UNIVERSITAT DE BARCELONA

A systematic and comprehensive approach for large-scale genome-wide association studies

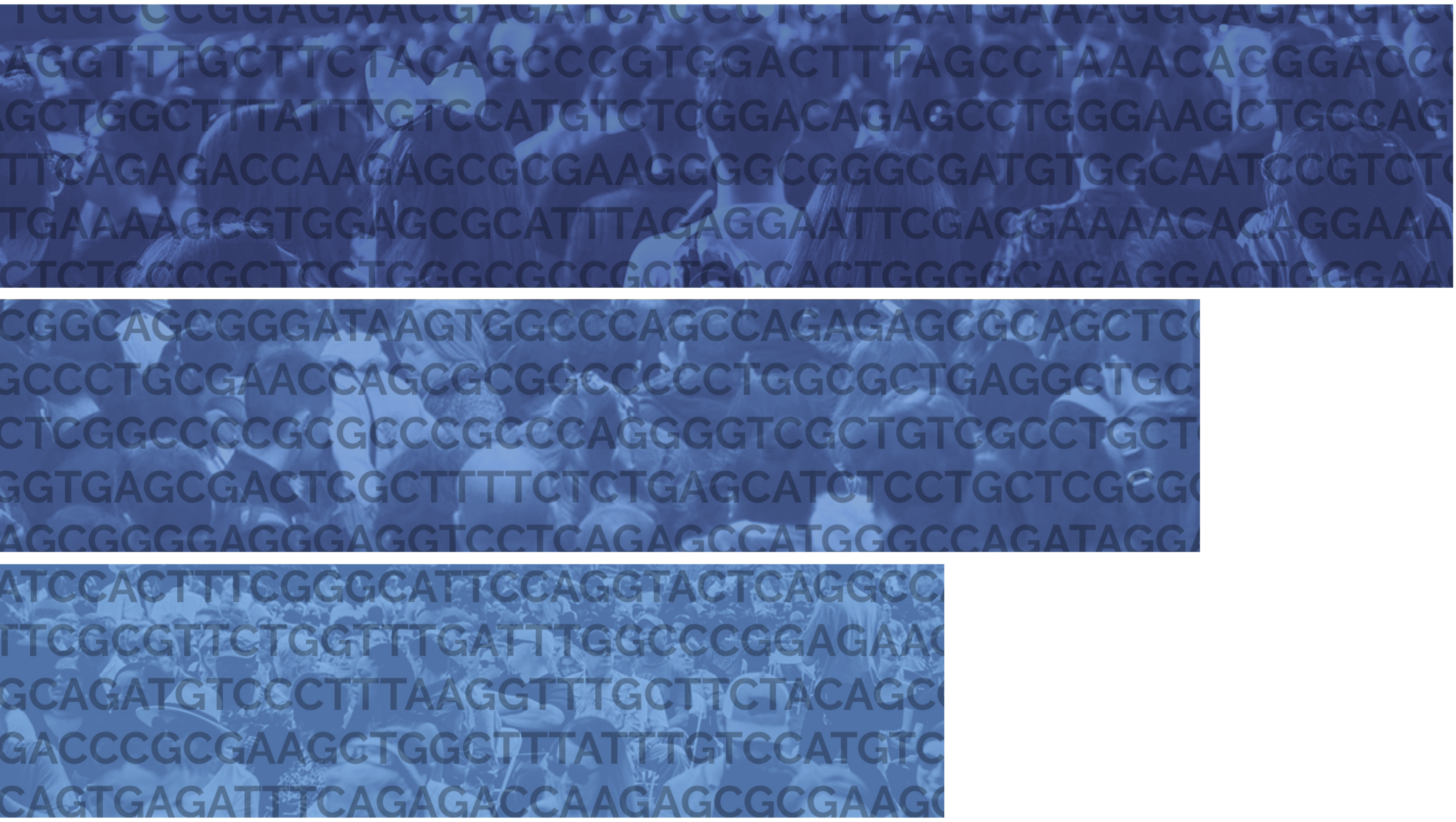
Unraveling non-additive inheritance models in age-related diseases

Marta Guindo Martínez

A systematic and comprehensive approach for large-scale genome-wide association studies

Unraveling non-additive inheritance models in age-related diseases

Ph.D. thesis



Marta Guindo Martínez
October 2019

UNIVERSITAT DE BARCELONA

Facultat de Biologia

Programa de Biomedicina (codi HDK05)

Línia de recerca 101114 – Bioinformàtica

Memòria presentada per Marta Guindo Martínez per optar al grau de doctor/a per la Universitat de Barcelona

A systematic and comprehensive approach for large-scale genome-wide association studies

Unraveling non-additive inheritance models in age-related diseases

DOCTORANDA:

Marta Guindo Martínez

DIRECTORS:

Dr. David Torrents Arenales

Dr. Josep María Mercader Bigas

TUTOR:

Dr. Josep Lluís Gelpi Buchaca

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Agraïments

No concebo aquesta tesi com una fita personal sinó de grup del qual en sóc la cara visible. Aquesta memòria per optar al grau de doctora que aquí presento no hagués estat possible sense el treball i suport constant de moltes persones que he tingut la sort de creuar-me. Sóc conscient que em serà impossible anomenar-les a totes i cadascuna d’elles, i espero però, que totes siguin conscients que els hi estic infinitament agraïda. No obstant això, no vull deixar passar l'oportunitat de mostrar personalment el meu agraïment a alguns dels que considero pilars d'aquesta tesi.

En primer lloc voldria agrair als membres del tribunal haver acceptat la invitació a l'acte de defensa de la meva tesi. És un plaer i un honor per mi poder comptar amb ells i la seva opinió[*].

Però, sobretot, voldria agrair als meus supervisors, al David Torrents i al Josep Maria (Txema) Mercader la oportunitat que m’han donat amb aquesta tesi. Recordo quan ens vam conèixer, on vam tenir la primera conversa sobre la possible tesi, i estic infinitament agraïda per la confiança que van dipositar en mi per dur a terme aquest projecte en un entorn privilegiat, tant per la part tècnica com la humana. Ha estat un plaer treballar tots tres, sempre buscant el millor pel projecte. A tots dos els hi estic molt agraïda pel seu suport i l'impuls que m’han donat per formar-me tant dins com fora del grup, podent participar en congressos i projectes que m'han permès créixer al llarg del doctorat, guanyant confiança en el projecte i en mi mateixa. D’entre totes aquestes experiències vull destacar una estada al Broad Institute que va acabar de perfilar la idea que vull perseguir en el meu futur a la recerca. Per tant, gràcies també al José Carlos Florez, al Jordi i a tota la gent del Broad i el Massachusetts General Hospital per l'ambient tan estimulant en el qual vaig ser tan ben rebuda.

Però tornant als meus supervisors...

[*] First of all, I would like to thank the members of my thesis committee for accepting the invitation to evaluate my thesis. It is a pleasure and an honor for me to have them in my thesis committee and to have their feedback.

En particular, al David vull agrair especialment el seu tracte humà i la seva paciència, les converses, la motivació, i el seu respecte i èmfasi en la vida que ens passa a tots més enllà de la feina. Tant de bo tots els supervisors sabessin fer-ho.

Al Txema li dec especialment, i en gran part, tot el que he après aquests anys i que queda escrit en aquesta tesi. Tots els que hem treballat amb ell hem tingut la sort de viure la constància i passió que hi posa a la recerca, el seu humor, la seva franquesa i els seus consells, que han anat més enllà dels aspectes tècnics. D'ell he après també com treballar més hores que un rellotge i que precisament això et faci feliç. Mostra de tot el que he mencionat anteriorment és el compromís que va adquirir amb el projecte, pel qual li estic infinitament agraïda. El fet que marxés del grup fa tres anys per continuar amb la seva carrera als Estats Units no s'ha notat gens en la seva implicació en tot el temps que ha durat la tesi. Sí que hem notat tots, però, que faltava el so de la seva bicicleta entrant a la sala cada matí. Perquè el Txema, a part d'haver sigut un excel·lent company de feina i mentor, és un amic. Per estar sempre, 24/7, a l'altra banda del telèfon, gràcies!

En aquest entorn privilegiat on s'ha desenvolupat el meu doctorat vull destacar a la Sílvia, una sort de germana gran a l'equip GWAS. A ella li agraeixo el seu acolliment quan era nouvinguda, la llavor d'aquest projecte, i tot el que m'ha ensenyat i inspirat.

El mateix passa amb l'Elias i la Mercè, amb els quals he tingut el plaer indescriptible de coincidir. Ells, de principi a fi, han fet d'aquest període de la meva vida una festa, en el seu sentit més ampli. Les converses, tant de recerca com mirant d'arreglar el món, els viatges, els sopars, els cafès, ... han sigut un motor per mi i per aquesta tesi, un caldo de cultiu excel·lent, un lloc segur on resguardar-me quan venien magres, i on celebrar les bones notícies. Si no hagués coincidit amb ells, ni jo ni la meva tesi hauríem estat el que som ara.

Totes les persones que en algun moment han fet acte de presència al Life Science Department, i especialment al Computational Genomics Group, durant aquests 5 anys han aportat, tant en la part tècnica com personal, en el resultat final de la tesi que aquí presento. També totes les persones que he conegut al llarg d'aquest anys en congressos i cursos. Algunes d'aquestes persones tinc la sort de considerar-les amigues.

Pel seu aport en aquesta tesi, d'entre elles vull d'estacar a la Montse, la informàtica per excel·lència del grup, per la seva infinita paciència ajudant-me amb les meves primeres passes amb Java, i tot el que m'ha ensenyat sobre bash i, en general, en la computació. Sempre la tindré associada al moment del cafè, i a la Romina! A qui he d'agrair la seva generositat mantenint-nos sempre tan ben alimentats.

Gràcies també als membres de Support per la seva paciència i flexibilitat amb els problemes que ha comportat el desenvolupament de GUIDANCE i un anàlisi de la magnitud del presentat en aquesta tesi.

Gràcies també al grup de COMPSs perquè m'han fet encarar de ple la informàtica més pura, i especialment al Ramon, co-autor de l'anàlisi que aquí es presenta, per oferir-se a portar la part més computacional de GUIDANCE. Sense ell segurament ara només tindríem una amalgama de scripts i idees però cap "integrated framework on top of COMPSs".

Si tiro la vista enrere, vull també donar les gràcies a l'equip de l'ICCC, la Lina Badimon, la Gemma Vilahur, la Rosa Suades, el Pablo, la Mari, la Sílvia... Amb només 18 anyets, quan tot just començant la carrera però encara no sabia que faria amb la meva vida, em van agafar per "ajudar-los" cada estiu durant 5 anys en diferents projectes de recerca. Va ser en aquell primer estiu quan vaig decidir que el Ph.D. seria l'objectiu d'estudiar biologia, i no vaig dubtar més en tota la carrera. No puc estar més contenta d'aquella decisió. D'ells vaig aprendre a fer les primeres passes.

Que canviés els guants i la bata pel teclat va ser conseqüència de la meva experiència durant el màster, i d'en Marc, que precisament va aparèixer en aquella etapa, i que em va animar a fer el salt a la bioinformàtica. El canvi de ruta que vaig fer, tant en l'aspecte pràctic com en el personal, li dec en gran part a la inspiració que em va aportar la seva visió de la vida, la seva curiositat, i també al seu suport. Suport que es mantingué durant anys, i tinc la sort de poder dir que roman encara. Gràcies per les xerrades, el teu seny i la confiança mútua mantinguda.

D'aquesta etapa també vull agrair als meus gats la seva paciència escoltant-me hores i hores practicant xerrades davant d'ells. I per descomptat la seva companyia en els dies més llargs. Un és fruit directe del BSC, ja que me'l vaig trobar a les portes, així que no seran menys. Pocs saben què ha passat de principi a fi "de puertas adentro" com ho saben ells (o ho sabrien si els gats se'n recordessin de coses així...).

I no hauria arribat tan serenament a la tesi, si no fos per la sorpresa de creuar-me amb el Xavi. Gràcies per la teva paciència infinita, i per mirar tant per mi i la meva tesi. No és fàcil aguantar a un doctorant en el seu últim any! No tinc proutes paraules per agrair-te el haver apostat per mi.

Vull agrair també a la meva família i amics el seu suport i la seva comprensió, inclús sense tenir molt clar que comporta fer una tesi en alguns casos. Al llarg dels anys han sabut respectar l'espai que de vegades he necessitat, mostrar interès en la meva tesi i animar-me a aconseguir el que fos que volia, com, d'altra banda, han fet sempre. Gràcies per la vostra eterna presència incondicional, també la d'aquells que ja no hi són. Laia, fa un parell d'anys us vaig voler fer saber a tu i a la Montse que sou per mi un exemple, que són les persones com vosaltres, amb aquesta determinació i força per ser una mateixa, les que ens feien a tots més lliures. I n'he necessitat, de llibertat i de força per decidir fer carrera a la recerca i posar la primera pedra en aquesta tesi. Eternament, gràcies.

Per últim, per descomptat m'agradaria agrair en general a tota la gent que dóna el seu consentiment per participar en la recerca pública, i als professionals que lliuren dades i eines al benefici de tots. Sense l'accés públic a dades i eines aquesta tesi no hagués estat possible. Gràcies a tots pel vostre altruisme.

Abstract

Genome-wide association studies (GWAS) have been proven useful for identifying thousands of associations between genetic variants and human complex diseases and traits. However, the identified loci account for a small proportion of the estimated heritability (i.e., the proportion of variance for a particular phenotype that can be explained by genetic factors).

The usually small effect size of common variants and the low frequencies of some variants with potentially larger effect sizes limit the statistical power of GWAS. The identification of common variants with small effects and low-frequency variants with large effects can be overcome with the analysis of larger sample sizes and imputing genotypes using dense reference panels. However, there is still room for improvement beyond increasing the sample size and the number of variants. As current GWAS are predominantly focused on the autosomes and only test the additive model, current strategies still constrain the full potential of GWAS.

In this thesis, we hypothesized that performing a comprehensive analysis improving current GWAS strategies by 1) implementing the analysis of the X chromosome alongside the autosomes, 2) including genetic variants from a broader allele frequency spectrum and type of variants, such as small insertions and deletions (INDELs) through genotype imputation using multiple reference panels, and 3) testing different models of inheritance in the association test, would improve our understanding of the genetic architecture of complex diseases.

To test these hypotheses we developed an integrated framework including our methodology, called GUIDANCE. Hence, GUIDANCE integrates state-of-the-art tools for GWAS analysis, including the analysis of X chromosome, a two-step imputation with multiple reference panels, the association testing including additive, dominant, recessive, heterodominant and genotypic inheritance models, and cross-phenotype association analysis when more than one disease is available in the cohort under study.

We used GUIDANCE to analyze the Genetic Epidemiology Research on Adult Health and Aging (GERA) cohort, a publicly available cohort that includes 62,281 subjects

from European ancestry with an average age of 63 years for 22 diseases, representing the largest cohort for age-related diseases to date.

After quality control, we analyzed 56,637 subjects from European descendant populations. Following our methodology, we imputed genotypes using 1000 Genomes Project (1000G) phase 3, the Genome of the Netherlands project (GoNL), the UK10K project22, and the Haplotype Reference Consortium (HRC) as reference panels.

Using this strategy, we identified 26 new associated loci for 16 phenotypes ($p < 5 \times 10^{-8}$), with 13 showing significant dominance deviation ($p < 0.05$).

Importantly, we identified three recessive loci with large effects that could not have identified by the additive model. This include a region let by an INDEL associated with cardiovascular disease in *CACNB4* (rs201654520, minor allele frequency [MAF] = 0.017, odds ratio [OR] = 19.02, $p = 4.32 \times 10^{-8}$), a lous near *PELO* associated with type 2 diabetes with the greatest odds ratio for type 2 diabetes in Europeans reported to date (rs77704739, MAF= 0.036, OR = 4.32, $p = 1.75 \times 10^{-8}$), and a rare INDEL associated with age-related macular degeneration near *THUMPD2* (rs557998486, MAF= 0.009, OR = 10.5, $p = 2.75 \times 10^{-8}$).

Despite the phenotype discrepancies and different demographical characteristics of the GERA cohort and UK Biobank, four of the novel loci were replicated with an equivalent phenotype in UK Biobank, and we found additional supporting associations in related traits, treatments or biomarkers in UK Biobank for the remaining novel loci.

Of note, *PELO* and *THUMPD2* recessive loci were replicated using the recessive model in UK Biobank (combined results: *PELO*, rs77704739, OR = 2.46, $p = 4.68 \times 10^{-11}$, and *THUMPD2*, rs557998486, OR = 26.51, $p = 3.29 \times 10^{-8}$), which could not have been found with the additive model.

Overall, these results highlight the importance of performing a comprehensive analysis of the full spectrum of genetic variation and considering non-additive models when performing GWAS, especially with well-powered biobanks and the increasing ability to impute low-frequency variants.

For the benefit of the research community, we make available both GUIDANCE to boost the analysis of existing and ongoing GWAS projects, and the GERA cohort results, which constitute the largest non-additive genetic variation association database to date, through the Type 2 Diabetes Knowledge Portal (http://www.type2diabetesgenetics.org).

Table of contents

Abbreviations and acronyms

1000G	The 1000 Genomes Project
ABF	Approximate Bayes Factor
BMI	Body Mass Index
CD/CV	Common Disease / Common Variant principle
CHR	Chromosome
CI	Confidence Interval
COMPSs	COMP Superscalar
dbGaP	The database of Genotypes and Phenotypes
DEPICT	Data-driven Expression Prioritized Integration for Complex Traits framework
DNA	Deoxyribonucleic Acid
EGA	The European Genome-phenome Archive
ENCODE	The Encyclopedia of DNA Elements
eQTL	Expression quantitative trait loci
eQTLGen	eQTLGen Consortium
ExAC	The Exome Aggregation Consortium
GC	Genomic control
GERA	The Resource for Genetic Epidemiology Research on Aging cohort
GoNL	The Genome of the Netherlands project
GTEx	The Genotype-Tissue Expression project
GWAS	Genome-Wide Association Study
HapMap	The International HapMap (Haplotype Map) Project
HGP	The Human Genome Project
HLA	The Human Leukocyte Antigen complex
HPC	High Performance Computing
HRC	The Haplotype Reference Consortium
HWE	Hardy–Weinberg Equilibrium
IBD	Identity By Descent
IBS	Identity By State
INDEL	Short Insertion-Deletion polymorphisms
K	Thousand
LD	Linkage Disequilibrium
LSF	Load Sharing Facility
M	Million
MAF	Minor Allele Frequency
Mb	Megabase pair
MDS	Multidimensional scaling
MSR	Male-Specific Region

NGS	Next-Generation Sequencing
OR	Odds ratio
PAR	Pseudoautosomal Region
PC	Principal Components
PCA	Principal Component Analysis
PHESANT	PHEnome Scan ANalysis Tool
QC	Quality Control
Q-Q plot	Quantile-Quantile plot
RPGEH	The Kaiser Permanente Research Program on Genes, Environment, and Health
SLURM	Simple Linux Utility for Resource Management
SNP	Single Nucleotide Polymorphism
SV	Structural Variant
TopMED	The Trans-Omics for Precision Medicine program
UK10K	The UK10K project
WGS	Whole-Genome Sequencing
WTCCC	The Wellcome Trust Case Control Consortium

Introduction

Complex diseases are the result of the combination of genetic factors, environmental exposures, and lifestyle choices, and they are also known as “common diseases” since they are spread medical problems worldwide (King et al., 1992). In fact, complex diseases are the leading cause of mortality worldwide, and compromise economies and life quality due to their chronic nature (WHO, 2018). For these reasons, after the successful health policies towards controlling and even eradicating infectious diseases by vaccination and antibiotics, current health care policies are predominantly focused on complex diseases.

Between prevention and treatment, prevention is the most cost-effective approach to address complex disease, and fortunately, many of them are potentially preventable at different levels. Hence, adopting healthy lifestyles, and even higher levels of education, which can influence decision-making patterns (e.g., smoking and obesity are inversely related to education), have been associated with population health (Cutler, 2007). Nevertheless, the genetic background that each person carries from birth (i.e., germline variations) also modifies the risk of developing a complex disease, and the identification of individuals at higher risk due to genetic factors will ultimately lead to better health prevention policies for them.

