size due to fragmentation, the power to discover the underlying genomic factors could be increased due to a reduction in the dilution of the relevant signals as a consequence of the homogeneity and less variability in the data [79].

Genetic heterogeneity is also significant between individuals of different ancestral backgrounds due to differences in variant frequencies (e.g., a rare variant in one ancestry might be common in another) and LD patterns. Early GWAS were performed with individuals of predominantly European or Caucasian ancestry, which raised the question of their relevance for individuals of other ancestries. Moreover, the possibility remained that common variants were only associated with complex diseases because they were in LD with rare, high-impact variants that were specific to the studied ancestry and thus that these associations would not replicate in other ancestries.

Since then, trans-ancestry (also named trans-ethnic) studies, which analyse samples of multiple ancestries together, have shown that the variants that were associated with the complex traits and diseases that were identified in these studies were predominantly consistent with those identified in ancestry-specific studies [80–82]. These findings suggest that these phenotypes are indeed driven by common variants and that their genetic architecture is mostly shared across different ancestries.

Albeit burdened with further increased sample collection and analytical complexities, these large studies have succeeded in the development of population genomics and have increased the genetic understanding of complex traits [82,83].

4.4. Complex Interactions

GWAS are typically applied to capture the effect of single independent variants on a phenotype. However, complex traits are understood to be caused by multiple genomic variants that interact with environmental variables [84,85]. Therefore, other analytical frameworks are needed to interrogate more complex interactions, such as gene-gene interactions (GxG) or gene-environment interactions (GxE) [86]. Given the computational and data acquisition challenges of these studies, these have only recently become feasible,

thus providing a novel avenue to reveal new understanding of the aetiology of complex traits and diseases.

4.4.1. Gene–Gene Interactions (GxG) and Genomic Variant Epistasis

Complex phenotypes arise due to the combined effects of multiple genes. For example, 16 different genes have so far been linked to the determination of the eye colour phenotype [87]. In some cases, the effects on the phenotype of one of the genes might be enhanced, diminished, or changed by variability in a different but interacting gene. These effects are known as gene–gene (GxG) interactions. Particularly, the term epistasis can be used to describe the result of the interaction of multiple genomic variants in different loci when it is not just a linear combination of the individual gene effects.

Variation interaction models present a framework to analyse the combined effect of multiple genomic loci on complex traits. These focus on finding groups of interacting variants and compute the relative contribution of these subsets of variants to the total phenotypic variability [88–90]. However, the combinatorial nature of the problem leads to very computationally expensive analyses, given the large number of genomic variants in a genome. For example, hundreds of billions of tests will need to be performed just to inspect the association for pairwise combinations of 500,000 SNVs [84]. Further, additional measures need to be applied to solve issues such as the power needed to detect epistasis [84] or to scale the problem to a higher order interaction of genetic factors [88].

GxG interaction analysis can be extended from the methods proposed in Section 3.2. For example, in the case of a logistic regression model, in a population with N individuals, for each individual j , the phenotype takes the values $P_j \in \{0, 1\} = \{\text{control}, \text{case}\} = \{\text{diseased}, \text{non-diseased}\}$ and follows a Bernoulli distribution $P_j \sim B(p_j)$, where p_j is the unknown probability of an individual being diseased. Thus, the phenotype of the individuals of the population follows a binomial distribution $P \sim \text{Bin}(p_j, n_j)$. Based on the observation of the $m \in \{1, \dots, M | M < \infty\}$ genomic variants V_i , $i = 1, \dots, m$, where the variants genotype can take a value from the space $G_{ij} = \{AA, AB, BB\} = \{\text{hom.ref}, \text{het}, \text{hom.alt}\}$, the probability of an individual being diseased given their genotype can be expressed as $p_j = E\left(\frac{P}{n_j} \mid G_{ij}\right)$. Thus, for a pair of variants $G_{ij,1}, G_{ij,2}$, this probability becomes $p_j = E\left(\frac{P}{n_j} \mid G_{ij,1}, G_{ij,2}\right)$. Under this scenario, the *logit* function can be applied to the ratio between the probability of the individual j having the disease or not given a pair of genotypes (1). As such, the logistic regression model for the main effects can be expressed as

$$\text{logit}(Y_j) \sim \beta_0 + \beta_1 G_{ij,1} + \beta_2 G_{ij,2}$$

to test whether the genotype is associated with the disease. As a result of that, the logistic regression model with main effects and pairwise interactions can be formulated [91] as

$$\text{logit}(Y_j) \sim \beta_0 + \beta_1 G_{ij,1} + \beta_2 G_{ij,2} + \beta_3 G_{i,1} G_{ij,2}.$$

More recently, this problem has also been approached using machine learning methods, where the relationship between multiple variants and disease risk can be evaluated at once [88,92]. Several machine learning algorithms are commonly applied for solving classification, regression, or ranking problems, such as support vector machines, stochastic gradient descent, nearest neighbours, naive Bayes, Gaussian processes, neural networks, or decision trees. These methods can be applied within a supervised learning framework to find a list of variants with an effect on the disease and their combined effects. However, while these approaches have opened a new avenue for GxG analysis, they also suffer from problematic computational costs.

To work around this limitation, most studies have been forced to reduce the dimension of their input set, which is generally accomplished using multifactor-dimensionality reduction [93–95] or Bayesian inference [96,97]. Therefore, to facilitate the integration of multi-dimensionality reduction in GxG analysis, some bioinformatic tools have integrated

this methodology in their software [23,98]. In addition, most studies also resort to restricting the genomic variants to test a selected subset of candidates based on prior biological knowledge, with the hypothesis that these are more likely to provide relevant biological insights. As a result, GxG and epistatic studies are generally limited in size and scope. This field remains open, and it is likely to provide further insights on the genomics of complex traits.

4.4.2. Gene-Environment Interactions (GxE)

The effect on complex phenotypes resulting from the environment (defined as all the non-genomic components) is often overlooked, but it plays a significant role in determining both the strength and the variability of a trait or disease. For example, even if type 2 diabetes is understood to have genomic causes, one of the best clinical predictors for risk is simply age, which is independent from the genomic components of the disease. However, the effects of environmental variables on an individual also can depend on their particular genomic background, e.g., the same food consumed by two individuals might have a different impact on their weight. This effect called named gene-environment (GxE) interaction.

Specifically, GxE interaction analyses focus on studying the environmental factors, such as diet, lifestyle, psychosocial stress, or airborne agents, and their relation with different genotype groups in terms of disease associations [99,100]. In an extension of the GWAS concept, Environment-Wide Association Studies (EWAS) analyse multiple environmental factors and compare them between different genotype subgroups of a complex disease in large-scale GxE multi-studies [101]. The most common approaches to study these GxE interactions are regression-based methods (Section 3.2), which are usually preceded by a filtering step [102–104].

Thanks to these studies, the genotype group information can be used to build better prognostic models and to identify possible high-penetrance or high-exposure subgroups to build better treatments [99,105]. However, much larger sample sizes are needed for the detection of interactions compared to marginal effect sample sizes. In addition, the complexity of measuring the environmental exposure, the difficulty of incorporating environmental measures to the models, the heterogeneity of the environmental exposures, and the lack of publicly available data represent important hurdles that limit the advancement of this field of study [99,100,105–107].

4.5. Biological Interpretation and Clinical Implications

GWAS have been successful in identifying multiple loci that are associated with complex traits. However, the biological interpretation and clinical application of these findings has proven to be very challenging.

First, because of linkage disequilibrium, GWAS can only provide associated genomic loci, encompassing multiple correlated variants. In addition, GWAS identify statistical associations, but it is well established that association does not imply causation. To attempt to overcome these limitations, further computational and experimental studies need to be pursued. Computational approaches include gene expression studies and enrichment analyses of gene, pathway, epigenomic, and regulatory elements or Mendelian randomisation analyses, which are used to gain further biological insights [108,109]. Simultaneously, wet-lab experiments with cell lines, model organisms, or further human studies also need to be used to answer the biological hypotheses that are inferred from these analyses.

As an attempt to produce some clinical insight directly from GWAS results, Polygenic Risk Scores (PRS) have recently been developed. PRS are based on the premise of evaluating the total risk of disease of a genome by considering all of its genomic variants with known disease associations [110].

Particularly, PRS compute the relative risk of an individual from the population of study to develop a disease. Therefore, in a study of a population with N individuals, for each individual j in the population of study, given $m \in \{1, \dots, M | M < \infty\}$ genomic

variants V_i , $i = 1, \dots, m$, where the variants genotype can take a value from the genotypes space $G_{ij} = \{AA, AB, BB\} = \{hom.ref, het, hom.alt\}$, GWAS models can be applied to estimate the effects β_i for each genotype (Section 3.2). Then, a PRS can be calculated based on the sum of the individual genotypes G_{ij} weighted by the estimated effects for that genotype β_i , resulting from the GWAS analysis [111]. Thus, each individual score S_j is calculated using the equation $S_j = \sum_{i=1}^M G_{ij}\beta_i$. As each individual j will have an associated score S_j , the score can be observed as an independent variable explaining the phenotype P of the individual. Consequently, under a similar scenario to the one explained in Section 3.2.2 for binary traits, $P \in \{0, 1\} = \{control, case\}$, with $P \sim Bin(p_j, n_j)$ and p_j being the probability of an individual being diseased. For example, the probability of an individual being diseased can be explained by the score as $p_j = E\left(\frac{P}{n_j} | S_j\right)$. Therefore, the logit can be applied to the ratio between the probability of the individual having the disease or not, given a particular score, to fit the logistic regression model $logit(p_j) \sim \beta_0 + \beta_1 S_j$. For quantitative traits, where the individual phenotype takes values $P_j \in \sigma(\mathbb{R})$, with $\sigma(\mathbb{R})$ the Borel set, a linear regression model could then be fitted to explain the phenotype based on the individuals score as $P_j \sim \beta_0 + \beta_1 S_j$.

The distribution of the scores across the population of study follows a normal distribution, in which the left tail contains the individuals with the lowest risk of developing the disease, and the right those with the highest risk (Figure 5). However, although the use of PRS has shown potential, statistically significant differences in disease risk are typically only found when comparing the individuals at the tails of the distributions (e.g., the individuals with the highest 5% of scores have a 3x higher risk of disease than those with the lowest 5% scores), thus only providing limited insights for the majority of the population.

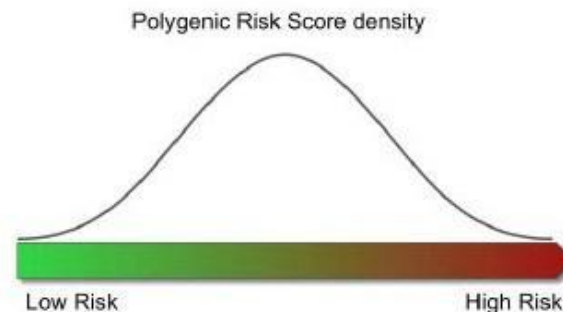


Figure 5. Relationship between risk of disease and Polygenic Risk Score. The distribution of the scores obtained from the individuals across the population follows a normal distribution. The left tail of the distribution contains the individuals with the lowest risk of developing the disease, and the right represents those with the highest risk.

Overall, the combination of cell biology studies [112,113] with GWAS results have produced a greater understanding of the biology behind complex diseases [56]. However, the study of the specific biological mechanisms that mediate the association between genotype and disease remains one of the main open fields of study in biomedicine, and the advancement of personalised medicine depends on its success.

4.6. Comprehensive GWAS Strategies for New Discoveries: An Example

As detailed in the previous sections, different strategies can be put in place to achieve good power and to produce discoveries in GWAS. Here, we describe an example of how an improved, comprehensive methodology for GWAS can reveal novel association loci in a previously analysed, publicly available cohort. In this study [14], 22 age-related diseases were analysed in 62,281 subjects from the GERA cohort. Ninety-four significant loci were

identified, of which twenty-six had never been reported before, despite the fact that the data had already been previously analysed.

A first essential feature in driving novel discovery was an extended imputation step. Imputation was performed using four reference panels yielding 16,059,686 variants to test for association. The variants encompassed a broad spectrum of frequencies and types, including 2.6 M low-frequency and 5.5 M rare variants as well as 1.6 M small insertion/deletions (indels), which are normally absent from DNA microarrays and were thus excluded from analysis. Indeed, 3 of the 26 new loci corresponded to low-frequency variants, and 7 corresponded to rare variants. Further, only a fraction of the 26 new loci would have been genome-wide significant if the imputation had been performed with only one of the individual haplotype panels.