Even before the beginning of “genetics” as a discipline, many have wondered how traits and diseases are inherited since many of them seem to run in families. For some diseases, called Mendelian, monogenic or rare diseases, the answer did not take long. However, the observations and methods applied to Mendelian diseases were not equally successful for complex, polygenic, common diseases. Even so, significant progress has been made in the field, and nowadays, we have methods to explore, and we better understand, which regions of the human genome (i.e., loci) compromise our health in the long run.

The introduction of this thesis will explain the necessary concepts to understand the theoretical framework on which this thesis has been developed. The introduction begins with an explanation of the concept of heritability and the utility and misconceptions of this measure of the genetic component behind complex diseases (Chapter 1). Right after, there is an historical overview that seeks to explain basic concepts that appear countless times throughout this thesis as well as to put in context how the study of the genetics behind complex diseases began, from the first description of the Mendelian laws of inheritance in 1866 until the first complete genome-wide association study (GWAS) in 2007 (Chapter 2). Chapter 3 starts with a summary about what we have learned from GWAS since 2007 and continues describing the current scenario of these analyses, specifying and detailing the steps that constitute a current GWAS. Finally, Chapter 4 is focused on GWAS results, explaining their interpretation, and exploring possible further analysis after GWAS to clarify the mechanisms through which associated loci might be influencing complex diseases risk.

1 The genetic basis of human traits and disease

1.1 Why family members look alike and its relation with diseases

Independently of environmental factors and lifestyle choices, relatives tend to be more alike compared to random individuals within the same population. The offspring tend to resemble their progenitors, their siblings, and their extended family to a lesser degree. By observing the height, the eye colors, and the hair types of some progenitors, one can make a good guess about the height, the eye color, and the hair of the offspring. In the same way, many common complex diseases are known to run in families (Lobo, 2008), and related individuals display a major risk of developing the same diseases compared to other individuals for a given population.

Shared environmental factors can explain many resemblances between the offspring and parents, but many can also be explained to some extent by genetics. During conception, half of the genetic information of each progenitor is passed to the offspring. Therefore, even though in the presence of the same environment, identical genetic twins look more alike than siblings or adopted children. If a trait can be inherited from the progenitors, it is heritable. Some traits, such as blood type, are fully heritable, but most of the traits and complex diseases are partially heritable, and the environment also contributes to them (King et al., 1992). Hence, up to what extent does genetics influences complex disease risk?

1.2 The variability explained by genetic factors; Heritability

To distinguish between genetic and environmental contributions to a trait, the concept of “heritability” was first introduced by Sewall Wright (1889-1988) and Ronald Fisher (1890-1962) a century ago (Fisher, 1918; Wright, 1920). Heritability is defined as the proportion of variation in a particular trait that is attributable to genetic factors (Visscher et al., 2008). In other words, heritability measures how much of the variability of a specific trait (e.g., a complex disease) can be explained by genetic differences. However, since some genetic factors may indirectly affect the phenotype through environmental factors, heritability has to be carefully interpreted.

Heritability is formally defined as a ratio of variances, where the dominator contains the total observed variation, usually excluding fixed factors and covariates such as age and sex, and the numerator contains the variation that is due to additive genetic values in the population, traditionally called “breeding values” (Falconer and Mackay, 1996).

Although shared environmental factors in related individuals can lead to inflated estimates (Zaitlen et al., 2013) and novel approaches have been developed to estimates heritability based on unrelated and admixed individuals (Zaitlen et al., 2014), most heritability estimates are based on family and twin studies (Visscher et al., 2008).

1.3 Pits and falls of the heritability estimation

As it is an estimate, heritability has many limitations since it is a simplification that does not accurately reflect the complex nature of phenotype-genotype interaction. The complexity and misconception of heritability have led to erroneous assumptions (Visscher et al., 2008). If a trait is difficult to measure or if it depends on who measures it (e.g., self-reported, physician-reported, or empirically measured), the heritability estimate can be lower than the real one. In addition, heritability is population-specific, and the heritability for a particular trait in a population does not predict its heritability in a different one in theory (Visscher et al., 2008). However, in practice, the heritability of similar traits are often similar in different populations (Visscher et al., 2008).

A high heritability implies a strong correlation between genotype and phenotype, i.e., the observed variation is mainly due to genetic factors. Hence, the genetic contributors for a disease with a high heritability can be more easily identified. However, a high heritability does not mean that the genotype can predict the phenotype (i.e., genetic determination) since environmental factors can modify the phenotype (Visscher et al., 2008).

The misconception of the heritability estimates can lead to unfair conclusions trying to justify differences by genetics. An example of this is the controversy about the heritability of intelligence quotient (IQ), a measurement of the performance in a series of mental ability tests. Beyond the problems of measuring cognitive ability, it can be perceived that intelligence could not be modifiable by intervention strategies since its heritability estimate is high (in the range of 0.50 - 0.80) (Visscher et al., 2008). However, it has been demonstrated that the heritability estimate of IQ increases when increasing the socioeconomic status of the group under study (Turkheimer et al., 2003), which demonstrates the importance of environmental factors and to do not extrapolate heritability estimations from one group to others.

In summary, heritability depends on the complexity of the trait under study, how and who measures it, and it is specific to the group under study, including its environment. Hence, it is not valid to use heritability as evidence for "inherited" differences between groups or populations or to predict the individual predisposition to a particular trait. However, heritability is a crucial parameter that puts an upper limit on the efficiency of the possible prediction of the genetic risk of a trait (e.g., a disease).

1.4 What is the heritability of complex diseases?

Heritability ranges between 0 (genetics explains nothing about the trait) and 1 (genetics explains everything). In complex diseases, heritability estimates can be as low as 0.01 for stomach cancer to values as high as 0.81 for schizophrenia (Table 1).

These heritability estimates justify projects to study the genetics behind complex diseases, and these projects lead to a better understanding of the genetic architecture and the pathways involved, for better prevention policies, and better treatments.

Table 1. Heritability estimates for common diseases.

Disease	Heritability	Population	Reference
Schizophrenia	0.81	European	Sullivan et al. (2003)
Chronic Obstructive Pulmonary Disease	0.60	Danish and Swedish	Ingebrigtsen et al. (2010)
Alzheimer	0.58-0.79	Swedish	Gatz et al. (2006)
Melanoma Skin Cancer	0.58	Nordic	Mucci et al. (2016)
Prostate Cancer	0.42-0.57	Nordic	Lichtenstein et al. (2000); Mucci et al. (2016)
Thyroid Cancer	0.53	Swedish	Czene et al. (2002)
Coronary Artery Disease	0.53	Black-defined ancestry from USA	Katzmarzyk et al. (2000)
	0.34-0.49	White-defines ancestry from USA and Germany	Katzmarzyk et al. (2000); Fischer et al. (2005)
Macular Degeneration	0.49-0.71	Dutch	Klaver et al. (1998)
Ovarian Cancer	0.39-0.40	USA and Nordic	Schildkraut et al. (1989); Mucci et al. (2016)
Stroke	0.32	Danish	Bak et al. (2002)
Type 2 Diabetes	0.26	Danish	Poulsen et al. (1999)

Breast Cancer	0.25-0.56	USA and Nordic	Czene et al. (2002);
			Mucci et al. (2016);
			Schildkraut et al. (1989);
			Lichtenstein et al. (2000)

Lung Cancer	0.08	Swedish	Czene et al. (2002)
Leukemia	0.08	Swedish	Czene et al. (2002)
Stomach Cancer	0.01	Swedish	Czene et al. (2002)

Nevertheless, even though hundreds of genetic regions have been associated with complex diseases during the last decade, they only account for a small fraction of the heritability estimates (Manolio et al., 2009). Fortunately, there is room for improvement in the current methods.

To boost our knowledge about the genetics behind complex diseases, find new therapeutic targets or predict their risk by improving current methodologies, we first have to set up what is the genetic material we are made of, how have we learned what it is, and how are we studying the genetic factors that modify diseases risk from birth.

2 Historical overview [†]

2.1 How are traits inherited by the offspring? The early years of genetics

The history of genetics consensually starts with Gregor Mendel (1822-1884). In 1866, this Augustinian monk from the Austro-Hungarian Empire, who was trained in physics, mathematics and philosophy, published his study, "Experiments in Plant Hybridization" (originally "*Versuche über Pflanzenhybriden*"), on how certain traits are inherited following certain rules based on his observation on pea plants. However, the "Mendelian Laws of Inheritance" were put aside until the 20th century. The reason is a contemporary novel theory about evolution and the inability to adequately reconcile both theories at that time (Charlesworth and Charlesworth, 2009).

Nevertheless, in 1900, everything took a turn when Mendel's work was "rediscovered" by Hugo de Vries (1848-1935) in Holand, Carl Correns (1864-1933) in Germany and Erik Tschermak (1871-1962) in Austria through independent works that pointed out to Mendel's studies.

The rediscovery of Mendel's work originates a new debate about inheritance since some inherited traits, such as human height, cannot be explained by Mendelian principles. On the one hand, "biometricians", such as Francis Galton (1822-1911) and Karl Pearson (1857-1936), developed statistical methods to estimate the genetic component of the phenotypic variance of traits, to further decompose genetic variance into additive and non-additive components, which would explain the inheritance of

[†] Before entering into the heart of the matter, I would like to clarify that, beyond the names of the prominent scientists that will appear in this chapter, the important thing to highlight is their ideas, not their names. Probably, their works were the result of a team. Unfortunately, little is known about the researchers with whom they worked and their contributions since their names have been diluted with the storytelling. Unfairly, many names have been forgotten, paving the way to the collective prototype of the white-male scientist (Richmond, M.L. (2007). Opportunities for women in early genetics. Nat Rev Genet *8*, 897-902.). I am afraid, and I apologize, that my summary of the story will not be an exception.

traits such as human height. On the other hand, "Mendelians", such as William Bateson (1961-1926), worked on estimating the effects and models of inheritance of strong effects (e.g., dominant/recessive) (Stranger et al., 2011).

Between 1903 and 1905, Walter Sutton (1877-1916) and Theodore Bovery (1862-1915) independently described in their works the chromosome theory of inheritance, describing chromosomes as the physical carriers of the heredity units that Mendel described (Dahm, 2005). The word "genetics" was used by Bateson in 1905, meaning the study of heredity and variation (Bateson, 2002). In addition to that, in 1905, Bateson and Reginald Punnet (1875-1967) realized that a physical coupling mechanism connected genetic characters increasing the occurrence of purple-long and red-round peas compared to other possible combinations, which contradicts the Mendelian law of independent segregation (Griffiths, 2000). In 1910 Thomas Hunt Morgan (1856-1945) further confirmed this physical coupling mechanism when he found a similar deviation studying Drosophila (white-eyed phenotype tied to males) (Griffiths 2000). In later works, Morgan correctly proposed that the strength of linkage between two genes depends on the distance between the genes on the chromosome (Lobo and Shaw, 2008). This is the basis of why we were able to map genes to specific chromosomes long before sequencing, and the beginning of gene-mapping and human genome maps.

In 1908, Godfrey Harold Hardy (1877-1947) and Wilhelm Weinberg (1862-1937) independently hypothesized what it is now known as the "Hardy-Weinberg Equilibrium" (HWE). This principle states that frequencies of the different forms of a genetic unit (alleles) do not vary over time in a population in the absence of evolutionary forces (Griffiths, 2000).

Finally, in 1918, Ronald Fisher reconciled biometricians and Mendelians in the article "The Correlation between Relatives on the Supposition of Mendelian Inheritance", showing that the debate could be solved by assuming that multiple genes obeying Mendelian rules contribute to variation in a population (Visscher and Goddard, 2019).

During the earlier years of the 20th century, the succession of publications and the iteration of different opinions build up the scientific work that fostered the foundation of population genetics as a discipline (Figure 1).

2.2 The (re)discovery of DNA as the substance of heredity

It is commonly considered in the collective thinking that Watson and Crick discovered the deoxyribonucleic acid (DNA) in 1953. However, we certainly have to go back to 1869, when the Swiss chemist Friedrich Miescher (1844-1895) isolated DNA accidentally while studying the proteins in leucocytes. Hence, he noticed that there was something, which he called "nuclei", that did not match protein's properties (Miescher, 1871). Between 1884 and 1885, Oscar Hertwig (1849-1922), Albrecht von Kölliker (1817-1905), Eduard Strasburger (1844-1912) and August Weismann (1834-1914) independently provide evidence that the cell nucleus contains the basis for inheritance (Dahm, 2005). Unfortunately, DNA was dismissed as the "substance of heredity" until the middle years of the 20th century for not being complex enough compared to proteins (Dahm, 2005; Hargittai, 2009). Until the mid-1940s and early 1950s, when Oswald T. Avery (1877-1955), Colin MacLeod (1909-1972), and Maclyn McCarthy (1911-2005) (Avery et al., 1944), and Al Hershey (1908-1997) and Martha Chase (1927-2003) (Hershey and Chase, 1952) demonstrated that DNA is indeed the carrier (Dahm, 2005).

At this point James Watson (1928-) and Francis Crick (1916-2004), based on the work in crystallography of Maurice Wilkins (1916-2004) and Rosalind Franklin (1920-1958), and putting all the evidences together, including those from Phoebus Levene (1869-1940) (Levene, 1919) and Erwin Chargaff (1905-2002) (Chargaff, 1950), wrote the famous article describing the molecular structure of DNA, published in Nature on April 25, 1953.

Their description of the double helix and the hydrogen bonds provided the first insight into how DNA works (Watson and Crick, 1953). Shortly after, Marshall W. Nirenberg (1927-2010) and J. Heinrich Matthaei (1929-) "cracked" the genetic code (Nirenberg and Matthaei, 1961), deciphering how every three nucleotides, i.e., a codon, build the blocks of proteins, i.e., the amino acids. By 1970s, knowing that genes encode proteins through transcription and translation, and by mimicking the mechanism by

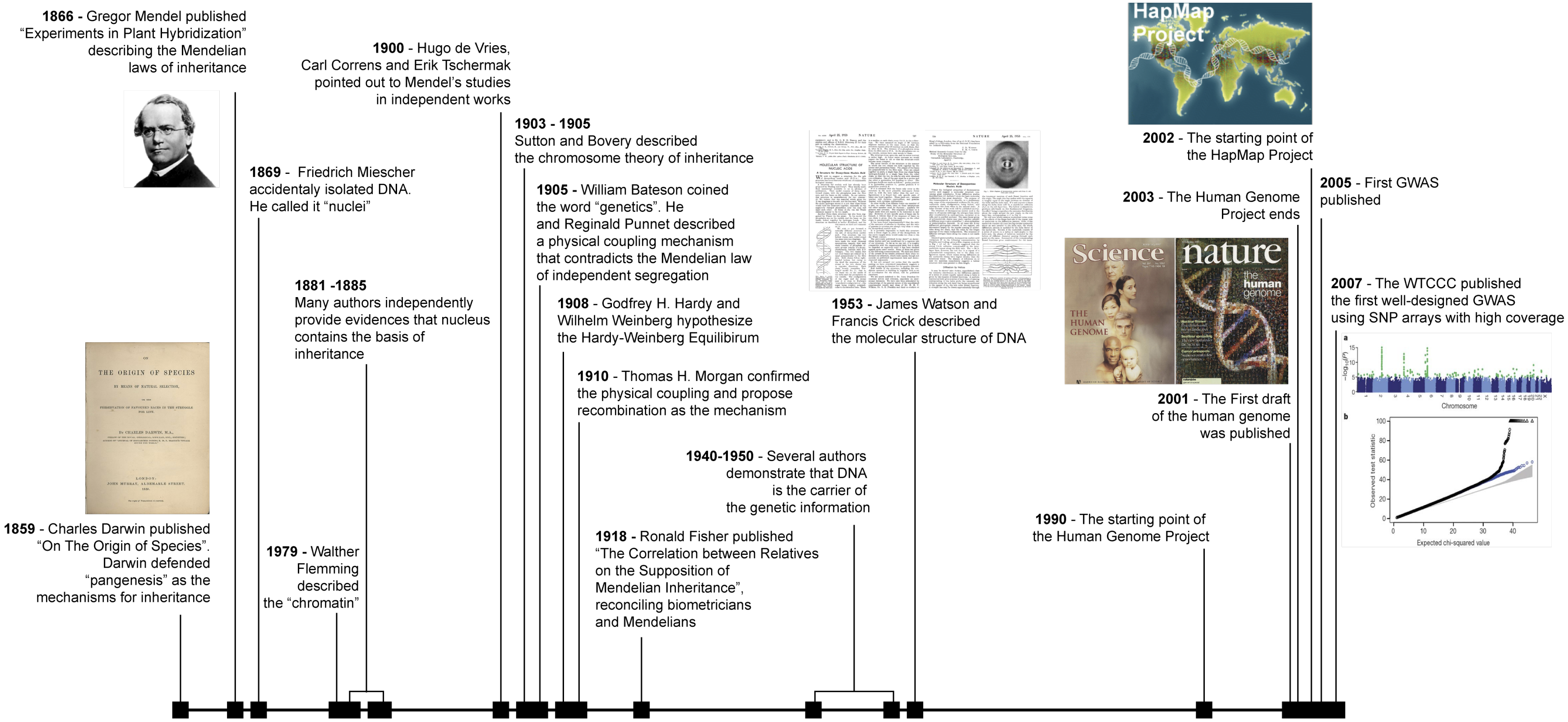

1866 - Gregor Mendel published "Experiments in Plant Hybridization" describing the Mendelian laws of inheritance
1859 - Charles Darwin published "On The Origin of Species". Darwin defended "pangenesis" as the mechanisms for inheritance
1869 - Friedrich Miescher accidentaly isolated DNA. He called it "nuclei"
1979 - Walther Flemming described the "chromatin"
1881 -1885 Many authors independently provide evidences that nucleus contains the basis of inheritance
1900 - Hugo de Vries, Carl Correns and Erik Tschermak pointed out to Mendel's studies in independent works
1903 - 1905 Sutton and Bovery described the chromosome theory of inheritance
1905 - William Bateson coined the word "genetics". He and Reginald Punnet described a physical coupling mechanism that contradicts the Mendelian law of independent segregation
1908 - Godfrey H. Hardy and Wilhelm Weinberg hypothesize the Hardy-Weinberg Equilibirum
1910 - Thomas H. Morgan confirmed the physical coupling and propose recombination as the mechanism
1918 - Ronald Fisher published "The Correlation between Relatives on the Supposition of Mendelian Inheritance", reconciling biometricians and Mendelians
1940-1950 - Several authors demonstrate that DNA is the carrier of the genetic information
1953 - James Watson and Francis Crick described the molecular structure of DNA
1990 - The starting point of the Human Genome Project
2001 - The First draft of the human genome was published
2002 - The starting point of the HapMap Project
2003 - The Human Genome Project ends
2005 - First GWAS published
2007 - The WTCCC published the first well-designed GWAS using SNP arrays with high coverage
HapMap Project
Science
THE HUMAN GENOME
nature
the human genome
THE ORIGIN OF SPECIES
a
b
Chromosome
Observed test statistic
Expected chi-squared value

Figure 1. Historical overview that illustrates the early years of genetics.

The timeline starts with the first theories of inheritance and highlights some of the events that made possible the first Genome-Wide Association Study (GWAS) in 2005.

which cells read the information in genes to translated it into proteins, researchers were able to clone genes for known proteins using recombinant DNA technologies of cloning and sequencing (Lander et al., 2001). This technology was used to find genes for known proteins.

However, for the majority of diseases, the pathways and genes involved were unknown. How we identified genes associated with diseases without any prior knowledge about the etiology?

2.3 The Linkage Studies in family pedigrees

The fact that co-located genes are linked and inherited together (co-segregate) was firstly described for sex-linked traits since they are easier to track through pedigree analysis (Morgan, 1910). This principle allows us to map genes to chromosomes by following co-segregating traits (or phenotypes, e.g., diseases) in family pedigrees (Donahue et al., 1968).

Hence, using the DNA recombination technology to define a locus, linkage analysis map genetic regions, or loci, by observing the segregation of a single-gene trait or disease through related individuals (Botstein et al., 1980).

The gene for cystic fibrosis (a Mendelian monogenic disease) was mapped at chromosome 7 in 1985 by linkage analysis (Knowlton et al., 1985), and ultimately found in 1989 (Kerem et al., 1989; Riordan et al., 1989; Rommens et al., 1989). The identification of the Huntington disease gene took even longer. The Huntington disease gene was mapped on chromosome 4 in 1983 (Gusella et al., 1983). However, *HD* was not identified until 1993 (MacDonald et al., 1993).