A second feature ensuring an increased discovery power was the use of multiple inheritance models in association testing. Typical GWAS only consider the additive model, according to which disease risk is proportional to the number of risk alleles in a genotype. However, dominant, recessive, or even more complex allelic interactions are known to exist. Indeed, 20 of the 94 loci only showed genome-wide significance when non-additive tests were applied. When focusing on the novel findings, 13 out of 26 (50%) would have been missed if considering the additive model only, indicating again the strength of this approach in pushing discovery. Three of the thirteen non-additive signals corresponded to rare variants with large recessive effects (OR 4.3–19.0).

This study highlighted the value of open access and data sharing since the re-analysis using more refined and extensive methodologies led to the discovery of novel loci and disease insights. The entire GWAS strategy for this comprehensive methodology was integrated into a publicly available framework named GUIDANCE in order to facilitate further studies.

5. Conclusions

In the recent years, the increasing availability of DNA and phenotypic information and the ease of access to computational power and tools, combined with the statistical methods that we have discussed here, have greatly advanced our understanding of the genomic basis of complex traits and diseases. In this review we have presented an overview of Genome-Wide Association Studies, a broadly successful method that can be used to find associations between genomic variation and complex traits. Specially, the application of these methodologies has led to the discovery of more than 276 thousand genomic associations, for more than 4 thousand traits and diseases [49–51].

However, a significant proportion of the underlying genetic causes is still known to be missing, an effect termed missing heritability [114]. Here, we presented the main known GWAS limitations and discussed their consequences, which might partially explain this effect. The need for statistical power is forcing studies to increase the size of their samples, which comes at the expense of increasing computational and statistical challenges, which impose important limitations to these approaches [13,14,90]. However, future gene–gene, epistasis, and gene–environment interaction studies might also be able to recapitulate some of this missing heritability and provide new insights for a better understanding of the genetic basis of complex traits and diseases.

Despite providing knowledge and relevant candidate markers for diseases, an important limitation of this type of analysis is still the low applicability of the results that are obtained into clinical practice. In the case of rare diseases, variants are identified on patients with the disease to obtain an accurate diagnosis. In contrast, in the case of complex diseases, the aim is to generate maps of genetic predictors for disease risk and to apply them before the disease phenotype appears, ideally as we are born, allowing the design of preventive clinical protocols. But unfortunately, the multifactorial nature of complex diseases makes the prediction of their risk highly challenging. Current efforts include the generation of polygenic risk scores to predict risk and disease by combining multiple genetic signals identified through GWAS. It is therefore necessary to improve the methodological and

statistical frames around association studies to align with the increase of samples and with the growing computational limitations.

Similarly, the functional interpretation of associated variants to contribute to this applicability into the clinics is also challenging and has not been well resolved. Currently, the vast majority of variants that are significantly associated with a specific disease or trait through GWAS do not directly disrupt gene sequences. Rather, these are found between genes, regulating the expression of these genes [56,115,116] and not their specific function, as is often the case in rare diseases. This makes the functional interpretation of associated variants a tedious task that also requires experimental validation.

Finally, it is important to be aware that around 79% of GWAS participants are of European ancestry, despite Europeans representing only 16% of the global population [117]. As a consequence, GWAS-derived results are predictably biased; for example PRS show lower predictive accuracies in non-Europeans [82,118]. Thus, extending GWAS to under-represented ancestries, including minority groups and isolated or indigenous populations might help improve our understanding of complex diseases. Indeed, some studies have shown how African/American and Hispanic/Latino populations contribute disproportionately to GWAS discovery, providing more signals than European samples with similar sample sizes [117]. This is likely due to their genetic specificities, in terms of allele frequencies or LD patterns, which would also favour the functional interpretation and the discovery of causal variants in known loci. Several recent initiatives in this direction include the H3Africa consortium [119] or the human pangenome project [120].

Altogether, GWAS have proven to be an efficient strategy to identify the genetic factors behind complex diseases. But despite the efforts, we believe we have uncovered only the tip of the iceberg, considering the amount of different factors, including genetic variants, that are involved in the risk, offset, and progression of these complex diseases. Coordinated work across disciplines, including deep mathematical and statistical expertise, are thus required to advance and to start building clinically relevant models for disease prediction based on solid genetic architectures.

Author Contributions: L.A. drafted the manuscript. I.M. and C.S. provided guidance and revisions. L.A., I.M., C.S. and D.T. wrote the final version of the text. D.T. supervised the work. All authors have read and agreed to the published version of the manuscript.

Funding: L.A. was supported by grant BES-2017-081635. This publication is part of R&D and Innovation grant BES-2017-081635 funded by MCIN and by "FSE Investing in your future". I.M. was supported by grant FJCI-2017-31878. This publication is part of R&D and Innovation grant FJCI-2017-31878 funded by MCIN. C.S. received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement H2020-MSCA-COFUND-2016-754433.

Acknowledgments: The entire Computational Genomics group at the BSC is thanked for their helpful discussions and valuable comments on the manuscript.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

EWAS	Environment Wide Association Studies
GxE	Gene–environment interactions
GxG	Gene–gene interactions
GLM	Generalized Linear Models
GRM	Genetic Relationship Matrix
GWAS	Genome Wide Association Studies
HMM	Hidden Markov Model
LD	Linkage Disequilibrium
OR	Odds Ratio
PRS	Polygenic Risk Score
SNV	Single Nucleotide Variation