By 1991, about 1,900 genes were already mapped to specific chromosomes (McKusick, 1991). However, it was estimated that the human genome could have between 50,000 - 100,000 expressed genes based on density and gene sizes (McKusick, 1991). At that time, it was hypothesized that the complete sequencing of the human genome, a technic under development, would be the best way to find the remaining genes and linked them to diseases. The interest of deciphering genes of entire genomes, and mapping genes to their chromosome location, laid the foundations of the Human Genome Project (HGP), the ultimate map (McKusick, 1991), to precisely locate every human gene into a particular region of a human chromosome. The access to that map would have considerably reduced the time and cost that was required to find the cystic fibrosis gene (Tsui and Dorfman, 2013).

2.4 The Human Genome Project

With the sequencing techniques under development, the HGP was a big challenge. It required the contribution of numerous people and considerable financial support, and doing so, the HGP also laid the foundations of international collaboration and data sharing to achieve common goals.

The first draft of the human genome was published in 2001 (Lander et al., 2001; Venter et al., 2001), and the HGP officially ends in 2003 with the reference sequence completed for almost every human chromosome.

The HGP was a monumental achievement, and it changes biology and medicine irrevocably. After the completion of the HGP, we learned that, in fact, in our more than 3 billion base-pair, we "only" have between 20,000 and 25,000 protein-coding genes, far from the 50,000-100,000 proposed a decade before (McKusick, 1991). Since other organisms, including plants such as the sunflower (Badouin et al., 2017), seem to have more protein-coding genes than humans, these findings suggested new layers of complexity.

Nowadays, estimates about the number of genes in the human genome continue to fluctuate, and we still do not know how many genes the human genome actually has (Willyard, 2018). In addition, little is known about the non-coding regions, which accounts for most of the human genome and was once controversially called "junk DNA" (Pennisi, 2012).

Although the completion of the sequencing of a human reference genome was a milestone in the history of science, it was just the beginning.

As a result of the HGP, other international efforts arise, such as ENCODE (or "The Encyclopedia of DNA elements"), which aims to identify the functional elements of the human genome (Encode Project Consortium, 2012), or HapMap (that stands for

“Haplotype Map”), with the goal of determining the common patterns of human genetic variation in the human genome (International HapMap Consortium, 2003).

While focused on around 99.9% of the human genome that is equal between individuals, the HGP revealed millions of genetic variants. Concretely, the initial draft sequence identified around 1.42 million Single Nucleotide Polymorphisms (SNPs) distributed throughout the human genome (Sachidanandam et al., 2001). In this scenario, the international HapMap project raised as a logical next-step (Manolio et al., 2008).

2.5 The International HapMap Project

Launched in 2002, the HapMap project aimed to characterize the 0.1% genetic differences between humans and to boost the understanding of complex diseases that do not follow a full Mendelian inheritance (International HapMap Consortium, 2003).

At the end of the 20th century, the location of genes linked to diseases improves both the diagnosis and the understanding of the pathogenesis for a vast number of monogenic (i.e., Mendelian) diseases. However, due to the inner complexity of common diseases, the reliable results for complex diseases (i.e., non-Mendelian) were anecdotes (Altmuller et al., 2001). Therefore, it was hypothesized that moving the analytical approaches from localized genome sections (reverse genetics; from gene to phenotype) to the whole genome (forward genetics; from phenotype to genes) would lead to a higher yield in the analysis of complex diseases (Risch and Merikangas, 1996; Risch, 2000).

Testing 1 million SNP alleles (i.e., genotyping) in 270 samples from four populations with diverse geographic ancestry, and analyzing patterns of association among SNPs (i.e., linkage), the HapMap project phase I produced a human haplotype (i.e., SNP alleles located in a chromosome that are associated and inherited together more often than expected by chance) map.

Hence, providing a public genome-wide database of common human variation and linkage patterns, HapMap was a shortcut to carry out genome-wide association studies (GWAS), which were unfeasible before the HapMap project completion (International HapMap Consortium, 2005) because, to do so, it was needed a catalog of human genetic variation and the quantification of the correlation structure (Linkage Disequilibrium or LD) of the genetic variants.

2.5.1 Cataloguing the different types of human genetic variation

Unlike mutations, genetic variants occurred in the population at or above a minimal frequency (Kruglyak and Nickerson, 2001). According to its frequencies, genetic variants can be classified as common variants (minor allele frequency (MAF) $\geq$ 5%), low frequency variant (5% < MAF $\geq$ 1%), rare variants (1% < MAF $\geq$ 0.1%), and very rare variants (MAF $\leq$ 0.1%) (Figure 2a).

The sequencing of the human genome estimated 7.1 million of common single nucleotide polymorphisms (SNPs), (i.e., with a MAF of at least 5% across the entire human population) in a 3.2 billion base pair sequence (Kruglyak and Nickerson, 2001). Shortly after that, a significant fraction (1.6 million SNPs) was identified and genotyped among population samples (Hinds et al., 2005; International HapMap Consortium, 2005).

Beyond their frequency, variants can also be classified according to their impact. Broadly, genetic variants can be coding or non-coding if they affect or not a protein-coding gene (Figure 2b). Coding variants tend to have larger effects since they may alter the structure, and therefore the function, of the resulting protein. Not surprisingly, coding variants with large effects are typically rare as a result of purifying selection (International HapMap et al., 2007). In contrast, the effects of non-coding variants tend to be smaller, and the characterization of their consequences is not straightforward. As most of the genome does not contain genes that code for proteins, non-coding variants are the most common (International HapMap et al., 2007).

However, although SNPs are the most abundant form of variation in the human genome (Kruglyak and Nickerson, 2001), there are genetic variations involving more than a single base pair, which can be broadly grouped as “structural variants” (Feuk et al., 2006) (Figure 2c). Based on the patterns of DNA sequence and their sizes, structural variations include insertions and deletions (INDELs), inversions, duplications and translocations of DNA sequences, including copy number variants (CNVs) (Hinds et al., 2005).

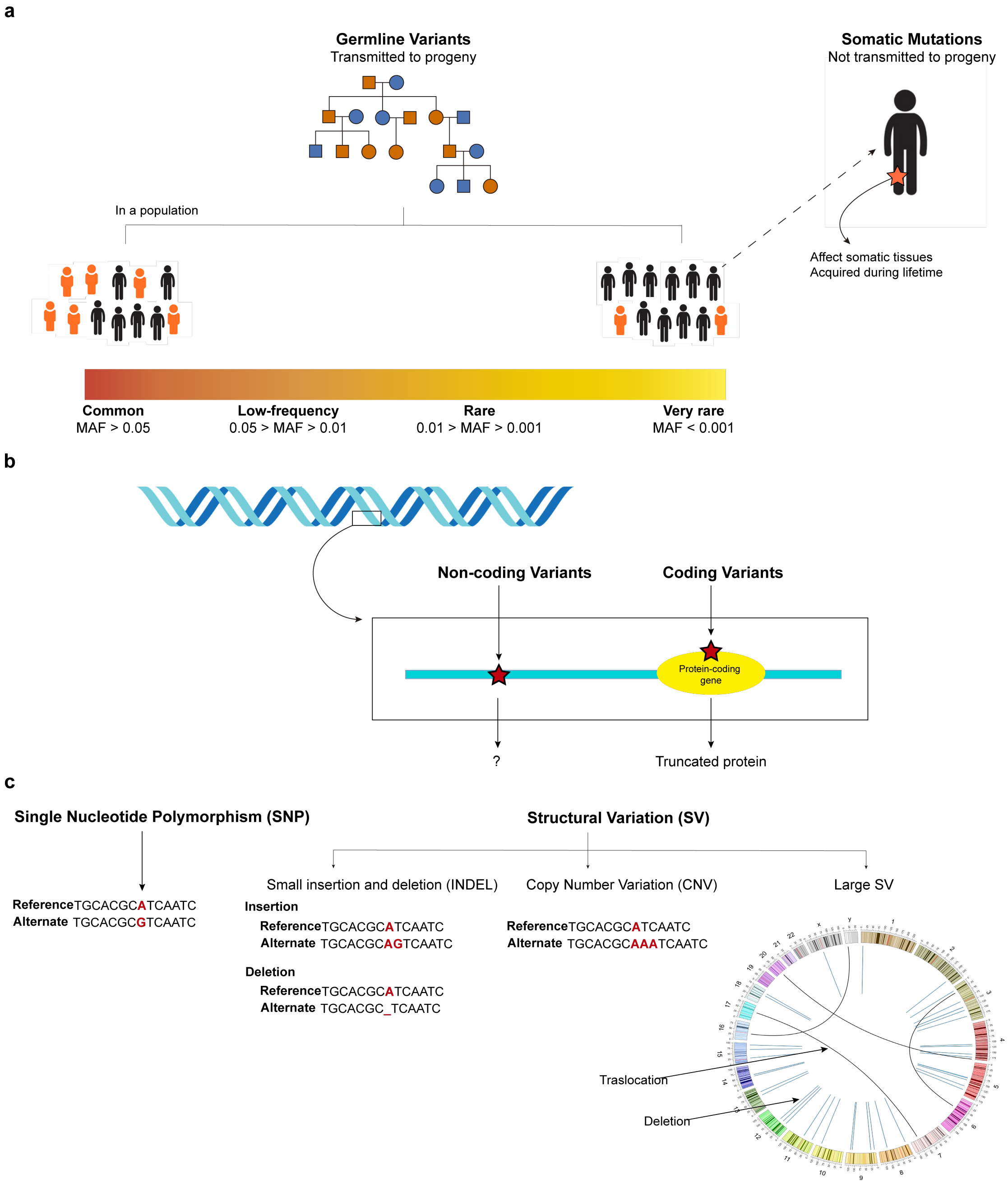


Figure 2. The landscape of human genome variation.

a Genetic variation can be classified according to the affected tissue as germline or somatic. Germline variations are in germ cells and are therefore transferred to the offspring. In contrast, somatic mutations are in somatic cell lines, and can not be transmitted to the progeny. As a result of this, germline variants are found in the population at or above minimal frequency. Although it is a continuum, germline variants are typically classified as common, low-frequency or rare. **b** Genetic can also be classified according to their impact as coding or non-coding if they occur or not in protein-coding genes. Coding variants may truncate proteins, and therefore the interpretation of their consequences is usually more straightforward than the consequences of non-coding variants. **c** According to the DNA sequence pattern and the number of nucleotides involved, genetic variants can be broadly classified as single nucleotide polymorphisms or structural variants. The genomic variation ranges from Single Nucleotide

Polymorphisms (SNPs), affecting only a base pair, to structural variation than can involved few base pairs or large chromosomal rearrangements such as translocations.

2.5.2 Studying the correlation structure of the genome and tag SNPs

HapMap genotyped data allowed the examination of linkage disequilibrium (LD). The LD is defined as a property of genetic variants on a genomic sequence that describes the degree of correlation between two variants within a population (Bush and Moore, 2012). The rationale behind the LD is related to the chromosomal linkage already observed by Bateson, Punnet, and Morgan in the early years of genetics, where genes on chromosomes remain physically attached over generations (Figure 3).

Hence, recombination events break chromosomal segments during meiosis, and all the variants in each of the segments are inherited together (i.e., linked). Therefore, in a population of fixed size and random mating, the chromosomes would be continuously breaking into fragments until all the variants are in linkage equilibrium (i.e., they are independently inherited) (Bush and Moore, 2012). This correlated with the observation that the most ancestral populations (i.e., African-descendent populations) have smaller regions of LD segments than European, Chinese, and Japanese – descendent populations (International HapMap Consortium, 2005).

In genetic analysis, LD is commonly reported as r^2, a statistical measure of correlation. Hence, high r^2 reports a high correlation between variants, and thus only one variant is needed to capture all the allelic information in that segment. These selected variants that capture the variation in the nearby sites are called tag variants (usually “tag SNPs”) (Bush and Moore, 2012).

The HapMap project determined that > 80% of common SNPs in European-descendant populations can be captured using a subset of 500,000 to 1 million SNPs scattered across the genome (Bush and Moore, 2012; Li et al., 2008). Therefore, the tag SNPs that capture most of the common genetic variation in the human genome enabled the production of SNP arrays, and thus the GWAS era began (Visscher et al., 2012).

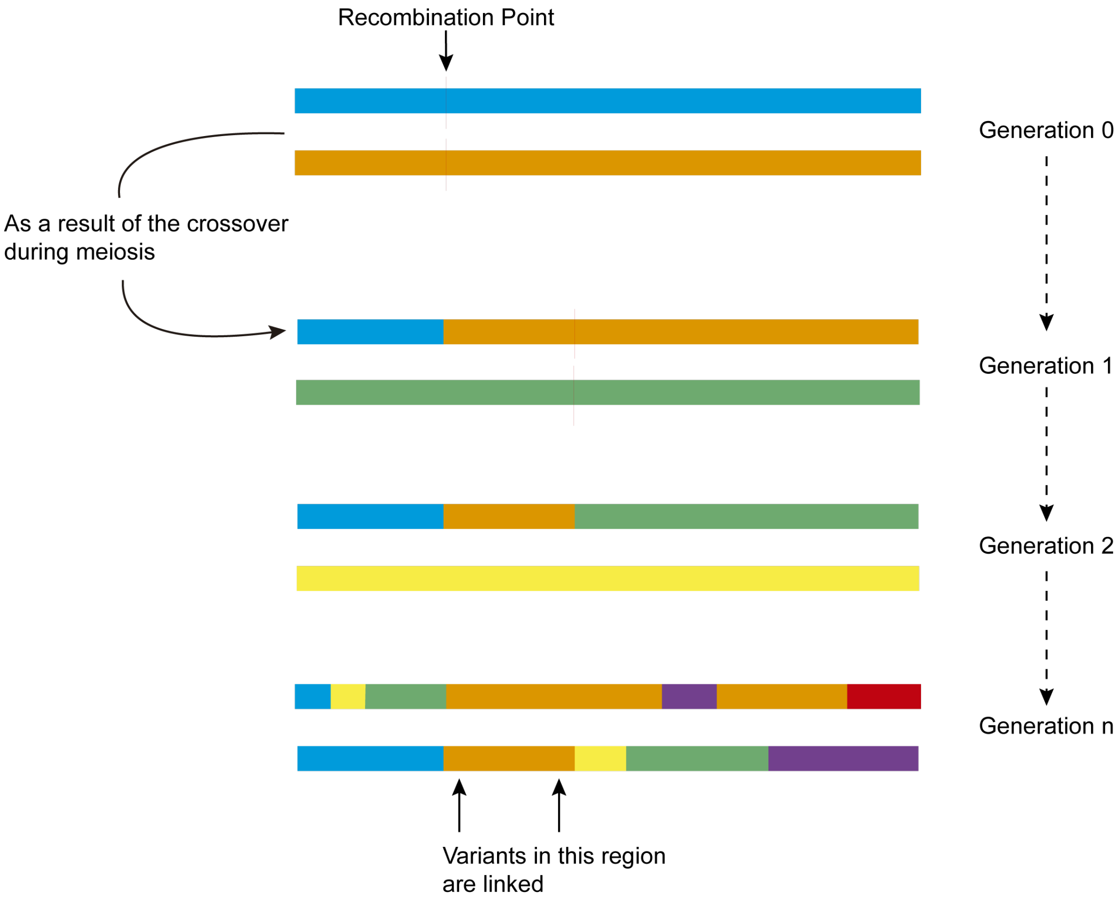

Recombination Point
As a result of the crossover during meiosis
Generation 0
Generation 1
Generation 2
Generation n
Variants in this region are linked

Figure 3. Linkage disequilibrium between variants.

Haplotype blocks in the population are the result of recombination during meiosis. Over generations, chromosomes are continuously breaking into fragments (haplotype blocks), and the variants inside these fragments remain physically attached (i.e., linked).

2.6 The Genome-Wide Association Studies Era

The study of affected families for a particular disease, and examining how the genetic variants segregate with the diseases in multiple families-trees, was successful for rare disorders. By 2003, around 1,400 genes affecting diseases were already known (Botstein and Risch, 2003).

However, the analysis has not fared as well when applied to complex common diseases. The reason is that the underlying mechanisms that influence complex common diseases are different from those causing rare diseases (Hirschhorn and Daly, 2005). By 2003, linkage studies only identified eight associated regions with modest effect sizes (typically between 1.1 and 2) for complex diseases (Lohmueller et al., 2003), in contrast with the around 1,400 genes with large effects found in Mendelian diseases at that time.

Single genes with high penetrance variants usually cause rare diseases, which means that the presence of a particular genetic variation is highly correlated with the presence of the phenotype (i.e., diseases). However, this does not apply to common diseases.

This makes sense in light of natural selection. Hence, deleterious mutations with high effects tend to be rare by purifying selection. In contrast, common diseases have a typically late onset, which means that the genetic variants behind those diseases do not affect, or modestly affect, the reproduction fitness of the individuals carrying them (Altshuler et al., 2008) and, therefore, those variants tend to be common.

For that reason, the first hypothesis was that common diseases were mainly influenced by common variants with low penetrance (common disease/common variant [CD/CV] principle) (Reich and Lander, 2001) (Figure 4). That would explain why the presence of associated variants is not always correlated with the presence of the disease, and this suggested that multiple common variants may influence disease risk with small effect sizes (Bush and Moore, 2012).

These observations pointe out that family-based linkage analyses were not appropriate for complex diseases, and promote the shift toward population-based association studies (Bush and Moore, 2012) using genetic markers to recognize common ancestor linked segments of DNA, or haplotypes (Lander and Botstein, 1986).

Following the HapMap project, the first GWAS, focused on age-related macular degeneration, was published in 2005 (Klein et al., 2005). Nevertheless, the publication of the Wellcome Trust Case Control Consortium (WTCCC) (Wellcome Trust Case Control, 2007) is considered the starting point, since it was the first well-designed GWAS using SNP arrays with high coverage (Visscher et al., 2012).

During the following decade, more than 50,000 associations between genetic variants and complex traits were reported (MacArthur et al., 2017).

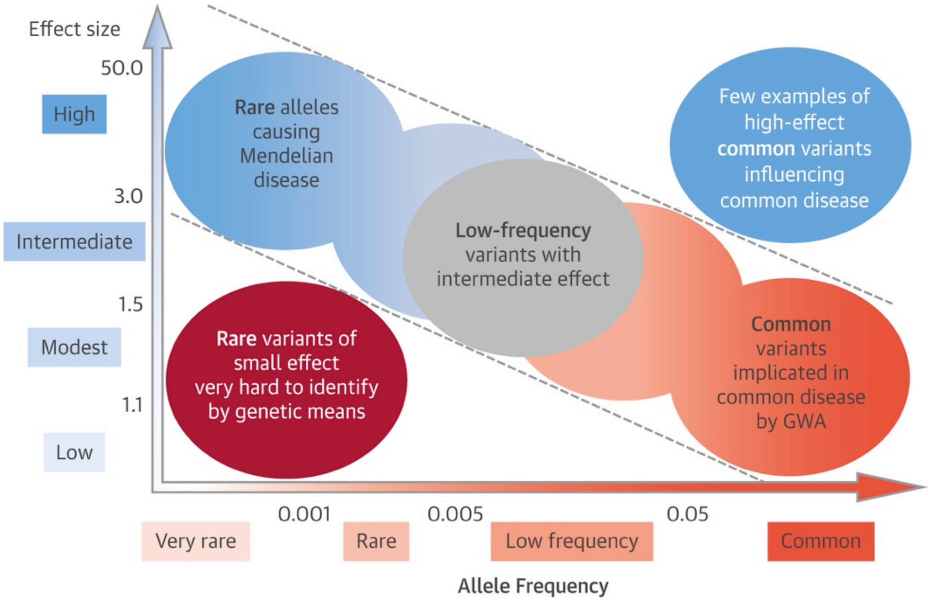

Effect size
50.0
High
3.0
Intermediate
1.5
Modest
1.1
Low
Rare alleles causing Mendelian disease
Few examples of high-effect common variants influencing common disease
Low-frequency variants with intermediate effect
Rare variants of small effect very hard to identify by genetic means
Common variants implicated in common disease by GWA
0.001
0.005
0.05
Very rare
Rare
Low frequency
Common
Allele Frequency

Figure 4. Genetic variants frequencies and their effect size in human disease.

Large effect sizes are expected for rare Mendelian diseases, while the common diseases / common variant hypothesis have been the most accepted for complex diseases with few exceptions. GWA = genome-wide association (Obtained from Assimes and Roberts (2016)).

3 The present: Genome-Wide Association Studies

3.1 What have we learned from GWAS?

GWAS have revolutionized the study of the genetics of complex diseases over the past decade (Visscher et al., 2012; Visscher et al., 2017). Beyond variant-trait associations, many conclusions about the genetic architecture of complex diseases have emerged from GWAS.

For example, it is unambiguously accepted that many loci contribute to the genetic variation for almost any complex disease that has been studied. Therefore, complex diseases are polygenic, and many genes play a role in the final outcome, but also each individual carries a unique pattern of variants, with a mixture of alleles that increase and decrease the disease risk. Hence, since GWAS measure the effect sizes across an averaged background, the effect sizes are found to be small (Visscher et al., 2017). Increasingly larger sample sizes are needed to find new associated regions with small effects (Visscher et al., 2017).

Throughout the years, it has been demonstrated that the number of discovered variants in GWAS is strongly correlated with the sample size (Tam et al., 2019; Visscher et al., 2012). Since 2005, GWAS have evolved from 146 individuals included in the first GWAS (Klein et al., 2005), which found one locus associated with age-related macular degeneration, to some current GWAS including over 1 million of samples (Karlsson Linner et al., 2019), identifying hundreds of associated loci.

Since the number of genetic variants in the human genome is large but finite, and as it was seen in Mendelian diseases (Bamshad et al., 1997; Hodgkin, 1998; Paul, 2000), the same genetic variants contribute to multiple phenotypes (Sivakumaran et al., 2011), a phenomenon known as “pleiotropy” (McKusick, 1976). In addition, studies about the genetic correlation between traits using GWAS results have also demonstrated that many variants affect the same trait consistently (Bulik-Sullivan et al., 2015a; Pickrell et al., 2016).

Despite the success of GWAS identifying genes and pathways involved in complex disease, there is still some level of misunderstanding about the purpose of GWAS and the value of GWAS results. However, many criticisms go beyond the GWAS means, unfairly underestimating its contribution (Visscher et al., 2012). The following section will describe what is the rationale behind GWAS, thus clarifying the interpretation of GWAS results.

3.1.1 What we can expect from GWAS and what we cannot

The general goal of GWAS is to find disease-associated regions to disentangle the genetics behind complex diseases. Very briefly for now, the GWAS rationale is the comparison of allele frequencies between cases and controls for a particular disease, seeking for genetic differences that may explain the phenotypic variance seen in a particular population (i.e., heritability) (Figure 5) (see section 3.3.3 for a detailed explanation of the association test in GWAS).

GWAS do not inform about the specific gene or biological mechanism that links the associated loci with the disease. The association can be both direct (i.e., the associated variant is the causal one) and indirect (i.e., the associated variant is in LD with the causal one) (Bush and Moore, 2012). Hence, “association” does not mean “causality”, and the path from GWAS to biology is not straightforward.

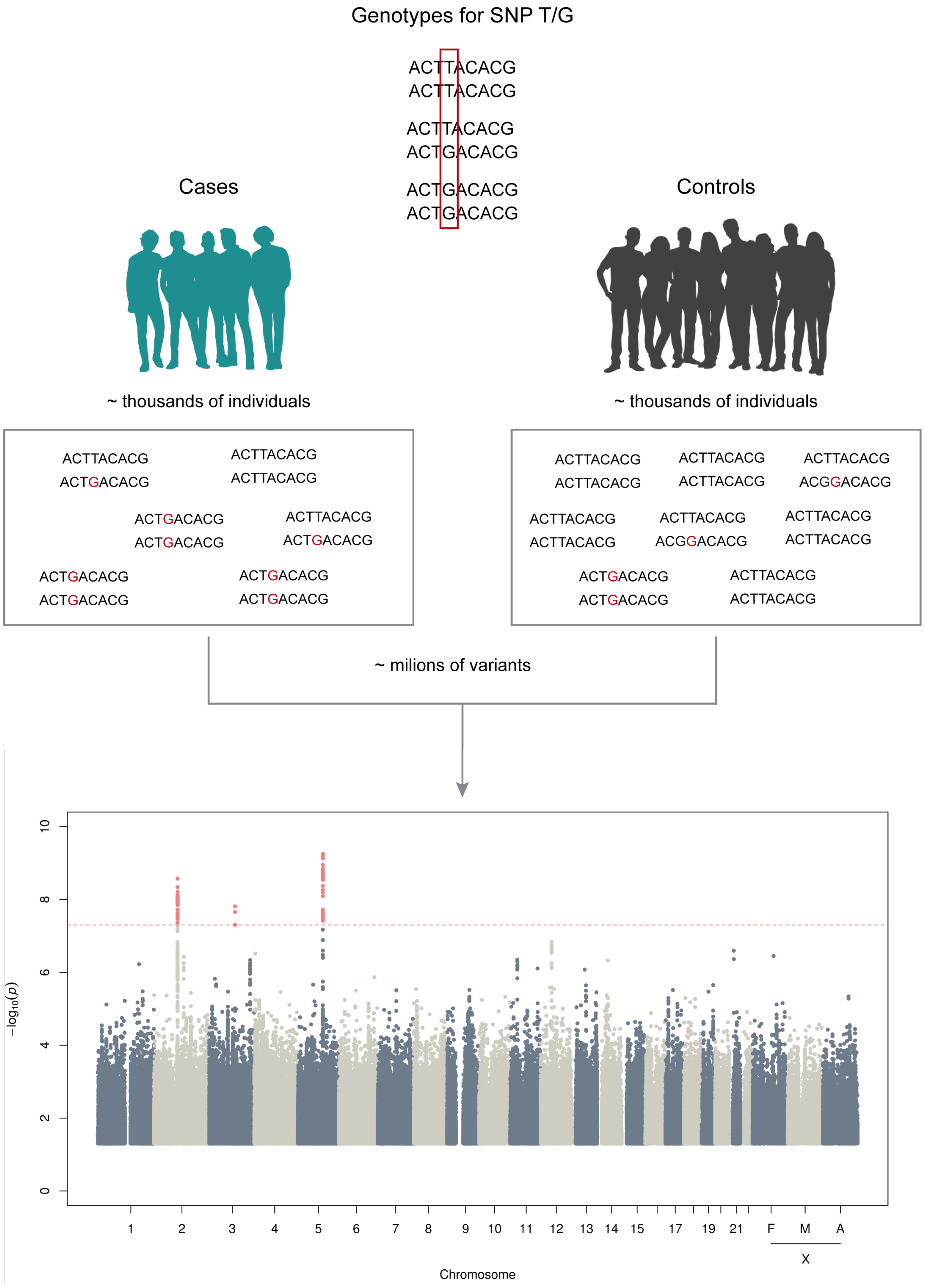


Figure 5. The rationale behind GWAS for complex diseases.

Comparing the allele frequencies between a group of cases and controls, GWAS point out discrepancies in the allele frequencies between the two groups, assessing the allele frequency differences of millions of variants for thousands of individuals. The final plot (i.e., the Manhattan plot), display the –$\log_{10}$(p-value) (y-axis) across the genome (x-axis), highlighting (in red) all the variants above the p-value threshold (in general p-value $< 5 \times 10^{-8}$ for genome-wide analysis), which are considered associated with the disease.

Furthermore, some criticisms argued that GWAS did not explain the heritability observed for complex diseases (Manolio et al., 2009). However, GWAS aim to detect loci-trait associations, not to explain all the genetic variation. Despite explaining the total heritability is not the main objective of GWAS, their findings have contributed to explaining a substantial proportion of the genetic variation behind complex traits. For example, in 2012, the heritability explained for type 2 diabetes, multiple sclerosis and Crohn disease was 10%, 20%, and 20%, respectively, while before GWAS it was essentially zero (Visscher et al., 2012).

GWAS have delivered meaningful biological knowledge, not only theoretical but of utility. Through little more than a decade, GWAS findings have been shown to be useful for preventing diseases through the identification of individuals at higher risk as well as to apply better treatments through classifying and subtyping diseases, delivering biomarkers for diagnosis, and informing drug development and prognosis (Tam et al., 2019; Visscher et al., 2012; Visscher et al., 2017). Hence, powerfully contributing to the emerging field of "personalized medicine" (Hamburg and Collins, 2010).

3.2 Current GWAS scenario: From the HapMap Project to the availability of sequenced catalogs of human genetic variation and biobanks

The scenario in which GWAS are performed today has completely changed compared to the scenario in 2005. The incorporation of High-Performance Computing (HPC) into the genetic field, as well as public initiatives to generate new catalogs of human genetic variation and biobank, has boosted GWAS and constitute the basis of any current GWAS today.

3.2.1 Bioinformatics as an essential tool to study the human genome

Sequencing technologies were under development when the HGP started, and the whole project required millions of dollars and years to reach completion during the earlier 2000s. With the development of the next-generation sequencing (NGS) technologies, sequencing a human genome in 2019 cost $1,301 (Wetterstrand, 2019) and takes less than a week (Gauthier et al., 2018).

Computation and biology are nowadays highly connected. "Bioinformatics", broadly defined as the discipline to develop and use computational tools to answer biological questions, is now needed in almost any field in biology and is essential in genomics.

Moreover, the massive genotyping and sequencing of thousands of genomes and the accumulation of their analytical results stressed the need for large-scale storage. Not surprisingly, genomics is considered today a Big Data science and has been defined as a “four-headed beast”, where acquisition, storage, distribution, and analysis are highly computational demanding (Stephens et al., 2015). For that reason, HPC was incorporated into the genetic field. HPC, such as clusters and clouds, allows the execution of workflows (sequential tasks involving multiple tools) taking profit of the parallelism in distributed computational environments (Spjuth et al., 2015).

GWAS are not an exception to this. The combination of whole-genome sequencing (WGS) data and genotyping arrays with the increasing sample sizes and the number of variants to be analyzed in current GWAS, makes HPC essential. Hence, the study of genomes, sequenced or genotyped, ranging from one genome to thousands, becomes feasible only with computers.

3.2.2 Data availability: data sharing and “biobanks”

Another direct benefit from the HGP was the popularization of data sharing. The HGP was about sharing data immediately among the participating groups to achieve the final goal of sequencing the human genome, as well as making the results from the project publicly available (Hood and Rowen, 2013).

Although sharing data is not yet the general way of proceeding, it is increasingly becoming popular to share the data, results, and computational codes with the community.

Several initiatives have recently emerged promoting data sharing, such as the database of Genotype and Phenotypes (dbGAP) (Tryka et al., 2014) and the European Genome-Phenome Archive (EGA) (Lappalainen et al., 2015). The availability of these resources boosts research by making data accessible to third investigators, who can apply new methodologies to existing data. Although some researchers complained about data sharing, even calling third investigators “research parasites” (Longo and Drazen, 2016), the whole society benefits from it. Besides, if public funding has been used to generate new genomic data, after an embargo time to give enough credit to those who generated the data, and always taking into account ethical concerns, data must be of public use.

GWAS can be largely benefited from data sharing, increasing discoveries with little cost. As demonstrated in a recent study, the re-analysis of publicly available cohorts for type 2 diabetes, including 70K individuals, identified seven new associated loci (Bonas-Guarch et al., 2018). In 2014, 924 publications registered a secondary use of dbGaP data, and 25% of them appeared in journals with an impact factor greater than 10 (Paltoo et al., 2014).

Moreover, the access to the increasing number of catalogs of human genetic variation, such as the HapMap project (International HapMap Consortium, 2003, 2005), or the newer sequencing-based ones, such as the 1000 Genome Project (1000G) (1000 Genomes Project Consortium et al., 2010; 1000 Genomes Project Consortium et al., 2012; 1000 Genomes Project Consortium et al., 2015; Sudmant et al., 2015), The UK10K project (UK10K Consortium et al., 2015), the GoNL project (The Genome of the Netherlands Consortium, 2014) or the Haplotype Reference Consortium (HRC) (McCarthy et al., 2016), enables the analysis of millions of genetic variation for a broad frequency spectrum in GWAS (Table 2)

Table 2. The evolution of reference panels of human genetic variation.

Cohort	Release	Sample Size	Depth	Number of variants	Ancestry
HapMap	2005	269	Genotyped	1.1M	Multi-Ethnic (4 populations)
HapMap 2	2007	270	Genotyped	3.8M	Multi-Ethnic (4 populations)
HapMap 3	2008	1,115	Genotyped	1.6M	Multi-Ethnic (11 populations)
1000G phase 1	2012	1,029	5x/Exome	39.5M	Multi-Ethnic (14 populations)
UK10K	2015	3,781	7x/Exomes	26.6M	European (British)
GoNL	2014	250	13x	21.6M	European (Dutch)
1000G phase 3	2015	2,504	7.4x/Exome	88M	Multi-Ethnic (26 populations)
HRC	2016	32,488	Diverse	39,2M	Multi-Ethnic (mainly European)
TopMED	2019?	53,831	30x	240M	Multi-Ethnic (60% non-European)

The availability of these catalogs of human genetic variation based on sequenced data have taken GWAS one step further. Hence, even though GWAS were designed to test common SNPs, today GWAS can identify SNPs at lower frequencies as well as small structural variants that contribute to susceptibility (Tam et al., 2019).

In addition to the availability of these sequence-based catalogs, population-based "biobanks" have recently emerged to link genetics and epidemiological factors to diseases risk (Bahcall, 2018). These long-term prospective cohorts contained genetic data associated with extensive phenotypic information for hundreds of thousands of individuals (Price et al., 2015). As state in previous sections, GWAS will be directly benefited from such larger sample sizes. Moreover, since biobanks collect data in a more clinical context than in case-control studies, they make possible more accurate analyses about side effects and diseases interaction or comorbidities (Price et al., 2015). Enrolling 500,000 volunteers, UK Biobank is the first publicly available biobank with genetic data (Bycroft et al., 2018), and has already demonstrated its utility from the very beginning (Elliott et al., 2018). UK Biobank's genetic information is based on genotyping arrays. However, it was recently announced that UK Biobank will sequence the 500,000 volunteers as well (UK Biobank, 2019).

Besides UK Biobank, a new initiative will integrate WGS data from more than 100,000 individuals from different populations with extensive phenotypic information: The Trans-Omics for Precision Medicine (TOPMed) program. Doing so, TOPMed aims to improve our understanding of the genetic architecture and disease biology of heart, lung, blood, and sleep disorders (Taliun et al., 2019).

Other biobanks have been created, such as the China Kadoorie Biobank (Chen et al., 2011), the Biobank Japan (Nagai et al., 2017), the Finngen Project (Tabassum et al., 2019), the Estonian Biobank (Leitsalu et al., 2015), the Million Veteran Program biobank (Gaziano et al., 2016), the All of Us Research Program biobank from the National Institute of Health (NIH), the BioVU from the Vanderbilt University, and the BioMe biobank from the Icahn School of Medicine at Mount Sinai.

In addition, large cohorts have also been created and are of public domain, such as The Resource for Genetic Epidemiology Research on Aging (GERA) (Hoffmann et al., 2017), available at dbGaP, which integrates individual data from 78,419 participants of diverse ancestry and 22 diseases.

Furthermore, beyond the direct benefits of genotyped and sequenced data from databases repositories, GWAS can also be largely benefited from interactive portals. Projects such as the Genome Browser (Kent et al., 2002), the NIH Roadmap Epigenomics Mapping Consortium (Bernstein et al., 2010), The Genotype-Tissue Expression (GTEx) project (GTEx Consortium, 2013) or the eQTLgen (Võsa et al., 2018), allow the integration of –omics data for a better interpretation of the underlying mechanisms behind the variants identified by GWAS. In addition, disease-specific portals, such as the T2D Knowledge Portal, the Cardiovascular Disease Knowledge Portal, Cerebrovascular Disease Knowledge Portal and Sleep Disorder Knowledge Portal from the Knowledge Portal Network, or the Oxford Brain Imaging Genetics (BIG) Server (Elliott et al., 2018), which includes GWAS results from UK Biobank for multiple disease, allow browsing human genetic data.

Therefore, making data accessible to the broad research community enhances collaboration, which is essential for the rapid discovery of new associations between genetic variants and diseases as well as to fill the gap between genetics and clinical outcomes.

However, despite the benefits of data sharing for the entire community, the majority of research initiatives remain reluctant to make data accessible, including many of the projects already mentioned in this section, thus limiting the scientific progress.

3.3 Current GWAS workflow

GWAS have taken great advantage of the changes and improvements made in the field during the following years after the completion of the human genome.

GWAS have mainly evolved towards increasing the sample size to gain statistical power to detect new signals and introducing a higher number of variants with different frequencies and types, such as INDELs, to achieve a better understanding of the genetic architecture of complex traits. Moreover, and taking into account the lessons learned from population genetics, GWAS have increased their accuracy and reliability implementing methods to avoid false positives and spurious associations. Several protocols and techniques have been developed to do so, and the following sections will describe the pillars of any current GWAS.

3.3.1 Pre-GWAS Quality Control to avoid systematic bias

Many errors can emerge if the input data of a GWAS (in general, genotyping data) is not accurate since it will introduce a systematic bias. Therefore, data have to be quality controlled (QCed) to avoid errors at the level of both variants and individuals before GWAS.

The main purpose of the QC is to remove any systematic difference between controls and cases that may lead to false associations compromising the true ones (confounding) (Anderson et al., 2010). A clear situation where this occurs is when cases and controls are from different populations. Population stratification is a major source of spurious associations (Anderson et al., 2010; Campbell et al., 2005; Price et al., 2006; Zeng et al., 2015). Since genetic variant depends on the mutation rate, recombination, and immigration (Griffiths, 2000), people from nearby geographical areas have a similar profile of genetic variants. Therefore, if we compare cases from a particular population against controls from a different one, most of the associations will be due to population structure, and will not be related to the disease or trait of interest. Conventional methods to identify and remove individuals with divergent ancestries are principal component analysis (PCA) and multidimensional scaling (MDS) (Anderson et al., 2010; Zeng et al., 2015).

Moreover, individuals with a high proportion of variants unsuccessfully genotyped (missing rates > 1-5%) (Zeng et al., 2015), duplicated or hidden related individuals, and the discordance between the estimated and the reported sex information, which may indicate sample mix-ups, have to be carefully assessed before GWAS.

At the variant level, variants with a high proportion of individuals without called genotypes are removed. Similarly, variants with MAF < 0.1% are removed as well. Moreover, variants that show a large deviation from HWE are excluded. However, some departure from HWE is expected from variants in cases that are truly associated with the diseases (Wittke-Thompson et al., 2005). Therefore, only controls are usually filtered (e.g., HWE controls p-value $< 1 \times 10^{-6}$) (Anderson et al., 2010; Zeng et al., 2015). Besides, if more than one disease is available in the cohort, a looser threshold for the whole cohort is recommended (e.g., HWE cohort p-value $< 1 \times 10^{-20}$) (Bonas-Guarch et al., 2018).

3.3.2 Two-step genotype imputation to increase the power of GWAS

GWAS are based in commercial genotyping arrays that mainly typed common variation. In the earlier 2000s, it was thought that this would be sufficient, hypothesizing than common variants are the cause of common diseases following the CD/CV principle (Reich and Lander, 2001). However, it has been suggested that rare variants may have greater consequences than common variants (Bodmer and Bonilla, 2008; Pritchard, 2001; Pritchard and Cox, 2002; Schork et al., 2009), thus explaining part of the missing heritability (Gibson, 2012). Although rare variants can be analyzed using WGS given sufficient coverage (e.g., > 20x) (Alioto et al., 2015; Cirulli and Goldstein, 2010), again, its use is still prohibitively expensive (Goodwin et al., 2016). Nowadays, the most cost-efficient strategy to analyze rare variants (as rare as MAF = 0.001) is genotype imputation (Li et al., 2009; McCarthy et al., 2016), a method of estimating genotypes that have not been directly genotyped using reference haplotypes (Figure 6).

Briefly, imputation methods are based on the principle behind identity by descent (IBD) to identify related individuals, where shared haplotype blocks are used to describe the genealogical tree. Therefore, imputation methods identify shared haplotypes between the study individuals and the haplotypes in a reference set, and use these shared haplotypes to impute the missing alleles in the study individuals (Marchini and Howie, 2010) (Figure 6).