References

- Manolio, T.A.; Brooks, L.D.; Collins, F.S. A HapMap harvest of insights into the genetics of common disease. *J. Clin. Investig.* **2008**, *118*, 1590–1605. [[CrossRef](#)] [[PubMed](#)]
- Mitchell, K.J. What is complex about complex disorders? *Genome Biol.* **2012**, *13*, 237. [[CrossRef](#)]
- Robinson, M.R.; Wray, N.R.; Visscher, P.M. Explaining additional genetic variation in complex traits. *Trends Genet.* **2014**, *30*, 124. [[CrossRef](#)] [[PubMed](#)]
- Hodge, S.; Greenberg, D. How Can We Explain Very Low Odds Ratios in GWAS? I. Polygenic Models. *Hum. Hered.* **2016**, *81*, 173–180. [[CrossRef](#)]
- Mahajan, A.; Taliun, D.; Thurner, M.; Robertson, N.R.; Torres, J.M.; Rayner, N.W.; Payne, A.J.; Steinthorsdottir, V.; Scott, R.A.; Grarup, N.; et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat. Genet.* **2018**, *50*, 1505–1513. [[CrossRef](#)]
- Génin, E. Missing heritability of complex diseases: Case solved? *Hum. Genet.* **2020**, *139*, 103–113. [[CrossRef](#)] [[PubMed](#)]
- McCarthy, M.I. Genomics, Type 2 Diabetes, and Obesity. *N. Engl. J. Med.* **2010**, *363*, 2339–2350. [[CrossRef](#)] [[PubMed](#)]
- Vercelli, D. Discovering susceptibility genes for asthma and allergy. *Nat. Rev. Immunol.* **2008**, *8*, 169–182. [[CrossRef](#)]
- O'Donnell, C.J.; Nabel, E.G. Genomics of Cardiovascular Disease. *N. Engl. J. Med.* **2011**, *365*, 2098–2109. [[CrossRef](#)]
- Van Cauwenbergh, C.; Van Broeckhoven, C.; Sleegers, K. The genetic landscape of Alzheimer disease: Clinical implications and perspectives. *Genet. Med.* **2015**, *18*, 421–430. [[CrossRef](#)]
- American Diabetes Association. Economic Costs of Diabetes in the U.S. in 2017. *Diabetes Care* **2018**, *41*, 917–928. [[CrossRef](#)] [[PubMed](#)]
- Vansteelandt, S.; Goetgheuk, S.; Lutz, S.; Waldman, I.; Lyon, H.; Schadt, E.E.; Weiss, S.T.; Lange, C. On the adjustment for covariates in genetic association analysis: A novel, simple principle to infer direct causal effects. *Genet. Epidemiol.* **2009**, *33*, 394–405. [[CrossRef](#)]
- Bonàs-Guarch, S.; Guindo-Martínez, M.; Miguel-Escalada, I.; Grarup, N.; Sebastian, D.; Rodriguez-Fos, E.; Sánchez, F.; Planas-Félix, M.; Cortes-Sánchez, P.; González, S.; et al. Re-analysis of public genetic data reveals a rare X-chromosomal variant associated with type 2 diabetes. *Nat. Commun.* **2018**, *9*, 321. [[CrossRef](#)] [[PubMed](#)]
- Guindo-Martínez, M.; Amela, R.; Bonàs-Guarch, S.; Puiggròs, M.; Salvoró, C.; Miguel-Escalada, I.; Carey, C.E.; Cole, J.B.; Rüger, S.; Atkinson, E.; et al. The impact of non-additive genetic associations on age-related complex diseases. *Nat. Commun.* **2021**, *12*, 2436. [[CrossRef](#)]
- The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **2015**, *526*, 68–74. [[CrossRef](#)] [[PubMed](#)]
- International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **2001**, *409*, 860–921. [[CrossRef](#)]
- McCarthy, M.I.; Abecasis, G.R.; Cardon, L.R.; Goldstein, D.B.; Little, J.; Ioannidis, J.P.A.; Hirschhorn, J.N. Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. *Nat. Rev. Genet.* **2008**, *9*, 356–369. [[CrossRef](#)] [[PubMed](#)]
- LaFramboise, T. Single nucleotide polymorphism arrays: A decade of biological, computational and technological advances. *Nucleic Acids Res.* **2009**, *37*, 4181–4193. [[CrossRef](#)]
- Uffelmann, E.; Huang, Q.Q.; Munung, N.S.; de Vries, J.; Okada, Y.; Martin, A.R.; Martin, H.C.; Lappalainen, T.; Posthuma, D. Genome-wide association studies. *Nat. Rev. Methods Prim.* **2021**, *1*, 59. [[CrossRef](#)]
- Lander, E.S.; Schork, N.J. Genetic dissection of complex traits. *Science* **1994**, *265*, 2037–2048. [[CrossRef](#)]
- Ozaki, K.; Ohnishi, Y.; Iida, A.; Sekine, A.; Yamada, R.; Tsunoda, T.; Sato, H.; Sato, H.; Hori, M.; Nakamura, Y.; et al. Functional SNPs in the lymphotoxin- α gene that are associated with susceptibility to myocardial infarction. *Nat. Genet.* **2002**, *32*, 650–654. [[CrossRef](#)] [[PubMed](#)]
- Klein, R.J.; Zeiss, C.; Chew, E.Y.; Tsai, J.-Y.; Sackler, R.S.; Haynes, C.; Henning, A.K.; SanGiovanni, J.P.; Mane, S.M.; Mayne, S.T.; et al. Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science* **2005**, *308*, 385. [[CrossRef](#)] [[PubMed](#)]