Hence, genotype imputation fills the gaps in the genotypes from SNP arrays adding more variants, thus increases the chances of finding a causal variant in GWAS. Besides, imputing genotypes is useful for fine mapping since imputation provides a higher resolution of the genomic regions. In addition, genotype imputation is useful for meta-analysis since it facilitates the combination of results across studies with different genotyping arrays generating a common set of variants.

Imputation has been a key step in GWAS (Marchini and Howie, 2010). Since the first studies using genotype imputation were published (Scott et al., 2007), many tools have been developed for genotype imputation, including IMPUTE (Bycroft et al., 2018; Howie et al., 2012; Howie et al., 2011; Howie et al., 2009; Marchini and Howie, 2010; Marchini et al., 2007) and MINIMAC (Das et al., 2016) among the most popular.

Nevertheless, genotype imputation is highly computational demanding. For that reason, the concept of "pre-phasing" was introduced in 2012, demonstrating that estimating (i.e., phasing) the haplotypes for each individual prior imputation reduces the computational cost without compromising the accuracy (Howie et al., 2012). To that end, phasing tools, such as SHAPEIT (Delaneau et al., 2013a; Delaneau et al., 2014; Delaneau et al., 2011; Delaneau et al., 2013b; O'Connell et al., 2014) and Eagle (Loh et al., 2016a; Loh et al., 2016b), have been developed.

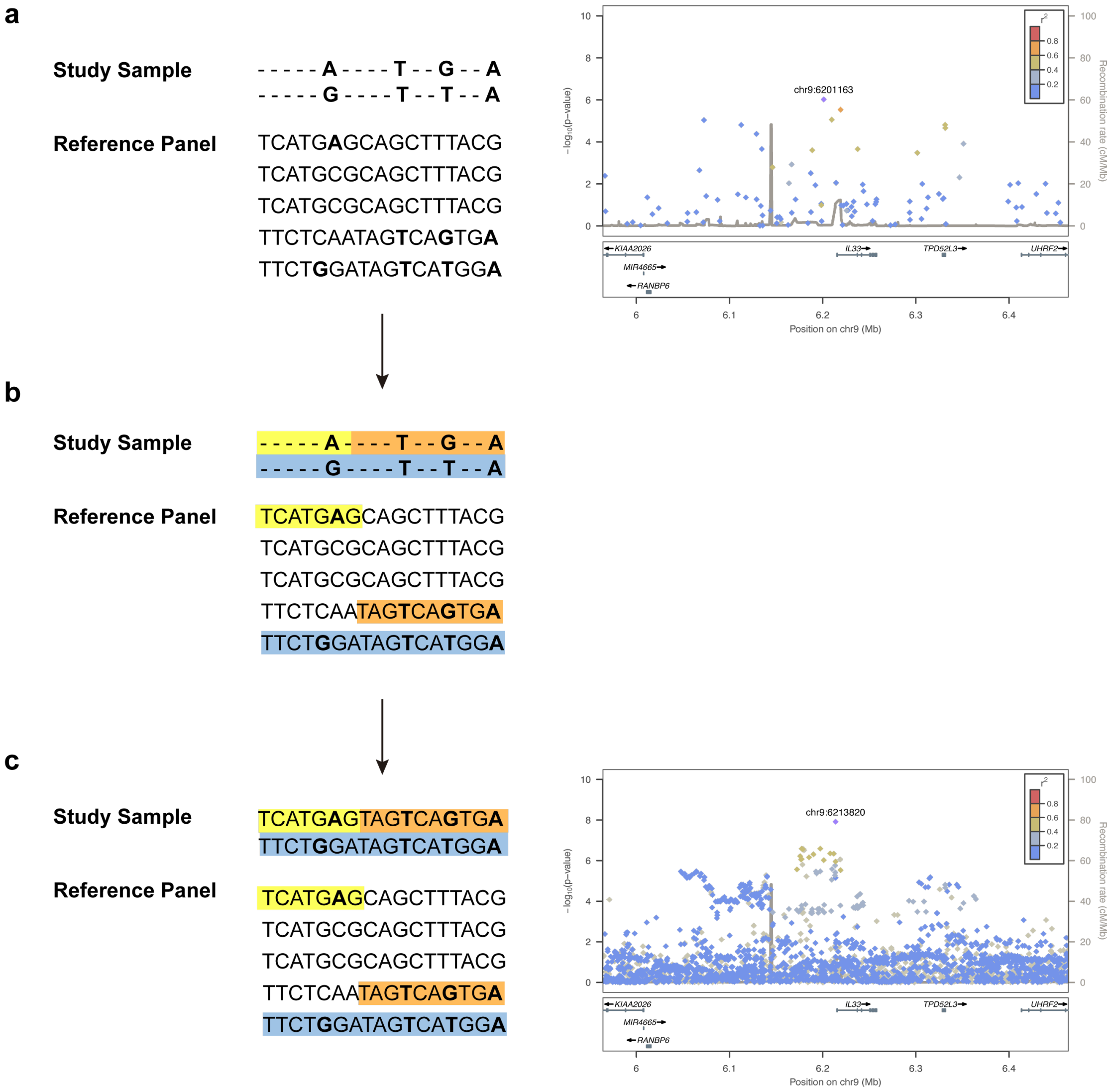


Figure 6. The rationale behind genotype imputation.

a Using the tag SNPs from the genotype data as a backbone, **b** and matching the haplotypes with those in a sequenced-based reference panel, **c** non-genotyped variants are inferred. The graphs show the increased resolution for a particular genomic region after genotype imputation, which was crucial to find this region as associated with genome-wide significance ($p < 5 \times 10^{-8}$).

Moreover, and beyond technological and methodological improvements (Das et al., 2016; Howie et al., 2012; van Leeuwen et al., 2015), genotype imputation has been benefitted notably from publicly available datasets over the years. The first widespread reference panel used for genotype imputation was HapMap 2, but over the years, it has been surpassed by sequence-based reference panels such as 1000G, UK10K, GoNL, HRC and soon TopMED (Table 2). The development of NGS technologies has rapidly increased the size of the reference panels that are continuously growing, ranging from 210 genotyped individuals in HapMap 2 (Huang et al., 2009) to the expected 100,00 sequenced individuals of TopMED. Therefore, genotype imputation using sequence-based reference panels enables GWAS to identify associations with rare variants using common SNP arrays (Das et al., 2018). Of note, nowadays, imputation using large reference panels is reasonably accurate for variants with MAFs as low as 0.1% (McCarthy et al., 2016).

Overall, current GWAS are commonly based on SNP arrays combined with genotype imputation using sequence-based reference panels. As the WGS cost decreases and more complete catalogs of genetic variation are released, including more variants and populations, the imputation of genotypes is constantly improving GWAS in a cost-effective manner (Tam et al., 2019).

3.3.3 The association testing

The core of GWAS when analyzing a disease is the comparison of allele frequencies between unrelated groups of balanced cases and controls. If the genetic background for both groups is the same (e.g., population), any frequency difference is assumed to be due to the disease status and is, therefore, associated with it. Hence, GWAS systematically analyze allele frequency discrepancies between groups all over the genome (hypothesis-free) to point out a genomic region associated with the diseases under study. A GWAS is a series of single-locus tests, examining each variant independently (Bush and Moore, 2012).

The necessary statistics proceeds by analyzing each SNP individually to test the null hypothesis of no association ($H_{0:}$ β_i=0) (Zeng et al., 2015). However, the statistical tests are different for quantitative (continuous) or dichotomous (i.e., case/control) traits. Hence, for quantitative traits, such as height or blood pressure, the standard method is, in general, the linear regression following the equation

$$Y_{ij} = \beta_0 + \beta_{ij} X_{ij} \, ,$$

where for a given SNP (j = 1, 2, 3, …, m) on a particular individual (i = 1, 2, 3, …, n), Y_{ij} is the trait value for the individual *i,* and X_{ij} is 1 if the individual has the effect allele *A* or 0 otherwise. Therefore, it tests the mean differences between *A* and *not-A* individuals, thus comparing the trend in the trait to the trend in the genotypes for each marker.

However, for case-control studies, the common statistical method is logistic regression since linear regression cannot be applied directly to the case-control status (Balding, 2006). Therefore, logistic regression estimates the probability of an outcome when it is binary, such as disease states.

Where the possible genotypes are AA, Aa, and aa (X_{ij}), and "a" is the minor allele (lowest frequency), and "A" is the major allele (highest frequency), the logistic regression is

$$logit(p_{ij}) = log\left(p_{ij}(1 - p_{ij})\right) = \beta_0 + \beta_{ij}X_{ij},$$

where p_{ij} is the disease risk (probability) for the *j* SNP on the *i* individual.

To reduce spurious associations, the association test should be adjusted for known influential factors such as age, sex or clinical measures such as body mass index (BMI), and logistic regression can accommodate these covariates. Besides, adjusting for genetic principal components is one of the most important covariates to take into account in order to adjust for any underlying population substructure (Balding, 2006; Bush and Moore, 2012).

When a logistic regression is calculated, the exponential function of the regression coefficient ($e^{\beta1}$) is the odds ratio (OR) associated with a one-unit increase in the exposure. The ORs are used to compare the relative odds of the occurrence of the outcome (e.g., disease), given exposure to the variable of interest (e.g., genotype). Therefore, the odds determine whether a particular exposure is a risk factor for a particular outcome and can be used to compare the magnitude of various risk factors for that outcome (Szumilas, 2010). Hence,

OR=1; Exposure does not affect odds of disease
OR>1; Exposure associated with higher odds of the disease
OR<1; Exposure associated with lower odds of the disease

In addition to the effect size (OR for logistic regression or β for linear regression) and the standard error, which explains the variability in the β estimate, the basic output of a test of association includes the p-value to asses the significance of the association.

Even after adjusting with known influential factors adding covariates to the association test, a fundamental problem in GWAS has always been how to distinguish between true and false associations. Thus, taking into account the size of the human genome, many associations could be found by chance (Balding, 2006). Moreover, genotyping errors, cryptic relationship between individuals and unrecognized population stratification can lead to spurious associations (Campbell et al., 2005; de Bakker et al., 2008; Hinrichs et al., 2009; Price et al., 2006; Wittke-Thompson et al., 2005; Yang et al., 2011) even though after the QC. Luckily, many approaches have been developed to address these concerns and ensure the GWAS results reliability.

3.3.4 Minimizing false positive associations.

There is a serious multiple comparison problem in GWAS than can inflate the Type I errors (false positives) if no measure is taken (Zeng et al., 2015). For instance, setting a significance level (p-value) α = 0.05, which means that 5% of the time the null hypothesis will be rejected when it is true (false positive), if 500,000 SNPs are tested for association, it is expected that about 25,000 false positive will be observed by chance.

In other to solve this, Bonferroni correction offers a way to control the Type I error inflation by dividing α by the number of independent tests (Zeng et al., 2015). Hence, after the completion of the HapMap project in 2005, it was estimated that the number of common (MAF > 5%) independent variants were 150 per 500 kilobase pairs (kb) in European, Japanese and Chinese population (International HapMap Consortium, 2005). Taking into account the whole genome (~3.3 Gb), this suggests a p-value threshold of 5×10^{-8} for these populations. However, some authors argued that the Bonferroni correction is too conservative (Bush and Moore, 2012) (Balding, 2006; Zeng et al., 2015) as it assumes that each association test for each variant is independent, an assumption that is in general untrue due to de LD among variants (Bush and Moore, 2012). However, some others suggested more stringent p-value for low frequency and rare variants recently added to GWAS (Fadista et al., 2016), but it is a discussion that remains unsolved. Nevertheless, the standard GWAS significant threshold has been established at 5×10^{-8} (Welter et al., 2014). Hence, any variant with a p-value lower than the adjusted significance level (p-value $< 5 \times 10^{-8}$) is considered to be associated with genome-wide significance.

Furthermore, beyond minimizing spurious associations through Bonferroni correction, any associated variant must also be replicated in independent cohorts when possible (Chanock et al., 2007; Price et al., 2015; Zeng et al., 2015). To do so, an equivalent independent cohort must be tested following the same association analysis to ensure consistency (Zeng et al., 2015). To this end, several criteria to ensure proper replication has been established (Chanock et al., 2007), including the need for enough sample size for the same population and identical phenotype criteria (Bush and Moore, 2012). However, it is challenging to find multiple studies that match perfectly. Nevertheless, even though there are many reasons for non-replication (Chanock et al., 2007; Zeng et al., 2015), any replicated variant will have additional evidence of its association.

The analysis of multiple GWAS results through meta-analysis (Evangelou and Ioannidis, 2013) is also a useful approach to examine and refine the significance and effects of the original study (Bush and Moore, 2012). Since only statistical results are needed to perform a meta-analysis (i.e., there is no need to share sensitive data), large consortia have been benefited from meta-analysis, increasing the sample size, and therefore the statistical power. However, as it happens with replication, all the studies in a meta-analysis have to examine the same hypothesis, and a perfect match is difficult to ensure. For that reason, meta-analysis usually quantifies the heterogeneity between studies as a guide to understand the different results from different studies. A popular measure to study the heterogeneity is the I^2, and $I^2>75$ are considered high (Bush and Moore, 2012). Several tools are available for meta-analysis, being METAL (Willer et al., 2010) among the most popular.

An additional measure of quality in GWAS is the “genomic control” (GC) (Yang et al., 2011). The logic behind GC relies on the fact that only a small fraction of variants are truly associated with the disease. Hence, most of the variants should follow the distribution under the null hypothesis of no association. Therefore, the observed median value of the χ^2 statistic divided by the expected median value of the χ^2 statistic (~0.456 for 1 degree of freedom) is the inflation factor or lambda (λ) (Hinrichs et al., 2009; Zeng et al., 2015) as shown in the following equation:

$$\lambda = \frac{median(\chi^2)}{0.456}$$

If $\lambda \leq 1$, there is no inflation and no adjustment is necessary. However, if $\lambda > 1$, all the following χ^2 statistics for the candidate variants have to be divided by λ (Hinrichs et al., 2009). In general, λ is used as an indicator of systematic bias, since it will be inflated if there are differences in the alleles frequencies due to unrecognized population stratification, cryptic relatedness and genotyping artifacts (de Bakker et al., 2008; Yang et al., 2011).

However, λ can also be inflated due to true associations in GWAS with large sample sizes (Lango Allen et al., 2010; Speliotes et al., 2010). A solution to this is the LD Score regression, which allows distinguishing between confounding biases or polygenicity (Bulik-Sullivan et al., 2015b). In fact, it has been demonstrated that polygenicity accounts for the majority of observed genomic inflation in large GWAS cohorts using the LD Score regression (Bulik-Sullivan et al., 2015b).

3.3.5 Graphical representation of GWAS results

Due to the inner complexity of GWAS results, visual presentation is undoubtedly useful to facilitate the interpretation of the results from a GWAS (Zeng et al., 2015). Such is the case that GWAS are always accompanied by graphical representations (Figure 7). Besides, graphs allow the visualization of any possible issue that could compromise the reliability of the results and needs to be addressed. Hence, many tools have been developed to generate quantile-quantile (Q-Q) plots and Manhattan plots (Turner, 2018) as well as regional association plots (Pruim et al., 2010).

A Q-Q plot is focused on discarding any systematic bias and is related to the GC described in the previous section. The Q-Q plot displays the observed distribution of the p-values to test if it follows the expected (null) distribution (Zeng et al., 2015), and it is always accompanied by the inflation factor λ (Figure 7a).

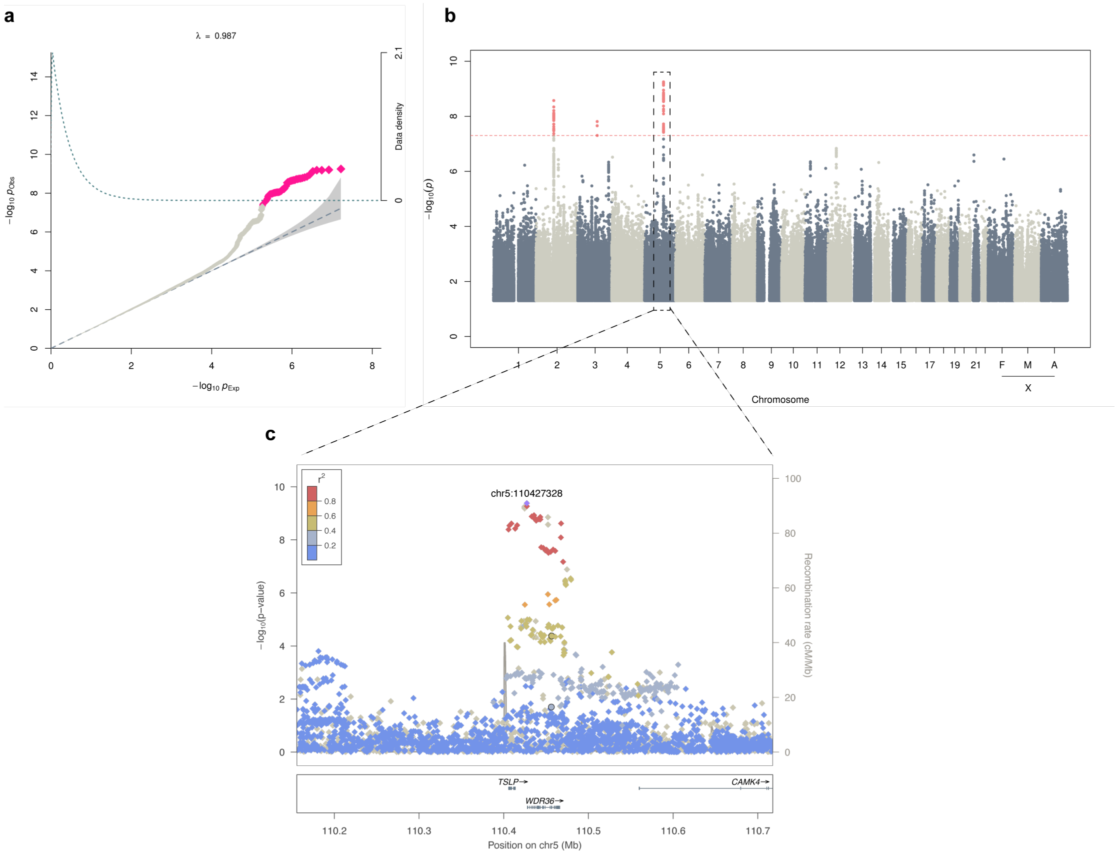


Figure 7. The typical graphical representations in GWAS.

a Q-Q plot showing the expected $-\log_{10}$ p-value under the null hypothesis of no association in the x-axis versus the observed $-\log_{10}$ p-value in the y-axis. The genomic inflation factor λ accompanied the plot as a measure of the extent of the false-positive rate. **b** Manhattan Plot highlighting the GWAS findings across the entire genome. The chromosomes are displayed along the x-axis, while the y-axis represents the $-\log_{10}$ p-value from the association test. Each variant (dot) above the red line (p-value threshold, and typically 5×10^{-8}) is considered a region of interest for further analysis. **c** An example of a regional association plot for the associated loci at chromosome 5. The x-axis represents the chromosomal location, and the y-axis shows the $-\log_{10}$ p-value from the association test. The colors displayed in the legend quantify and represents the linkage structure of the region.

In addition, after the association test of millions of variants, Manhattan plots facilitate the visualization of GWAS hits across the whole genome (Gibson, 2010). Briefly, a Manhattan plot is a scatter plot representing the $-\log_{10}$ p-values against chromosomal location (Zeng et al., 2015), named “Manhattan” as its typical shape for the whole genome reminds the skyline of Manhattan (Alder and Kass, 2017) (Figure 7b).

Finally, to properly inspect associated regions, graphical representation of locus-specific association results have been implemented in user-friendly tools such as LocusZoom (Pruim et al., 2010). These plots allow the visual inspection of the strength of the association and its extent as they include the LD information with nearby variants (Figure 7c). Moreover, it includes the position of the associated variants relative to genes, and it can also display previously described associations.

3.4 Future work needed in GWAS

Beyond continuously enlarging the sample sizes to gain statistical power to detect new associations, there are many gaps in current GWAS strategies and workflows that need to be addressed.

From strategic aspects, such as including more diversity in GWAS, to implementing innovative analyzes for existing and ongoing data, there is still a long way to go to fully squeeze the potential of GWAS.

3.4.1 The lack of diversity and its impact in GWAS findings

GWAS are biased towards wealthy European-descendant populations for both sample collection and authors. As a starting point, even the genotyping array designs are primarily based on European ancestry (Schaid et al., 2018).

Recently, a scientometric study reviewed GWAS publications from 2005 to 2018, and pointed out potential gaps in the current research that may impact GWAS findings (Mills and Rahal, 2019). In particular, participants of European populations account for 86.03% discovery, 76.69% replication, and 83.19% combined in GWAS.

Not representing human diversity have an obvious impact on research findings since it will not provide globalize prevention and therapeutic targets. Therefore, the field is increasingly aware of that problem (Bustamante et al., 2011; Editorial, 2017; Guglielmi, 2019a; Guglielmi, 2019b; Lariviere et al., 2013; Popejoy and Fullerton, 2016; Tam et al., 2019), and undergoing initiatives like TopMED have a more varied ancestry profile, with ~60% individuals of non-European ancestry (Taliun et al., 2019). However, the diversity in GWAS ancestry may decrease even further with the release of UK Biobank (94.23% participants from European ancestry) or the availability of 23andMe data (77% of European ancestry) (Mills and Rahal, 2019).

The study of diverse populations has demonstrated to be extremely useful in identifying critical variants (Wojcik et al., 2019). The advantage of not only analyze non-European populations is especially important for rare variants that are likely to be population-specific (Gravel et al., 2011). In fact, isolated-populations with higher frequencies for rare variants (including European-descendant populations such as the Finns) boosted the discovery of new associations with clinical relevance (Estrada et al., 2014; Gudmundsson et al., 2012; Han et al., 2016; Kenny et al., 2012; Manning et al., 2017; Moltke et al., 2014; Sidore et al., 2015).

On top of that, not representing human diversity leads to increased disparities in disease prevalence and healthcare.

3.4.2 Including non-additive models of inheritance

For both quantitative and qualitative traits, there are a variety of ways to encode the genotypes for the association testing, and the choice of how to encode the data can have implications for the statistical power (Bush and Moore, 2012). Hence, the association test examines the association between genotypic groups (*AA*, *Aa*, *aa*) and the phenotype.

The genotypes can be grouped in models, and each model makes a different assumptions about the genetic effect, such as additive (*AA* = 2*k* and *Aa* =*k*), dominant (*AA* and *Aa* > risk (*k)* than *aa*), recessive (*AA* > *k* than *Aa* and *aa*), heterodominant or heterozygote (*Aa* > *k* than *AA* and *aa*) and genotypic or general (which is parameterized with an additive effect and a heterozygote effect) (Table 3).

A general practice in GWAS is to examine the additive model only assuming a uniform, linear increased risk. As explained, if *A* is the allele of risk (*k*), the genotype *aa* has no risk, *Aa* genotype has *k* risk, and *AA* doubles its risk (2*k*). The usual choice of the additive model as the unique way of encoding the genotypes is partially explained by the fact that, for most common associations, the genetic model of inheritance is unknown a priori, and the additive model can capture most of the signals, especially for variants with dominant effect. Hence, typically the choice of a particular model of inheritance is not discussed since the additive model can capture most of the signals for common variants even in the case of non-additive effect, and global estimates indicate that the majority of effects may be additive (Zhu et al., 2015).

Table 3. Inheritance models and the coding of the possible genotypes for *A* and *a* alleles to study their effect

Inheritance Model	aa	Aa	AA
Additive	0	1	2
Dominant	0	1	1
Recessive	0	0	1
Heterodominant	0	1	0
Genotypic (Additive + Heterodominant)	0	1	2
	0	1	0

However, the importance of non-additive models has been shown for Mendelian diseases, where risk variants can have a dominant or recessive model of inheritance. For complex diseases, few examples have been reported, such as recessive effects in *FTO* for obesity (Wood et al., 2016), in *ITGA1* (Grarup et al., 2018), *TBC1D4* (Moltke et al., 2014) and *CDKAL1* (Steinthorsdottir et al., 2007; Wood et al., 2016) for type 2 diabetes, and widespread non-additive effects in *HLA* for autoimmune diseases (Lenz et al., 2015), including an heterodominant effect for ulcerative colitis (Goyette et al., 2015). Hence, it is likely that exclusive non-additive associations will be missed by the conventional additive approach, since it has been demonstrated that the additive model has limited statistical power to detect associations with complex traits showing a recessive effect (Lettre et al., 2007; Salanti et al., 2009) especially at lower frequencies since fewer homozygous are observed.

Rather than analyzing only one model (i.e., the additive), as it is a common practice in GWAS, the analysis of multiple models with an appropriate correction for multiple testing may lead to new findings (Tam et al., 2019).

3.4.3 The underestimated X chromosome

Beyond only performing additive association tests, the vast majority of GWAS exclude the sex chromosomes (Khramtsova et al., 2019; Wise et al., 2013). As they have been systematically ignored from GWAS, it is likely that variants on the sex chromosomes contribute to the missing heritability (Tukiainen et al., 2014), especially on the X chromosome (Figure 8).

There are no GWAS reporting signals in the Y chromosome (GWAS Catalog, 2019), although some studies found that the Y chromosome haplogroups contribute to diseases phenotypes (Charchar et al., 2012; Krementsov et al., 2017; Lu et al., 2016; Sezgin et al., 2009). However, Y chromosome is confined to males and contains the smallest number of genes (~80 coding-protein genes; chromosome 22, similar in size, contains ~400 coding-protein genes), most of which locate in the "male-specific region" (MSR) that constitutes 95% of the Y chromosome (Skaletsky et al., 2003). Moreover, and in contrast with the rest of the genome, there is no recombination with a partner chromosome in the MSR during meiosis. Thus, MSR is inherited unaltered. Hence, due to its haploid nature and the difficult to build linkage maps in it, the Y chromosome is routinely ignored in GWAS (Maan et al., 2017).

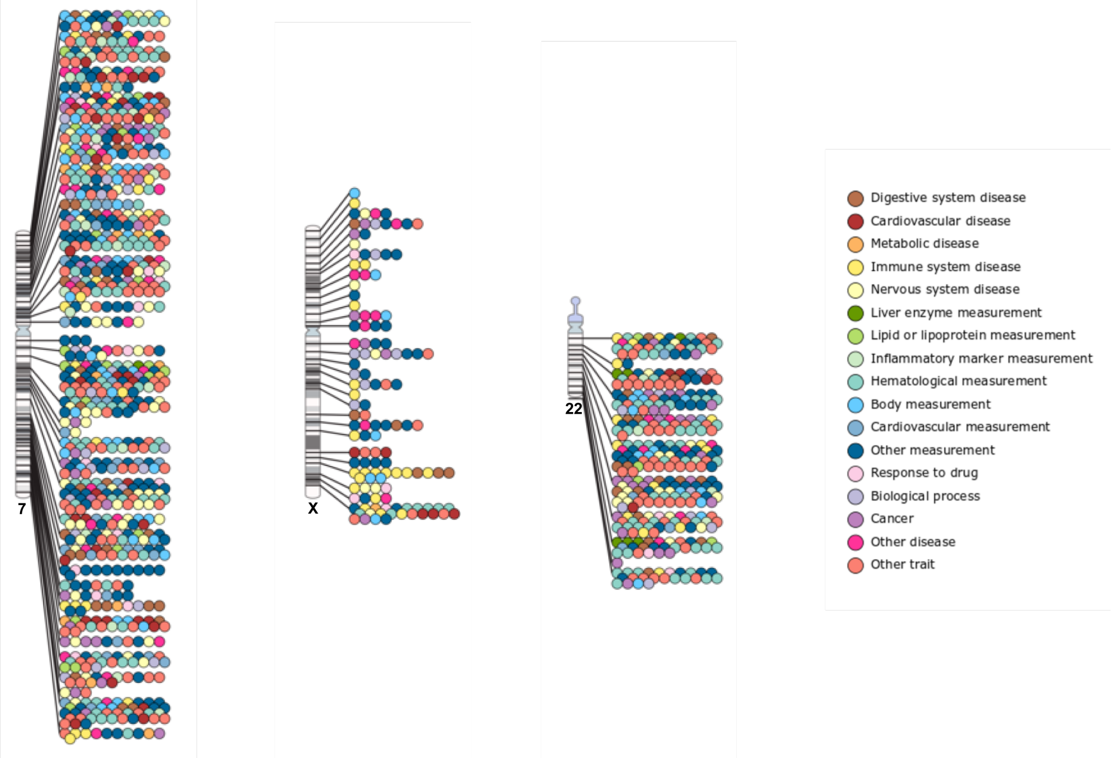


Figure 8. GWAS catalog diagram associations (p-value < 5×10^{-8}) on May 2018 for chromosome 7, chromosome X and chromosome 22.

Although chromosome 7 and chromosome X are comparable in size, few associations have been reported for chromosome X. Even the chromosome 22, one of the smallest chromosomes in the human genome, has a larger number of reported associations than chromosome X.

More remarkable is the systemic exclusion of the X chromosome in GWAS. With a size comparable to chromosome 7, the X chromosome contains more than 1,500 genes representing 5% of the genes in the human genome (Tukiainen et al., 2014; Wise et al., 2013). Since there is a correlation between the chromosome size and the number of reported associated loci, it is expected that analyzing the X chromosome, or reanalysis existing data that did not originally account for it, would lead to the discovery of new associations (Khramtsova et al., 2019). As an example of this, in a recent study that re-analyzed public available GWAS data for type 2 diabetes, including the X chromosome, a rare new variant that doubles the risk in men was identified (Bonas-Guarch et al., 2018).

Moreover, several Mendelian diseases are known to be X-linked (Wise et al., 2013). This provides evidence of the importance of the X chromosome in human diseases. Besides, for complex diseases, there is a broad appreciation that some disorders, including autism (Robinson et al., 2013) and autoimmune diseases (Kantarci et al., 2006; Ngo et al., 2014), are more diagnosed in one sex than in the other. In addition, even though several well-powered studies found largely similar heritability estimates for males and females for most traits (Ge et al., 2017; Polderman et al., 2015; Stringer et al., 2017; Traglia et al., 2017), there are notable exceptions such as post-traumatic stress disorder, hypertension, rheumatoid arthritis and allergic rhinitis (Duncan et al., 2018; Ge et al., 2017).

Taking all this into account, the X chromosome may influence disease risk directly or indirectly, and it is an open area of research with many opportunities. Therefore, this raises the question: why are researchers so discouraged to analyze the X chromosome?

First, biological sex is defined by the sex chromosomes that primarily determine the sexual differentiation of gonads (ovaries and testes) and the expression of sex hormones. Hence, women are typically XX and men XY. Therefore, men have half the dosage of women for the X chromosome. However, X chromosome complexity goes beyond this observation. The X chromosome is divided into the pseudoautosomal regions (PAR) and the non-pseudoautosomal region (non-PAR). As these names indicate, the PAR regions (PAR1 and PAR2) are homologous sequences between the X and the Y chromosomes that remind of the sequences in autosomes (men and women have two copies). Nevertheless, these regions are substantially smaller than the non-PAR, where there is indeed a difference in dosage between men and women since there is no homologous region in the Y chromosome. Hence, to ensure dosage compensation between both sexes, females have one copy silenced.

This process, called the X chromosome inactivation, results in approximately half of the cells expressing one copy and half expressing the other. Besides, this characteristic makes men more susceptible to some conditions than women, since men do not have the possibility of an alternative copy that may compensate a deleterious one. This is the case of hemophilia, Duchenne muscular dystrophy, Rett syndrome, fragile X syndrome, red-green color blindness and male-pattern baldness (Khramtsova et al., 2019). Moreover, to make matters more complicated, the X chromosome inactivation is not complete and around 23% of the X chromosome genes may escape from it, resulting in sex-biased gene expression of escape genes (Carrel and Willard, 2005; Tukiainen et al., 2017).

The common approaches used in GWAS, starting from genotyping platforms until association testing tools, have been mainly designed focused in the autosomal characteristics (Khramtsova et al., 2019; Wise et al., 2013), and as it has already been pointed out, specific considerations are needed to deal with the sex chromosomes (Konig et al., 2014). However, X-chromosome specific tools have been developed since 2007 (Gao et al., 2015; Marchini et al., 2007), but the availability of these tools has not increased the overall studies that include the X chromosome analysis (Wise et al., 2013).

The inner characteristics of the X chromosome make it difficult to analyze. Besides, there is a lack of power to detect GWAS significant associations in the X chromosome. The different dosage between males and females and male-female ratios in imbalanced cohorts impairs the statistical power to detect associations. Moreover, some imputation tools, such as MINIMAC (Das et al., 2016), cannot impute the X chromosome directly and need to be run separately for males and females. This hinders the incorporation of the X chromosome analysis together with the autosomes and impairs the statistical power by dividing the sample size in two in males-females balanced cohorts.

In summary, the X chromosome is routinely omitted from GWAS, mainly due to their unique characteristics that make its analysis challenging and requires additional expertise. Besides, important findings, and thus high-profile publications, can be achieved when analyzing autosomal chromosomes only (Wise et al., 2013).

Altogether, this partially explained why researches are discouraged about the incorporation of the X chromosome into the analysis. Nevertheless, in light of the evidence that highlights its importance in human diseases, the X chromosome deserved more attention and should be analyzed despite being more difficult.

4 Beyond GWAS results: Post-GWAS analysis to study the genetics behind complex diseases

4.1 Interpreting significant variants from GWAS

As explained in previous sections, GWAS results are commonly represented with Manhattan plots followed by regional association plots for the loci of interest. In general, all the variants with a p-value $< 5 \times 10^{-8}$ are considered associated at a genome-wide significant level, even though some researches use a weaker threshold of p-value $< 10^{-6}$ to highlight suggestive regions for further analysis (Schaid et al., 2018).

As stated before, since GWAS evaluate variants individually, the captured associated signal can be indirectly associated with the trait due to the complex LD patterns among variants. That is to say, associated variants from GWAS results can be merely correlated with the causal one, and it is challenging to determine which one is actually the casual variant for each associated region. One might think that the variant with the smallest p-value in a region, usually called lead or top variant, is the causal one. Nonetheless, the causal variant is not likely to be the one with the smallest p-value partially due to the small effects sizes of variants on complex traits (Schaid et al., 2018; Schaub et al., 2012), and to determine which variant in each region is the most likely to be functional, and therefore, causal, it is not straightforward.

In order to undergo costly and time-consuming lab functional studies to translate GWAS results into clinics, it is crucial to properly prioritize the variants that deserved to be further evaluated among the number of variants associated after a GWAS. This is when fine-mapping, functional annotations, gene expression associations, and pathway and gene enrichment analysis, can help.

4.1.1 Fine-mapping GWAS regions

Given statistical evidence of the association of a genomic region with a complex trait, and assuming that there is at least one causal variant there, fine-mapping is a statistical approach that seeks to elucidate the genetic variant (or variants) responsible for that trait (Schaid et al., 2018).

To improve the fine-mapping resolution, meta-analyzing multiple cohorts or increasing variants density through genotype imputation is crucial (Li et al., 2009; Marchini and Howie, 2010; Schaid et al., 2018). However, imputation quality depends on the LD structure, and low-frequency and rare variants, which are not in strong LD with neighboring variants, might require additional genotyping to evaluate their association and reduce measurement errors properly. Custom arrays or those targeting specific diseases, such as the OncoArray (Amos et al., 2017), the Metabochip (Voight et al., 2012) or the Immunochip (Parkes et al., 2013), may help to increase the sample size and the density of variants in a particular region in a cost-effective manner.

One of the most straightforward approaches to disentangle independent regions among the associated GWAS loci is conditioning on the lead variant, also called "forward stepwise conditional regression". Hence, independent regions inside the associated locus can be found by treating the top variant as an adjusting covariate in a regression model and testing the remaining variants in that region. However, multiple conditional tests might need to be done for multiple variants until no test is significant to point out all the independent regions, thus increasing the chance of a false-positive result or requiring the use of stringent thresholds. This situation can occur, for example, when the correlation between the variants is very high, diminishing the probability of finding secondary signals (Schaid et al., 2018).

To determine the causal variants for each independent region, credible sets determined through Bayesian methods are among the most used approaches. A credible set is defined as the minimum set of variants that contains all causal variants with probability α (e.g., $\alpha = 99\%$) (Hormozdiari et al., 2014; Schaid et al., 2018).

4.1.2 Functional annotations

Selected variants by fine-mapping can be classified as protein-coding or non-protein-coding depending on whether they are in a protein-coding sequence or not. When a variant is inside a gene that encodes a protein, genomic annotations are focused on the impact of this variant in the resulting protein. These variants are the easiest to interpret, and follow-up functional laboratory-based analyses are quite straightforward. However, most GWAS associations, are found in non-protein coding sequences often involved in gene regulation (Encode Project Consortium et al., 2007; Maurano et al., 2012; Schaid et al., 2018; Schaub et al., 2012). Some examples of non-coding annotations are promoters, enhancers, long non-coding RNAs, transcription factor binding sites, histone modifications, and DNAse I hypersensitivity sites. Further increasing complexity, gene regulation is highly tissue/cell-specific, varies through developmental stages, and environmental factors influence it. Moreover, variants in a particular position might modify the expression of a distant gene, even when it is not the nearest one (Figure 9).

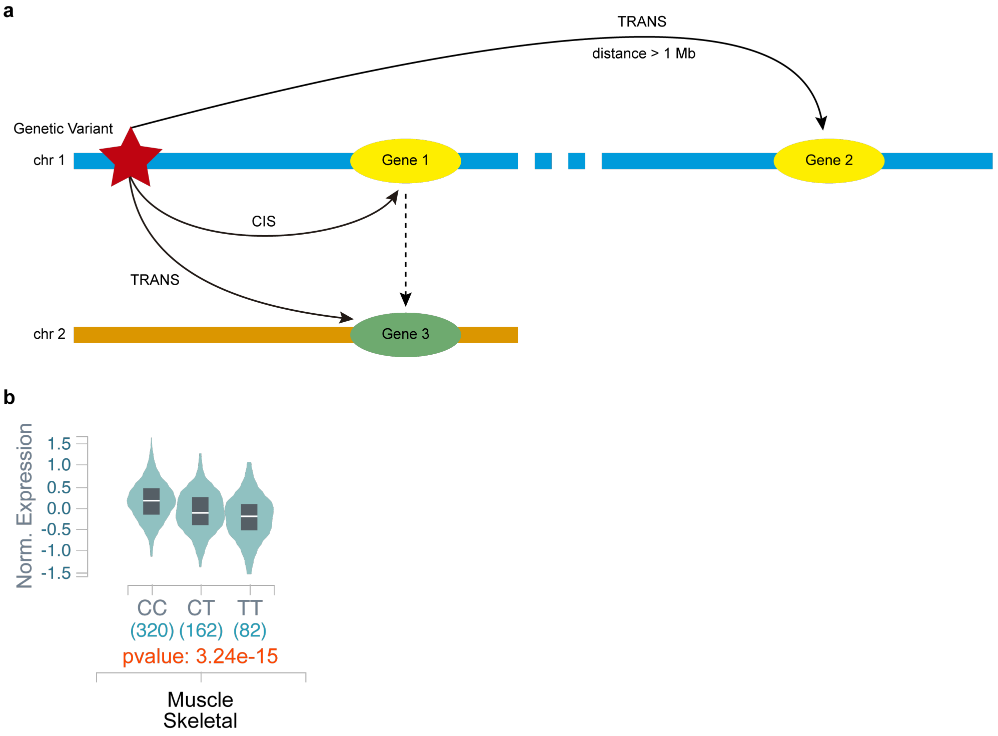
Our limited knowledge about the regulatory networks impairs our functional interpretation of GWAS results. However, many efforts to facilitate a direct and systematic interpretation have emerged. Hence, large public initiatives have increased the available databases for genomic annotation, including Gene Ontology (Rhee et al., 2008), GENCODE (Harrow et al., 2012), ENCODE (Encode Project Consortium, 2004), FANTOM5 (Andersson et al., 2014), The Roadmap Epigenomics Project (Roadmap Epigenomics et al., 2015) and GeneHancer (Fishilevich et al., 2017). Integrating multiple tissues and cell types, it is estimated that current functional annotation covers around 80% of the human genome (Pennisi, 2012; Schaid et al., 2018).