23. Purcell, S.; Neale, B.; Todd-Brown, K.; Thomas, L.; Ferreira, M.A.R.; Bender, D.; Maller, J.; Sklar, P.; De Bakker, P.I.W.; Daly, M.J.; et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **2007**, *81*, 559–575. [[CrossRef](#)]
24. Shah, S.; Henry, A.; Roselli, C.; Lin, H.; Sveinbjörnsson, G.; Fatemifar, G.; Hedman, Å.K.; Wilk, J.B.; Morley, M.P.; Chaffin, M.D.; et al. Genome-wide association and Mendelian randomisation analysis provide insights into the pathogenesis of heart failure. *Nat. Commun.* **2020**, *11*, 163. [[CrossRef](#)]
25. van Zuydam, N.R.; Ahlqvist, E.; Sandholm, N.; Deshmukh, H.; Rayner, N.W.; Abdalla, M.; Ladenvall, C.; Ziemek, D.; Fauman, E.; Robertson, N.R.; et al. A Genome-Wide Association Study of Diabetic Kidney Disease in Subjects with Type 2 Diabetes. *Diabetes* **2018**, *67*, 1414–1427. [[CrossRef](#)]
26. Aulchenko, Y.S.; Ripke, S.; Isaacs, A.; van Duijn, C.M. GenABEL: An R library for genome-wide association analysis. *Bioinformatics* **2007**, *23*, 1294–1296. [[CrossRef](#)] [[PubMed](#)]
27. Kutalik, Z.; Johnson, T.; Bochud, M.; Mooser, V.; Vollenweider, P.; Waeber, G.; Waterworth, D.; Beckmann, J.S.; Bergmann, S. Methods for testing association between uncertain genotypes and quantitative traits. *Biostatistics* **2011**, *12*, 1–17. [[CrossRef](#)]
28. Marchini, J.; Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **2010**, *11*, 499–511. [[CrossRef](#)]
29. Yang, J.J.; Li, J.; Williams, L.K.; Buu, A. An efficient genome-wide association test for multivariate phenotypes based on the Fisher combination function. *BMC Bioinform.* **2016**, *17*, 19. [[CrossRef](#)] [[PubMed](#)]
30. Nelder, J.A.; Wedderburn, R.W.M. Generalized Linear Models. *J. R. Stat. Soc. Ser. A* **1972**, *135*, 370. [[CrossRef](#)]
31. Loh, P.-R.; Kichaev, G.; Gazal, S.; Schoech, A.P.; Price, A.L. Mixed-model association for biobank-scale datasets. *Nat. Genet.* **2018**, *50*, 906–908. [[CrossRef](#)]
32. Loh, P.-R.; Tucker, G.; Bulik-Sullivan, B.K.; Vilhjalmsón, B.J.; Finucane, H.K.; Salem, R.M.; Chasman, D.I.; Ridker, P.M.; Neale, B.M.; Berger, B.; et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **2015**, *47*, 284–290. [[CrossRef](#)]
33. Browning, B.L.; Zhou, Y.; Browning, S.R. A One-Penny Imputed Genome from Next-Generation Reference Panels. *Am. J. Hum. Genet.* **2018**, *103*, 338–348. [[CrossRef](#)] [[PubMed](#)]
34. Mbatchou, J.; Barnard, L.; Backman, J.; Marcketta, A.; Kosmicki, J.A.; Ziyatdinov, A.; Benner, C.; O’Dushlaine, C.; Barber, M.; Boutkov, B.; et al. Computationally efficient whole-genome regression for quantitative and binary traits. *Nat. Genet.* **2021**, *53*, 1097–1103. [[CrossRef](#)] [[PubMed](#)]
35. Bycroft, C.; Freeman, C.; Petkova, D.; Band, G.; Elliott, L.T.; Sharp, K.; Motyer, A.; Vukovic, D.; Delaneau, O.; O’Connell, J.; et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **2018**, *562*, 203–209. [[CrossRef](#)]
36. Zhou, W.; Nielsen, J.B.; Fritsche, L.G.; Dey, R.; Gabrielsen, M.E.; Wolford, B.N.; LeFaive, J.; VandeHaar, P.; Gagliano, S.A.; Gifford, A.; et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* **2018**, *50*, 1335–1341. [[CrossRef](#)] [[PubMed](#)]
37. Rohan, L.F.; Dorian, G. Bayesian Methods Applied to GWAS. *Methods Mol. Biol.* **2013**, *1019*, 237–274. [[CrossRef](#)]
38. van Erp, N.; Gelder, P. van Bayesian logistic regression analysis. *AIP Conf. Proc.* **2013**, *1553*, 147. [[CrossRef](#)]
39. Meuwissen, T.H.E.; Hayes, B.J.; Goddard, M.E. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics* **2001**, *157*, 1819–1829. [[CrossRef](#)]
40. Benner, C.; Spencer, C.C.A.; Havulinna, A.S.; Salomaa, V.; Ripatti, S.; Pirinen, M. FINEMAP: Efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **2016**, *32*, 1493. [[CrossRef](#)]
41. Banerjee, S.; Zeng, L.; Schunkert, H.; Söding, J. Bayesian multiple logistic regression for case-control GWAS. *PLoS Genet.* **2018**, *14*, e1007856. [[CrossRef](#)] [[PubMed](#)]
42. Lloyd-Jones, L.R.; Zeng, J.; Sidorenko, J.; Yengo, L.; Moser, G.; Kemper, K.E.; Wang, H.; Zheng, Z.; Magi, R.; Esko, T.; et al. Improved polygenic prediction by Bayesian multiple regression on summary statistics. *Nat. Commun.* **2019**, *10*, 5086. [[CrossRef](#)] [[PubMed](#)]
43. Yang, Y.; Basu, S.; Mirabello, L.; Spector, L.G.; Zhang, L. A Bayesian Gene-Based Genome-Wide Association Study Analysis of Osteosarcoma Trio Data Using a Hierarchically Structured Prior. *Cancer Inform.* **2018**, *17*. [[CrossRef](#)] [[PubMed](#)]
44. Turchin, M.C.; Stephens, M. Bayesian multivariate reanalysis of large genetic studies identifies many new associations. *PLoS Genet.* **2019**, *15*, e1008431. [[CrossRef](#)]
45. Pe’er, I.; Yelensky, R.; Altshuler, D.; Daly, M.J. Estimation of the multiple testing burden for genome wide association studies of nearly all common variants. *Genet. Epidemiol.* **2008**, *32*, 381–385. [[CrossRef](#)] [[PubMed](#)]
46. Risch, N.; Merikangas, K. The future of genetic studies of complex human diseases. *Science* **1996**, *273*, 1516–1517. [[CrossRef](#)]
47. Visscher, P.M.; Wray, N.R.; Zhang, Q.; Sklar, P.; McCarthy, M.L.; Brown, M.A.; Yang, J. 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **2017**, *101*, 5–22. [[CrossRef](#)] [[PubMed](#)]
48. Goddard, M.E.; Wray, N.R.; Verbyla, K.; Visscher, P.M. Estimating Effects and Making Predictions from Genome-Wide Marker Data. *Stat. Sci.* **2009**, *24*, 517–529. [[CrossRef](#)]
49. Buniello, A.; MacArthur, J.A.L.; Cerezo, M.; Harris, L.W.; Hayhurst, J.; Malangone, C.; McMahon, A.; Morales, J.; Mountjoy, E.; Sollis, E.; et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **2019**, *47*, D1005–D1012. [[CrossRef](#)]