Although such data might not be perfect, functional annotation can help to prioritize variants for follow-up analysis after assigning a biological function to them.

4.1.3 Expression quantitative trait loci

Variants associated with complex diseases or traits are likely to be expression quantitative trait loci (eQTLs) (Nicolae et al., 2010). That is, they influence the amount of expression of genes, ultimately influencing the trait. Hence, identifying associations between GWAS significant variants and gene expression may help to point out the most plausible gene behind the condition. Variants can be associated with gene expression in *cis* or *trans*, where cis-eQTLs affect nearby genes, while trans-eQTLs affect distant genes, even in other chromosomes (Figure 9).

Figure 9. Cis- and trans- eQTLs and the expression distribution of genotypes.



a eQTLs are classified according to the distance of the associated gene. cis-eQTLs affect gene expression of local genes near the genetic variants ("Gene 1" in the figure) while trans-eQTLs alter the expression of distant genes ("Gene 2" in the figure) in the same chromosome (in blue) or even genes ("Gene 3" in the figure) located in different chromosomes (in orange). **b** Example of a violin plot from GTEx v8 data, showing the expression distributions of the three genotypes for *AP4B1* (ENSG00000134262.12) and rs10858023 (chr1_113906130_C_T_b38) in muscle skeletal tissue. chr = chromosome, Norm.Expression = normalized expression.

As gene expression is highly tissue-specific, a critical step when integrating GWAS and eQTLs results is the tissue where the expression was measured. To that end, the GTEx project facilitates the exploration of gene expression in multiple tissues, including data for 449 human donors across 44 tissues (GTEx Consortium, 2013). In addition, a recently released catalog of genetic effects on gene expression after analyzing 31,684 blood samples from the eQTLGen Consortium has been generated (Võsa et al., 2018).

However, in the GTEx project, 92.74% of common variants were found associated with the expression level of at least one nearby gene with a p-value < 0.05, and after controlling for the multiple tissues tested, 48.45% remain associated (GTEx Consortium, 2013). Therefore, coincidental overlaps are very likely. As some authors have pointed out, the abundance of eQTLs data and strong LD structures have increased the false positive rate of causal hypotheses for GWAS results (Liu et al., 2019; Nica et al., 2010). Hence, an overlap between an eQTL and a GWAS signal can lead to an incorrect causal hypothesis. Moreover, when a locus contains multiple eQTLs for different genes, a GWAS signal may not be caused by the most significant. To address this issue, colocalization analysis has been developed (Giambartolomei et al., 2014; Hormozdiari et al., 2016; Nica et al., 2010; Zhu et al., 2016), including tools for visualizing colocalization events (Liu et al., 2019). Briefly, colocalization compares the distribution of summary statistics from two association signals accounting for the LD structure of the region, thus mitigating the false-positive findings by analyzing multiple variants at a time.

4.1.4 Pathway and functional enrichment analyses

An additional methodology to link GWAS results to their likely biological function is to identify causal genes and pathways involved in the pathophysiology of complex diseases by functional enrichment analysis. Traditional approaches explored protein interaction maps, gene expression data and constructed gene networks with predefine genes and key pathways for the diseases (Raychaudhuri et al., 2009). Hence, the traditional approaches limit the discovery of new genes and pathways.

To fill this gap, computational approaches that allow the exploration of genes and pathways without preconceived hypotheses have been developed. Among the most used tools there is GRAIL, a gene prioritization framework based on text-mining from PubMeD abstracts (Raychaudhuri et al., 2009), MAGENTA (Segre et al., 2010), a gene-set enrichment framework that explores pathways from public databases, and recently DEPICT (Pers et al., 2015), based on predicted gene functions. DEPICT enables the exploration of poorly annotated genes and outperforms both GRAIL and MAGENTA in the prioritization of genes and the analysis of gene set enrichment, thus generating more accurate testable genes and pathways hypotheses from GWAS results.

In summary, the associated variants in a GWAS require further analyses to elucidate the ones that deserve cost and time consuming laboratory-based follow-up analysis. Due to the amount of information and the complicated LD structure of the human genome, the identification of the causal variant and genes, and its biological function, is not easy and unequivocal. Luckily, fine-mapping, large consortia of eQTLs, gene expression in multiple tissues in combination with colocalization analyses, and gene set enrichment analysis with the identification of potentially relevant pathways, are disentangling the complex world of GWAS associated variants, moving from the association signal to the biological function influencing the disease.

Objectives

Objectives

This thesis has the following objectives:

I. To increase the amount of genetic variation to be tested for association after genotype imputation combining the results from multiple sequence-based reference panels.

II. To analyze the X chromosome alongside with the autosomes.

III. To include additive and non-additive inheritance models for the association test to study their contribution.

IV. Develop GUIDANCE, an integrated pipeline with our GWAS strategy to facilitate, and thus promote, a comprehensive GWAS of the existing and the newly generated GWAS datasets.

V. To apply our approach to a large publicly available cohort; The Resource for Genetic Epidemiology Research on Aging (GERA) cohort.

Methods

This chapter has been split into two major blocks; the first one focused on 1) the programming framework on top which GUIDANCE was developed and its technical advantages, and the second one related to 2) the analysis of the GERA cohort and the posterior follow-up analyses done to date.

1 Genome-wide imputation and association testing for parallel computing (GUIDANCE)

1.1 GUIDANCE developed on top of COMPSs

Current GWAS workflows are sequential and require constant intervention between each step. However, some steps in GWAS, such as genotype imputation, can run in parallel after splitting each chromosome into chunks. This also represents a significant effort from the researcher since it requires manually managing the different computations in parallel using threads or another parallel environment. GUIDANCE, however, can execute in parallel without requiring a broad background in parallel environments. For that, GUIDANCE was implemented on top of the COMP Superscalar Programming Framework (COMPSs) (Lordan et al., 2014), which makes it feasible for non-HPC experts.

Therefore, we developed GUIDANCE by combining and integrating state-of-the-art GWAS analysis tools into COMPSs, releasing users from the responsibility of dealing with the computational complexity of the whole process.

COMPSs is a programming framework that aims to make more accessible the development of parallel workflows in computing platforms such as clusters or clouds, keeping the workflow code agnostic of the actual computing platform. Therefore, while COMPSs programs are expressed as sequential code in Java, Python or C/C++, the runtime can parallelize the code and make decisions such as scheduling the different workflow nodes and transferring the data to the nodes where the computing will take place.

With COMPSs, GUIDANCE workflow was implemented as a sequential Java program. The code contains the calls to the GWAS tools encapsulated in Java methods and selected as tasks (see Results chapter, section 1, for detailed information of GUIDANCE workflow). As a result of executing the workflow, a task dependency graph is dynamically generated by COMPSs (Figure 10), which controls the execution of those tasks on the underlying parallel infrastructure.

1.1.1 COMPSs runtime system

The COMPSs runtime system is in charge of controlling the parallelization of the application and managing its execution on a set of distributed resources. Hence, the runtime system interacts with the underlying infrastructure on behalf of GUIDANCE.

The main functionalities provided by the runtime system from which GUIDANCE benefits are the following:

- Data dependency analysis and task scheduling: the runtime system detects and enforces task dependencies. For example, as shown in Figure 10, genotype imputation cannot start before the haplotype phasing step has finished. Tasks are scheduled and submitted to the available resources, trying to exploit data locality when possible.

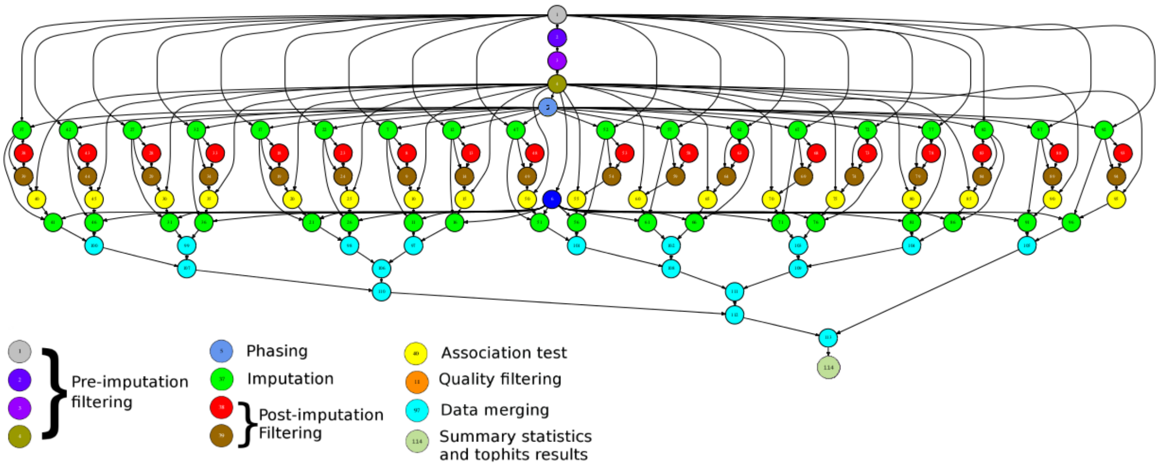


Figure 10. A representative example of task distribution and dependencies from a GUIDANCE execution corresponding to chromosome 22.

Each node represents a particular task, and each link represents a dependency between tasks. The type of task is defined by the color-code displayed below.

- On-demand resource allocation: the runtime system can work in cloud environments, where it exploits the elastic capabilities of the infrastructure. In particular, new resources are requested depending on the task load that the application generates at every moment. Hence, the number of resources is dynamically adapted to face the peaks and valleys of such load.

- Fault-tolerance: the runtime system implements fault-tolerance mechanisms for task submission and file transfer that ensure that the execution will continue even in the case of partial failure, thus guaranteeing the proper completion of the workflow execution.

- Monitoring and results collection: users can follow the progress of the GUIDANCE workflow execution, obtaining the resources usage information and a real-time execution graph. Additionally, results are collected as tasks generate them, so the user can verify the correctness of such results without waiting until the whole application ends. For example, when the association test for a chunk finishes, the outputs are written, and a file summarizing the task is generated.

- Performance analysis: users can generate execution traces, which graphically represent the execution behavior of the workflow. In such traces, the user can see which tasks ran and where, as well as data transferred between resources. Traces are useful to analyze the performance of the application and find possible improvement opportunities or performance issues.

2 The analysis of GERA cohort

2.1 GERA cohort description

The access to the Resource for Genetic Epidemiology Research on Aging (GERA) cohort data was obtained through dbGaP (phs000674.v1.p1).

GERA cohort was created by an RC2 "Grand Opportunity" awarded by the Kaiser Permanente Research Program on Genes, Environment, and Health (RPGEH) and the UCSF Institute for Human Genetics (AG036607; Schaefer/Risch, PIs). The RC2 project enables the genotyping of over 110,266 saliva samples from adults who are members of the Kaiser Permanente Medical Care Plan, Northern California Region (KPNC), and participating in its RPGEH.

The resulting cohort have an average age of 63 years, ranging from 18 to over 100 years old at the time of the survey (2007), with 42% of males and 58% of females.

The cohort is described as generally well-educated individuals with above-average income and ethnically diverse. However, only 19% (20,925) individuals are from non-European ancestry, while 81% (89,341) are described as white non-Hispanic participants.

In order to accurately capture the genetic variability in this multi-ancestry cohort, four custom genotyping arrays were designed using the Affymetrix Axiom array, one for each of the four majors ancestries in the cohort, i.e., African Americans, East Asians, Latinos, and Non-Hispanic Whites. Description of the array designs has been provided in two publications (Hoffmann et al., 2011a; Hoffmann et al., 2011b).

Health conditions in the participants are derived from summarizing ICD-9 coded diagnoses in Kaiser Permanente Electronic Medical Records. The criteria of inclusion as a case for a particular disease is the requirement of at least two diagnoses in a disease category recorded on separate days to reduce false positives.

After an explicit requirement of consent by email, data from 78,486 participants (Table 4) was deposited in dbGaP, with similar demographic characteristics to those of the initial genotyped cohort.

2.2 Quality control

A subset of 62,281 subjects from European ancestry from the GERA cohort was QCed before GWAS. For the QC, a QC pipeline for genotyped data that was previously applied in a Type 2 Diabetes study (Bonas-Guarch et al., 2018) was used. Briefly, the QC pipeline is a 3-step quality control protocol using PLINK, which includes two stages of SNP removal and an intermediate step in between for sample exclusion. The QC pipeline is described below in detail.

2.2.1 Variant based filtering

First, using PLINK, variants were filtered according to their proportion of missingness (--missing), the deviance of HWE (--hardy) and MAF (--freq). In case-control studies, the deviance of HWE is generally tested for the whole cohort as well as controls (e.g., a HWE cohort threshold of 1×10^{-20} and 1×10^{-6} for controls). However, since the GERA cohort includes multiple diseases, neither a filter for controls deviance of HWE nor a missingness test to analyze variant differences between cases and controls were applied.

Table 4. GERA diseases and sample size before QC.

Disease	Cases	Controls
Asthma	13,110	65,376
Allergic Rhinitis	19,939	58,547
Cardiovascular	19,963	58,523
Cancer	21,536	56,950
Major Depressive Disorder	9,732	68,754
Dermatophytosis	10,768	67,718
Type 2 Diabetes	10,572	67,914
Dyslipidaemia	41,587	36,899
Hypertensive	39,291	39,195
Hemorrhoids	12,574	65,912
Hernia Abdominopelvic Cavity	8,267	70,219
Insomnia	5,276	73,210
Iron Deficiency	3,429	75,057
Irritable Bowel	4,089	74,397
Macular Degeneration	4,624	73,862
Osteoarthritis	26,823	51,663
Osteoporosis	7,050	71,436
Peripheral Vascular	5,641	72,845
Peptic Ulcers	1,309	77,177
Psychatric	11,835	66,651
Stress	6,147	72,339
Varicose Veins	3,196	75,290

Hence, the filters and thresholds applied in this step were the following:

-MAF < 0.001

-Miss ≥ 0.05

-HWE cohort $\leq$ 1e-10

-HWE controls ≤ 0

-HWE cases ≤ 0

2.2.2 Sample based filtering

For sample exclusion, we considered the following criteria: gender discordance (--check sex) and variant call rates ≥ 2% (--mind 0.02) using PLINK.

Related subjects were also excluded following a published protocol (Anderson et al., 2010). After generating an identity by state (IBS) pair-wise comparison matrix for all the samples to determine the degree of shared ancestry from each pair of individuals, the identity by descent (IBD) was estimated. After that, we removed the individual with the highest proportion of missingness for third-degree relatives (pairs with PI-HAT > 0.125).

For IBS calculation, data was pruned to remove variants in LD, and these independent variants were merged with HapMap to study the population structure in the cohort under study. Hence, individuals showing more than four standard deviations within the distribution of the study population were also removed according to the first seven principal components (PCs). To generate the 7 PCs, the pipeline uses a multidimensional-scale analysis with PLINK (--read-genome --cluster --mds-plot 7). These PCs were then added to the regression model as covariates in the association test. In addition, a plot based on the first 4 PCs was generated (Figure 11), to manually inspect the data before GWAS.

Finally, an additional variant filtering was performed as described in the previous section.

This automatic QC pipeline ended with a summary report displaying all the thresholds applied and the number of variants and samples excluded in each step, pointing out the reason for their exclusion (Figure 12).

After applying this QC to the 62,281 subjects from European ancestry from the GERA cohort, 56,637 subjects remained for the GWAS (Figure 12 and Table 5).

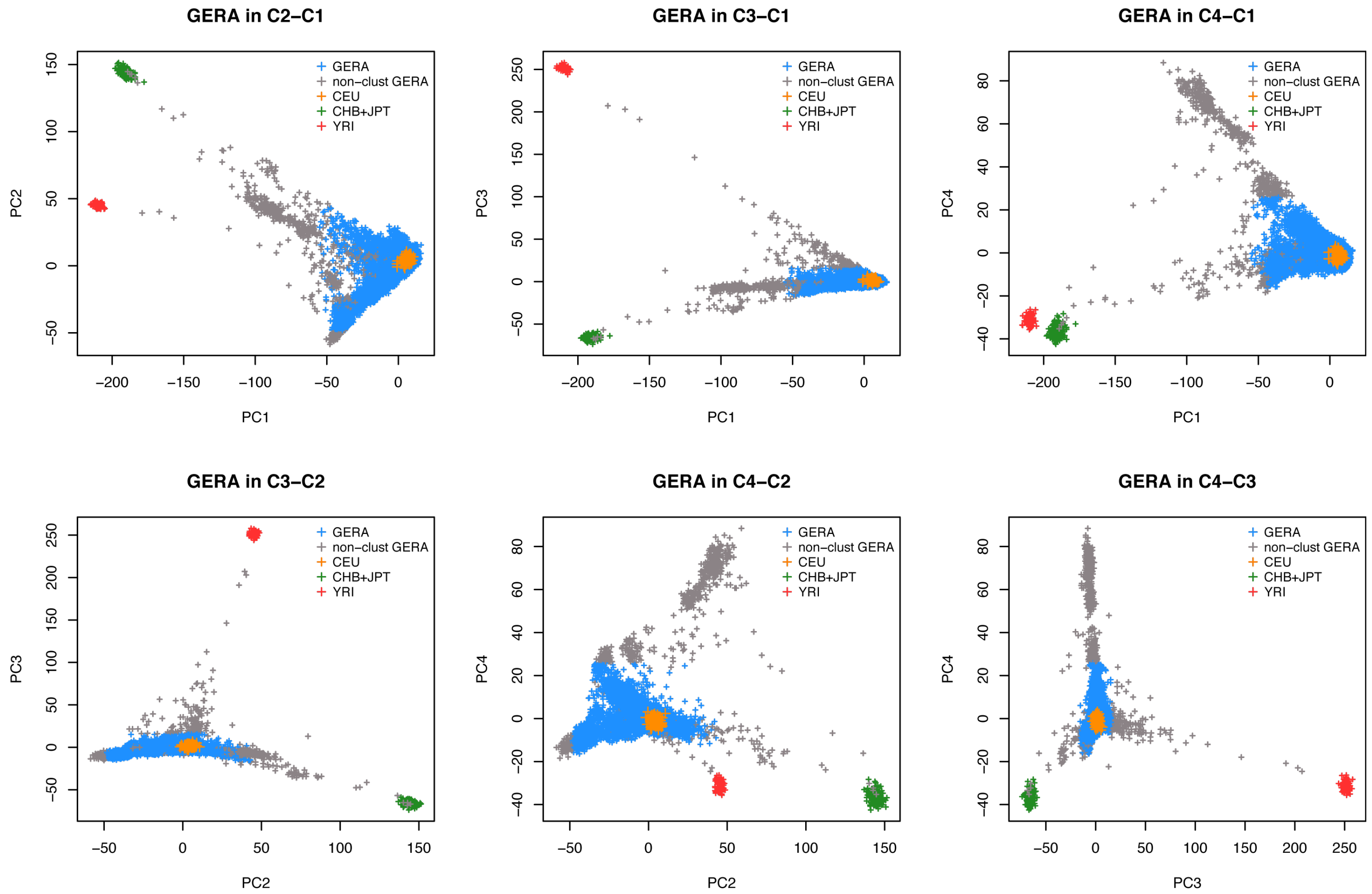
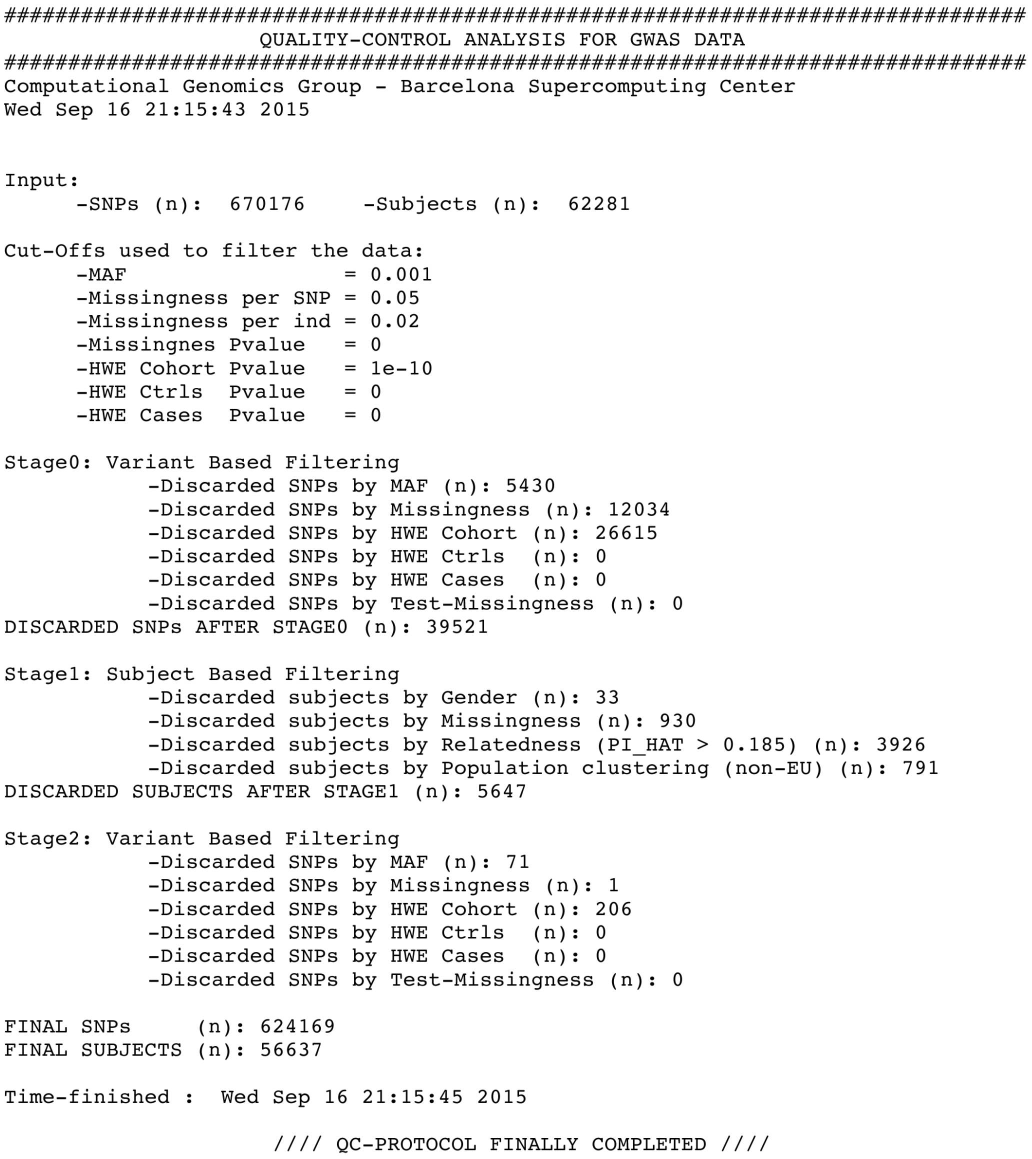


Figure 11. The first 4 PCs from the QC of GERA cohort.