50. Watanabe, K.; Stringer, S.; Frei, O.; Umičević Mirkov, M.; de Leeuw, C.; Polderman, T.J.C.; van der Sluis, S.; Andreassen, O.A.; Neale, B.M.; Posthuma, D. A global overview of pleiotropy and genetic architecture in complex traits. *Nat. Genet.* **2019**, *51*, 1339–1348. [[CrossRef](#)]
51. Beck, T.; Hastings, R.K.; Gollapudi, S.; Free, R.C.; Brookes, A.J. GWAS Central: A comprehensive resource for the comparison and interrogation of genome-wide association studies. *Eur. J. Hum. Genet.* **2014**, *22*, 949–952. [[CrossRef](#)] [[PubMed](#)]
52. Tam, V.; Patel, N.; Turcotte, M.; Bossé, Y.; Paré, G.; Meyre, D. Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* **2019**, *20*, 467–484. [[CrossRef](#)] [[PubMed](#)]
53. Ripke, S.; Neale, B.M.; Corvin, A.; Walters, J.T.R.; Farh, K.H.; Holmans, P.A.; Lee, P.; Bulik-Sullivan, B.; Collier, D.A.; Huang, H.; et al. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **2014**, *511*, 421–427. [[CrossRef](#)]
54. Steinthorsdottir, V.; Thorleifsson, G.; Sulem, P.; Helgason, H.; Grarup, N.; Sigurdsson, A.; Helgadóttir, H.T.; Johannsdóttir, H.; Magnusson, O.T.; Gudjonsson, S.A.; et al. Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. *Nat. Genet.* **2014**, *46*, 294–298. [[CrossRef](#)]
55. Sakaue, S.; Kanai, M.; Tanigawa, Y.; Karjalainen, J.; Kurki, M.; Koshihara, S.; Narita, A.; Konuma, T.; Yamamoto, K.; Akiyama, M.; et al. A global atlas of genetic associations of 220 deep phenotypes. *MedRxiv* **2020**, *46*, 20213652. [[CrossRef](#)]
56. Alonso, L.; Piron, A.; Morán, I.; Guindo-Martínez, M.; Bonas-Guarch, S.; Atla, G.; Miguel-Escalada, I.; Royo, R.; Puiggros, M.; García-Hurtado, X.; et al. TIGER: The gene expression regulatory variation landscape of human pancreatic islets. *Cell Rep.* **2021**, *37*, 109807. [[CrossRef](#)]
57. Sudlow, C.; Gallacher, J.; Allen, N.; Beral, V.; Burton, P.; Danesh, J.; Downey, P.; Elliott, P.; Green, J.; Landray, M.; et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Med.* **2015**, *12*, 1001779. [[CrossRef](#)] [[PubMed](#)]
58. Nagai, A.; Hirata, M.; Kamatani, Y.; Muto, K.; Matsuda, K.; Kiyohara, Y.; Ninomiya, T.; Tamakoshi, A.; Yamagata, Z.; Mushiroda, T.; et al. Overview of the BioBank Japan Project: Study design and profile. *J. Epidemiol.* **2017**, *27*, S2–S8. [[CrossRef](#)]
59. Borodulin, K.; Tolonen, H.; Jousilahti, P.; Jula, A.; Juolevi, A.; Koskinen, S.; Kuulasmaa, K.; Laatikainen, T.; Männistö, S.; Peltonen, M.; et al. Cohort Profile: The National FINRISK Study. *Int. J. Epidemiol.* **2018**, *47*, 696–696i. [[CrossRef](#)]
60. Panagiotou, O.A.; Willer, C.J.; Hirschhorn, J.N.; Ioannidis, J.P.A. The Power of Meta-Analysis in Genome-Wide Association Studies. *Annu. Rev. Genom. Hum. Genet.* **2013**, *14*, 441–465. [[CrossRef](#)]
61. Evangelou, E.; Ioannidis, J.P.A. Meta-analysis methods for genome-wide association studies and beyond. *Nat. Rev. Genet.* **2013**, *14*, 379–389. [[CrossRef](#)]
62. Hivert, V.; Sidorenko, J.; Rohart, F.; Goddard, M.E.; Yang, J.; Wray, N.R.; Yengo, L.; Visscher, P.M. Estimation of non-additive genetic variance in human complex traits from a large sample of unrelated individuals. *Am. J. Hum. Genet.* **2021**, *108*, 786–798. [[CrossRef](#)]
63. Lamy, P.; Grove, J.; Wiuf, C. A review of software for microarray genotyping. *Hum. Genom.* **2011**, *5*, 304–309. [[CrossRef](#)]
64. Marchini, J.; Howie, B.; Myers, S.; McVean, G.; Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **2007**, *39*, 906–913. [[CrossRef](#)]
65. Das, S.; Abecasis, G.R.; Browning, B.L. Genotype Imputation from Large Reference Panels. *Annu. Rev. Genom. Hum. Genet.* **2018**, *19*, 73–96. [[CrossRef](#)] [[PubMed](#)]
66. Li, Y.; Willer, C.; Sanna, S.; Abecasis, G. Genotype Imputation. *Annu. Rev. Genom. Hum. Genet.* **2009**, *10*, 387. [[CrossRef](#)]
67. Boomsma, D.I.; Wijmenga, C.; Slagboom, E.P.; Swertz, M.A.; Karssen, L.C.; Abdellaoui, A.; Ye, K.; Guryev, V.; Vermaat, M.; Van Dijk, E.; et al. The Genome of the Netherlands: Design, and project goals. *Eur. J. Hum. Genet.* **2014**, *22*, 221–227. [[CrossRef](#)] [[PubMed](#)]
68. The UK10K Consortium The UK10K project identifies rare variants in health and disease. *Nature* **2015**, *526*, 82–90. [[CrossRef](#)]
69. McCarthy, S.; Das, S.; Kretzschmar, W.; Delaneau, O.; Wood, A.R.; Teumer, A.; Kang, H.M.; Fuchsberger, C.; Danecek, P.; Sharp, K.; et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **2016**, *48*, 1279–1283. [[CrossRef](#)] [[PubMed](#)]
70. Taliun, D.; Harris, D.N.; Kessler, M.D.; Carlson, J.; Szpiech, Z.A.; Torres, R.; Taliun, S.A.G.; Corvelo, A.; Gogarten, S.M.; Kang, H.M.; et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **2021**, *590*, 290–299. [[CrossRef](#)]
71. Valls-Margarit, J.; Galván-Femenía, I.; Matias, D.; Blay, N.; Puiggros, M.; Carreras, A.; Salvo, C.; Cortés, B.; Amela, R.; Farre, X.; et al. GCAT1 Panel, a comprehensive structural variant haplotype map of the Iberian population from high-coverage whole-genome sequencing. *bioRxiv* **2021**, *21*, 453041. [[CrossRef](#)]
72. Marchini, J. Haplotype Estimation and Genotype Imputation. In *Handbook of Statistical Genomics*; Wiley: Hoboken, NJ, USA, 2019; Volume 1, pp. 87–114.
73. Scheet, P.; Stephens, M. A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* **2006**, *78*, 629–644. [[CrossRef](#)] [[PubMed](#)]
74. Burton, P.R.; Clayton, D.G.; Cardon, L.R.; Craddock, N.; Deloukas, P.; Duncanson, A.; Kwiatkowski, D.P.; McCarthy, M.L.; Ouwehand, W.H.; Samani, N.J.; et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **2007**, *447*, 661–678. [[CrossRef](#)]
75. Li, Y.; Willer, C.J.; Ding, J.; Scheet, P.; Abecasis, G.R. MaCH: Using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* **2010**, *34*, 816–834. [[CrossRef](#)]
76. Naj, A.C. Genotype Imputation in Genome-Wide Association Studies. *Curr. Protoc. Hum. Genet.* **2019**, *102*, e84. [[CrossRef](#)] [[PubMed](#)]

77. Lo, C. Algorithms for Haplotype Phasing. Available online: [https://cseweb.ucsd.edu/~\[chl107\]/pubs/re.pdf](https://cseweb.ucsd.edu/~[chl107]/pubs/re.pdf) (accessed on 30 April 2021).
78. Ahlqvist, E.; Storm, P.; Käräjämäki, A.; Martinell, M.; Dorkhan, M.; Carlsson, A.; Vikman, P.; Prasad, R.B.; Aly, D.M.; Almgren, P.; et al. Novel subgroups of adult-onset diabetes and their association with outcomes: A data-driven cluster analysis of six variables. *Lancet Diabetes Endocrinol.* **2018**, *6*, 361–369. [[CrossRef](#)]
79. Ahlqvist, E.; Prasad, R.B.; Groop, L. Subtypes of Type 2 Diabetes Determined From Clinical Parameters. *Diabetes* **2020**, *69*, 2086–2093. [[CrossRef](#)]
80. Waters, K.; Stram, D.; Hassanein, M.; Le Marchand, L.; Wilkens, L.; Maskarinec, G.; Monroe, K.; Kolonel, L.; Altshuler, D.; Henderson, B.; et al. Consistent association of type 2 diabetes risk variants found in Europeans in diverse racial and ethnic groups. *PLoS Genet.* **2010**, *6*, e1001078. [[CrossRef](#)]
81. Imamura, M.; Takahashi, A.; Yamauchi, T.; Hara, K.; Yasuda, K.; Grarup, N.; Zhao, W.; Wang, X.; Huerta-Chagoya, A.; Hu, C.; et al. Genome-wide association studies in the Japanese population identify seven novel loci for type 2 diabetes. *Nat. Commun.* **2016**, *7*, 10531. [[CrossRef](#)]
82. Chen, J.; Spracklen, C.N.; Marenne, G.; Varshney, A.; Corbin, L.J.; Luan, J.; Willems, S.M.; Wu, Y.; Zhang, X.; Horikoshi, M.; et al. The trans-ancestral genomic architecture of glycaemic traits. *Nat. Genet.* **2021**, *53*, 840–860. [[CrossRef](#)]
83. Chen, M.-H.; Raffield, L.M.; Mousa, A.; Sakaue, S.; Huffman, J.E.; Moscati, A.; Trivedi, B.; Jiang, T.; Akbari, P.; Vuckovic, D.; et al. Trans-ethnic and Ancestry-Specific Blood-Cell Genetics in 746,667 Individuals from 5 Global Populations. *Cel* **2020**, *182*, 1198–1213.e14. [[CrossRef](#)]
84. Marchini, J.; Donnelly, P.; Cardon, L.R. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat. Genet.* **2005**, *37*, 413–417. [[CrossRef](#)]
85. Álvarez-Castro, J.M. Gene–Environment Interaction in the Era of Precision Medicine—Filling the Potholes Rather Than Starting to Build a New Road. *Front. Genet.* **2020**, *11*, 6. [[CrossRef](#)]
86. Manolio, T.A. Genomewide Association Studies and Assessment of the Risk of Disease. *N. Engl. J. Med.* **2010**, *363*, 166–176. [[CrossRef](#)]
87. White, D.; Rabago-Smith, M. Genotype-phenotype associations and human eye color. *J. Hum. Genet.* **2011**, *56*, 5–7. [[CrossRef](#)]
88. Cordell, H.J. Detecting gene–gene interactions that underlie human diseases. *Nat. Rev. Genet.* **2009**, *10*, 392–404. [[CrossRef](#)] [[PubMed](#)]
89. Kirino, Y.; Bertias, G.; Ishigatsubo, Y.; Mizuki, N.; Tugal-Tutkun, I.; Seyahi, E.; Ozyazgan, Y.; Sacli, F.S.; Erer, B.; Inoko, H.; et al. Genome-wide association analysis identifies new susceptibility loci for Behçet’s disease and epistasis between HLA-B*51 and ERAP1. *Nat. Genet.* **2013**, *45*, 202–207. [[CrossRef](#)] [[PubMed](#)]
90. Monir, M.M.; Zhu, J. Comparing GWAS Results of Complex Traits Using Full Genetic Model and Additive Models for Revealing Genetic Architecture. *Sci. Rep.* **2017**, *7*, 38600. [[CrossRef](#)]
91. Wan, X.; Yang, C.; Yang, Q.; Xue, H.; Fan, X.; Tang, N.L.S.; Yu, W. BOOST: A fast approach to detecting gene–gene interactions in genome-wide case-control studies. *Am. J. Hum. Genet.* **2010**, *87*, 325–340. [[CrossRef](#)]
92. Behravan, H.; Hartikainen, J.M.; Tengström, M.; Pylkäs, K.; Winqvist, R.; Kosma, V.; Mannerman, A. Machine learning identifies interacting genetic variants contributing to breast cancer risk: A case study in Finnish cases and controls. *Sci. Rep.* **2018**, *8*, 13149. [[CrossRef](#)] [[PubMed](#)]
93. Ritchie, M.D.; Hahn, L.W.; Roodi, N.; Bailey, L.R.; Dupont, W.D.; Parl, F.F.; Moore, J.H. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am. J. Hum. Genet.* **2001**, *69*, 138–147. [[CrossRef](#)]
94. Hahn, L.W.; Ritchie, M.D.; Moore, J.H. Multifactor dimensionality reduction software for detecting gene–gene and gene–environment interactions. *Bioinformatics* **2003**, *19*, 376–382. [[CrossRef](#)]
95. Moore, J.H. Computational analysis of gene–gene interactions using multifactor dimensionality reduction. *Expert Rev. Mol. Diagn.* **2004**, *4*, 795–803. [[CrossRef](#)] [[PubMed](#)]
96. Zhang, Y.; Liu, J.S. Bayesian inference of epistatic interactions in case-control studies. *Nat. Genet.* **2007**, *39*, 1167–1173. [[CrossRef](#)]
97. Kerin, M.; Marchini, J. Gene–environment interactions using a Bayesian whole genome regression model. *bioRxiv* **2019**, *19*, 797829. [[CrossRef](#)]
98. Gayán, J.; González-Pérez, A.; Bermudo, F.; Sáez, M.E.; Royo, J.L.; Quintas, A.; Galan, J.J.; Morón, F.J.; Ramírez-Lorca, R.; Real, L.M.; et al. A method for detecting epistasis in genome-wide studies using case-control multi-locus association analysis. *BMC Genom.* **2008**, *9*, 360. [[CrossRef](#)]
99. Dempfle, A.; Scherag, A.; Hein, R.; Beckmann, L.; Chang-Claude, J.; Schäfer, H. Gene–environment interactions for complex traits: Definitions, methodological requirements and challenges. *Eur. J. Hum. Genet.* **2008**, *16*, 1164–1172. [[CrossRef](#)] [[PubMed](#)]
100. Bookman, E.B.; McAllister, K.; Gillanders, E.; Wanke, K.; Balshaw, D.; Rutter, J.; Reedy, J.; Shaughnessy, D.; Agurs-Collins, T.; Paltoo, D.; et al. Gene–environment interplay in common complex diseases: Forging an integrative model—Recommendations from an NIH workshop. *Genet. Epidemiol.* **2011**, *35*, 217–225. [[CrossRef](#)]
101. Patel, C.J.; Bhattacharya, J.; Butte, A.J. An environment-wide association study (EWAS) on type 2 diabetes mellitus. *PLoS ONE* **2010**, *5*, e10746. [[CrossRef](#)]
102. Thomas, D. Methods for investigating gene–environment interactions in candidate pathway and genome-wide association studies. *Annu. Rev. Public Health* **2010**, *31*, 21–36. [[CrossRef](#)] [[PubMed](#)]
103. Simon, P.H.G.; Sylvestre, M.P.; Tremblay, J.; Hamet, P. Key Considerations and Methods in the Study of Gene–Environment Interactions. *Am. J. Hypertens.* **2016**, *29*, 891–899. [[CrossRef](#)]

104. Han, S.S.; Chatterjee, N. Review of Statistical Methods for Gene-Environment Interaction Analysis. *Curr. Epidemiol. Rep.* **2018**, *5*, 39–45. [[CrossRef](#)]
105. McAllister, K.; Mechanic, L.E.; Amos, C.; Aschard, H.; Blair, I.A.; Chatterjee, N.; Conti, D.; Gauderman, W.J.; Hsu, L.; Hutter, C.M.; et al. Current Challenges and New Opportunities for Gene-Environment Interaction Studies of Complex Diseases. *Am. J. Epidemiol.* **2017**, *186*, 753–761. [[CrossRef](#)] [[PubMed](#)]
106. Thomas, D. Gene-Environment-Wide Association Studies: Emerging Approaches. *Nat. Rev. Genet.* **2010**, *11*, 259. [[CrossRef](#)] [[PubMed](#)]
107. Zheng, Y.; Chen, Z.; Pearson, T.; Zhao, J.; Hu, H.; Prosperi, M. Design and methodology challenges of environment-wide association studies: A systematic review. *Environ. Res.* **2020**, *183*, 109275. [[CrossRef](#)]
108. Cano-Gamez, E.; Trynka, G. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Front. Genet.* **2020**, *11*, 424. [[CrossRef](#)] [[PubMed](#)]
109. Lichou, F.; Trynka, G. Functional studies of GWAS variants are gaining momentum. *Nat. Commun.* **2020**, *11*, 6283. [[CrossRef](#)]
110. Lambert, S.A.; Abraham, G.; Inouye, M. Towards clinical utility of polygenic risk scores. *Hum. Mol. Genet.* **2019**, *28*, R133–R142. [[CrossRef](#)]
111. Choi, S.W.; Mak, T.S.H.; O'Reilly, P.F. Tutorial: A guide to performing polygenic risk score analyses. *Nat. Protoc.* **2020**, *15*, 2759–2772. [[CrossRef](#)] [[PubMed](#)]
112. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **2012**, *489*, 57–74. [[CrossRef](#)]
113. Lonsdale, J.; Thomas, J.; Salvatore, M.; Phillips, R.; Lo, E.; Shad, S.; Hasz, R.; Walters, G.; Garcia, F.; Young, N.; et al. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **2013**, *45*, 580–585. [[CrossRef](#)] [[PubMed](#)]
114. Manolio, T.A.; Collins, F.S.; Cox, N.J.; Goldstein, D.B.; Hindorf, L.A.; Hunter, D.J.; McCarthy, M.I.; Ramos, E.M.; Cardon, L.R.; Chakravarti, A.; et al. Finding the missing heritability of complex diseases. *Nature* **2009**, *461*, 747–753. [[CrossRef](#)]
115. Taylor, D.L.; Jackson, A.U.; Narisu, N.; Hemani, G.; Erdos, M.R.; Chines, P.S.; Swift, A.; Idol, J.; Didion, J.P.; Welch, R.P.; et al. Integrative analysis of gene expression, DNA methylation, physiological traits, and genetic variation in human skeletal muscle. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 10883–10888. [[CrossRef](#)]
116. Beesley, J.; Sivakumaran, H.; Moradi Marjaneh, M.; Shi, W.; Hillman, K.M.; Kaufmann, S.; Hussein, N.; Kar, S.; Lima, L.G.; Ham, S.; et al. eQTL Colocalization Analyses Identify NTN4 as a Candidate Breast Cancer Risk Gene. *Am. J. Hum. Genet.* **2020**, *107*, 778–787. [[CrossRef](#)]
117. Martin, A.R.; Kanai, M.; Kamatani, Y.; Okada, Y.; Neale, B.M.; Daly, M.J. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **2019**, *51*, 584–591. [[CrossRef](#)]
118. McGuire, A.L.; Gabriel, S.; Tishkoff, S.A.; Wonkam, A.; Chakravarti, A.; Furlong, E.E.M.; Treutlein, B.; Meissner, A.; Chang, H.Y.; López-Bigas, N.; et al. The road ahead in genetics and genomics. *Nat. Rev. Genet.* **2020**, *21*, 581–596. [[CrossRef](#)]
119. Mulder, N.; Abimiku, A.; Adebamowo, S.N.; de Vries, J.; Matimba, A.; Olowoyo, P.; Ramsay, M.; Skelton, M.; Stein, D.J. H3Africa: Current perspectives. *Pharmacogenomics Pers. Med.* **2018**, *11*, 59–66. [[CrossRef](#)] [[PubMed](#)]
120. Miga, K.H.; Wang, T. The Need for a Human Pangenome Reference Sequence. *Rev. Genom. Hum. Genet.* **2021**, *22*, 81–102. [[CrossRef](#)] [[PubMed](#)]