As seen in the plots, the individuals that do not cluster with the European samples from HapMap are depicted in grey and removed from the dataset. Samples in blue are considered from the same ancestry and remain for further analysis.



```
##############################################################################
                    QUALITY-CONTROL ANALYSIS FOR GWAS DATA
##############################################################################
Computational Genomics Group - Barcelona Supercomputing Center
Wed Sep 16 21:15:43 2015


Input:
      -SNPs (n):  670176    -Subjects (n):  62281

Cut-Offs used to filter the data:
      -MAF                    = 0.001
      -Missingness per SNP = 0.05
      -Missingness per ind = 0.02
      -Missingnes Pvalue    = 0
      -HWE Cohort Pvalue    = 1e-10
      -HWE Ctrls  Pvalue    = 0
      -HWE Cases  Pvalue    = 0

Stage0: Variant Based Filtering
           -Discarded SNPs by MAF (n): 5430
           -Discarded SNPs by Missingness (n): 12034
           -Discarded SNPs by HWE Cohort (n): 26615
           -Discarded SNPs by HWE Ctrls  (n): 0
           -Discarded SNPs by HWE Cases  (n): 0
           -Discarded SNPs by Test-Missingness (n): 0
DISCARDED SNPs AFTER STAGE0 (n): 39521

Stage1: Subject Based Filtering
           -Discarded subjects by Gender (n): 33
           -Discarded subjects by Missingness (n): 930
           -Discarded subjects by Relatedness (PI_HAT > 0.185) (n): 3926
           -Discarded subjects by Population clustering (non-EU) (n): 791
DISCARDED SUBJECTS AFTER STAGE1 (n): 5647

Stage2: Variant Based Filtering
           -Discarded SNPs by MAF (n): 71
           -Discarded SNPs by Missingness (n): 1
           -Discarded SNPs by HWE Cohort (n): 206
           -Discarded SNPs by HWE Ctrls  (n): 0
           -Discarded SNPs by HWE Cases  (n): 0
           -Discarded SNPs by Test-Missingness (n): 0

FINAL SNPs     (n): 624169
FINAL SUBJECTS (n): 56637

Time-finished :  Wed Sep 16 21:15:45 2015

                     //// QC-PROTOCOL FINALLY COMPLETED ////
```

Figure 12. Summary report of the GERA QC.

Summary of the thresholds used, and the variants and samples removed in each step.

Table 5. GERA diseases and sample size for European populations after QC.

Disease	Cases	Controls
Asthma	9,209	47,428
Allergic Rhinitis	13,936	42,701
Cardiovascular	15,009	41,628
Cancer	17,131	39,506
Major Depressive Disorder	7,264	49,373
Dermatophytosis	7,676	48,961
Type 2 Diabetes	6,967	49,670
Dyslipidaemia	30,244	26,393
Hypertensive	28,391	28,246

Hemorrhoids	9,129	47,508
Hernia Abdominopelvic Cavity	6,291	50,346
Insomnia	3,972	52,665
Iron Deficiency	2,439	54,198
Irritable Bowel	3,117	53,520
Macular Degeneration	3,685	52,952
Osteoarthritis	20,212	36,425
Osteoporosis	5,399	51,238
Peripheral Vascular	4,301	52,336
Peptic Ulcers	920	55,717
Psychatric	8,624	48,013
Stress	4,314	52,323
Varicose Veins	2,483	54,154

2.3 Running GUIDANCE to analyze GERA

Using GUIDANCE, genotypes were pre-phased into whole haplotypes with SHAPEIT2 and, after that, genotypes were imputed independently for each reference panel, i.e., 1000G phase 3, UK10K, GoNL and HRC, using IMPUTE2.

After excluding variants with an info score < 0.7 and MAF < 0.001, the imputed genotypes from each panel were tested for assotiacion separately using SNPTEST. For autosomes, additive, dominant, recessive, heterodominant and genotypic inheritance models were assessed. Seven principal components, sex and age were added as covariates in the logistic regression.

For chromosome X, the analysis was restricted to the non-pseudoautosomal (non-PAR) region and males and females were analyzed separately as well as in conjunction, but stratifying the association analysis by sex in order to account for hemizygosity for males, while allowing an autosomal model for females. We assumed the random X chromosome inactivation model by using the method “newml” from SNPTEST.

To maximize power and accuracy, the association results from the four reference panels were combined by choosing for each variant, the reference panel that provided the best IMPUTE2 info score. This final set of variants was then filtered for HWE in controls $p \leq 1 \times 10^{-6}$.

Further details about the analysis of GERA can be found in the configuration file from this GUIDANCE execution (Figure 15, Results chapter). As a large part of this thesis was focused on the development of GUIDANCE, detailed information on the pipeline used to analyze GERA can be found in the Results chapter. Likewise, details about the definition of each phenotype in the GERA cohort can be found in the following link: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/GetPdf.cgi?id=phd004308.1.

2.3.1 Identification of known and new associated loci

After the association testing, GUIDANCE provided a list of variants that passed the p-value threshold specified in the configuration file (i.e., 5×10^{-8}). Using "IRanges" R package, all the genome-wide significant variants are collapsed into ranges that define each associated locus.

To distinguish between known or new associated regions, a general and systematic approach was applied since an in-deep analysis for all the diseases included in this study would have required an exhaustive screening outside the scope of this thesis. Therefore, for each top variant we looked for any proxy variant with an LD $r^2 > 0.35$ in the GWAS catalog (accession 5 September 2019) associated with the same phenotype or a related one (for example, bone mineral density, cholesterol levels or diastolic/systolic blood pressure phenotypes for osteoporosis, dyslipidemia or hypertension, respectively). HLA regions at chromosome 6 were excluded since the particularities of these regions required further detailed studies on their LD pattern to clarify if our findings constitute new regions or not. Proxies were selected using LDlink (https://ldlink.nci.nih.gov/) (Machiela and Chanock, 2015).

2.4 Follow-up analysis

2.4.1 Dominance deviation test

The additive model can detect non-additive variants, and non-additive models can capture additive signals. To detect genuine differences between additive and non-additive signals, we performed a dominance deviation test for all the 93 autosomal genome-wide significant loci.

Dominance deviation was tested by a logistic regression analysis using PLINK. Sex, age and the first 7 PCs were included as covariates.

2.4.2 Replication using UK Biobank

When collecting and analyzing phenotypes from the UK Biobank (application number: 31063 and 27892), the curation and harmonization of the vast array of categorizations, variable scaling, and follow-up responses is a central challenge. In order to generate meaningful, interpretable phenotypes, we use the PHEnome Scan ANalysis Tool or PHESANT (https://github.com/MRCIEU/PHESANT) (Figure 13).

We performed the association testing for the curated phenotypes using SNPTEST for additive, dominant, recessive, heterodominant and genotypic inheritance models.

For each of the novel loci, we searched for equivalent phenotypes or for traits related to the phenotype where the novel association was discovered.

With the association testing results of both the GERA cohort and UK Biobank, we meta-analyzed the results using METAL when equivalent phenotypes were found in UK Biobank. For the meta-analysis, we use the inverse variance-weighted fixed-effect model for all the variants except for the rs557998486 variant associated with age-related macular degeneration since its beta, calculated with the "em" method from SNPTEST, was inflated. Therefore, we performed a sample size based meta-analysis, which converts the direction of the effect and the p-value into a Z-score. In order to compute the most accurate estimate of the odds ratio, we performed a mega-analysis, by merging the genotypes of the GERA cohort with the UK Biobank and testing the association with the " expected" method from SNPTEST.

To explore the assciations with biomarkers in UK Biobank, only the results from the first visit were taken into account since less than 10% of the cases were present in the second visit, and again, we assessed additive and non-additive inheritance models using SNPTEST.

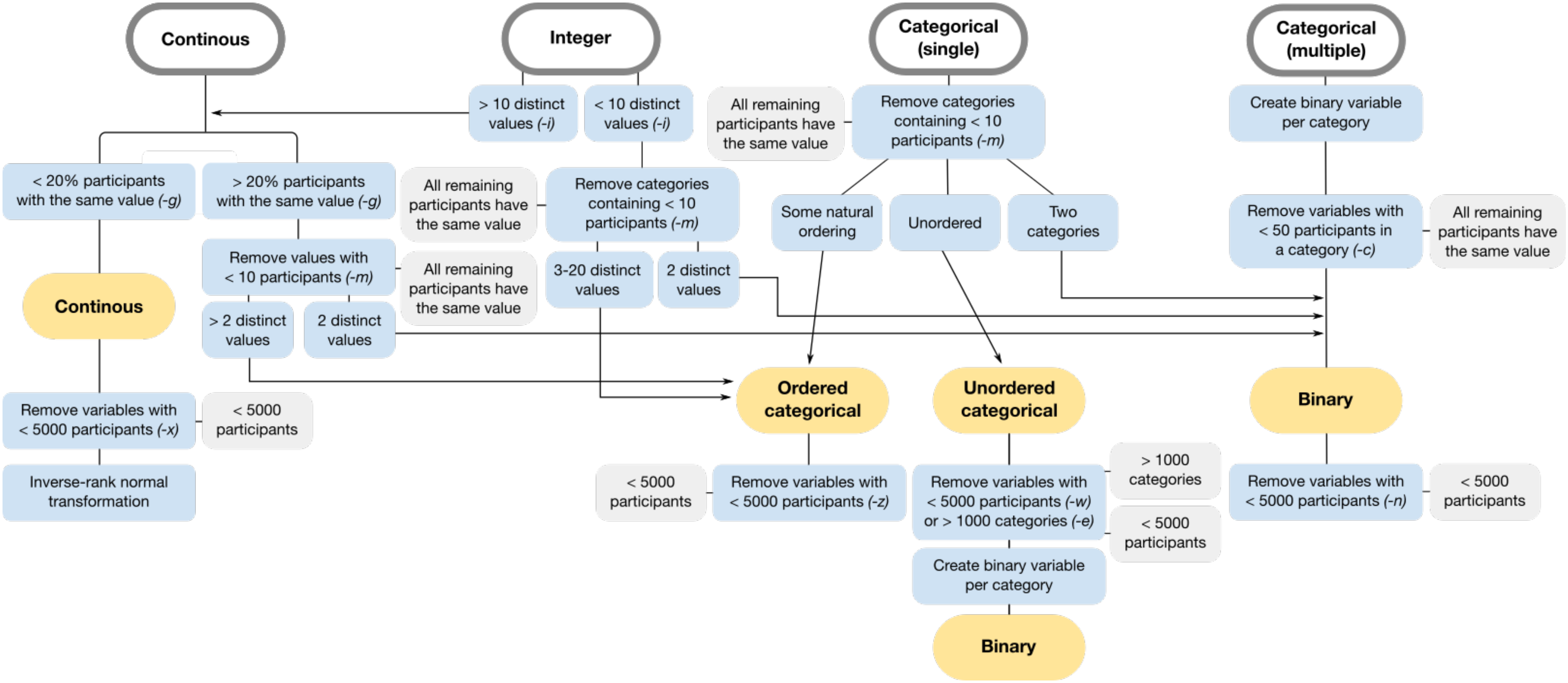

Continous
Integer
Categorical (single)
Categorical (multiple)
> 10 distinct values (-i)
< 10 distinct values (-i)
All remaining participants have the same value
Remove categories containing < 10 participants (-m)
Create binary variable per category
< 20% participants with the same value (-g)
> 20% participants with the same value (-g)
All remaining participants have the same value
Remove categories containing < 10 participants (-m)
Some natural ordering
Unordered
Two categories
Remove variables with < 50 participants in a category (-c)
All remaining participants have the same value
Remove values with < 10 participants (-m)
All remaining participants have the same value
3-20 distinct values
2 distinct values
Continous
> 2 distinct values
2 distinct values
Ordered categorical
Unordered categorical
Binary
Remove variables with < 5000 participants (-x)
< 5000 participants
Inverse-rank normal transformation
< 5000 participants
Remove variables with < 5000 participants (-z)
Remove variables with < 5000 participants (-w) or > 1000 categories (-e)
> 1000 categories
< 5000 participants
Create binary variable per category
Binary
Remove variables with < 5000 participants (-n)
< 5000 participants

Figure 13. Phenotype curation pipeline.

Raw phenotype data (gray outlined boxes) are passed to PHESANT, and a collection of filters (blue boxes) is applied. The thresholds shown here are the defaults in our modified version of PHESANT that can be altered in the phenomescan.r code using the flags displayed in parentheses. Gray filled boxes display the criteria for removal, and yellow filled boxes show the category of the variable after the rules in the blue boxes have been enforced.

2.4.3 Definition of 99% credible sets of GWAS significant loci

For each genome-wide significant region, the fraction of aggregated variants that have a 99% probability of containing the causal one was identified. The 99% credible set of variants for each region were defined with a Bayesian refinement approach (Wellcome Trust Case Control Consortium et al., 2012), considering variants with an $r^2 > 0.1$ with the leading one.

For each variant within a particular associated locus, the credible set provides a posterior probability of being the causal one (Wellcome Trust Case Control Consortium et al., 2012). The approximate Bayes factor (*ABF*) for each variant was estimated as

$$ABF = \sqrt{1-r}\, e^{(rz^2/2)} \ ,$$

where

$$r = \frac{0.04}{(SE^2 + 0.04)},$$

$$z = \frac{\beta}{SE}.$$

The β and the SE result from a logistic regression model testing for association. The posterior probability for each variant was calculated as

$$Posterior\ Probability_i = \frac{ABF_i}{T},$$

where ABF_i corresponds to the approximate Bayes' factor for the marker *i,* and *T* represents the sum of all the *ABF* values enclosed in the interval. As commonly employed by SNPTEST, this calculation assumes that the prior of the β is a Gaussian with mean 0 and variance 0.04.

Finally, the cumulative posterior probability was calculated after ranking the variants according to the *ABF* in decreasing order. Variants were included in the 99% credible set of each region until the cumulative posterior probability of association got over 0.99.

2.4.4 Functional annotation of novel findings

The eQTLGen Consortium (https://www.eqtlgen.org/cis-eqtls.html, last access on July 2019) and GTEx portal (https://gtexportal.org/, last access on July 2019) were used to find associations between our novel findings and gene expression. When the variant was not available in these resources, a proxy SNP from LDlink (https://ldlink.nci.nih.gov/?tab=ldproxy) was used instead.

Colocalization analysis was performed to determine whether the overlap between GERA associated loci and GTEx eQTLs was due to a true-shared association signal. Colocalization was assessed by a Bayesian test using summary statistics from the two studies (Giambartolomei et al., 2018); summary statistics from the GTEx study were downloaded from the GTEx portal (https://gtexportal.org/, last access on July 2019). The test was performed using the R package coloc v3.2-1 (https://cran.r-project.org/web/packages/coloc/). For each pair of GWAS locus-eQTL, the test provided a posterior probability for the two loci to share the same causal variants.

For the functional characterization of rs77704739 and rs557998486, we used the WashU EpiGenome Browser (http://epigenomegateway.wustl.edu/browser/, last access on July 2019). Public data from the reference human epigenomes from the Roadmap Epigenomics Consortium track hubs and the Roadmap Epigenomics Integrative Analysis Hub were used. These data were released by the NIH Roadmap Epigenomics Mapping Consortium.

Results

Following the hierarchy of topics in the previous section, the results have been split into two blocks; the first one related to 1) the development of GUIDANCE, including a detailed explanation of the workflow, and the second one focused on 2) the association findings as a result of the analysis of GERA cohort using GUIDANCE.

1 Genome-wide imputation and association testing for parallel computing

1.1 Developing GUIDANCE: An overview

As current GWAS workflows are computationally demanding and time-consuming, we developed GUIDANCE as an integrated framework to analyze genome-wide genotyped data in a single execution in parallel computing infrastructures without the need for extensive computational expertise or constant user intervention.

Integrating state-of-the-art tools with in-house code written in java, bash and R (Table 6), GUIDANCE efficiently performs large-scale GWAS, including 1) the pre-phasing of haplotypes, 2) the imputation of genotypes using multiple reference panels, 3) the association testing for different inheritance models and 4) cross-phenotype analysis when more than one phenotype is available in the cohort, to finally, 5) generate summary statistics tables and graphic representations of the results (Figure 14), for both the autosomes and the X chromosome.

1.1.1 First steps before phasing haplotypes

GUIDANCE starts splitting the QCed PLINK input files into the chromosomes specified in the configuration file as shown in the following command line:

```
plink –bed clean_snps_subjects_aut_hg19_chr1_chr23_final_for_impute.bed --bim
clean_snps_subjects_aut_hg19_chr1_chr23_final_for_impute.bim --fam
clean_snps_subjects_aut_hg19_chr1_chr23_final_for_impute.fam --chr 23 --out
mixed_GERA_ASTHMA_chr_23 --make-bed
```

Table 6. Tools and versions included in GUIDANCE.

Software	Version
Plink	1.9
SHAPEIT	v2 r727
Eagle	2.4
IMPUTE	2.3.2
Minimac	4
Qctool	1.4
SNPTEST	2.5
Tabix	1.9
Bgzip	1.9
Samtools	1.5
Bcftools	1.8
R	≥ 3.5.0

If the X chromosome analysis is required in the configuration file, the commands

```
/usr/bin/plink --noweb --bed mixed/Chr_23/mixed_GERA_ASTHMA_chr_23.bed -bim
mixed_GERA_ASTHMA_chr_23.bim --fam mixed_GERA_ASTHMA_chr_23.fam --filter-females --
out mixed_GERA_ASTHMA_chr_23_females --make-bed
```

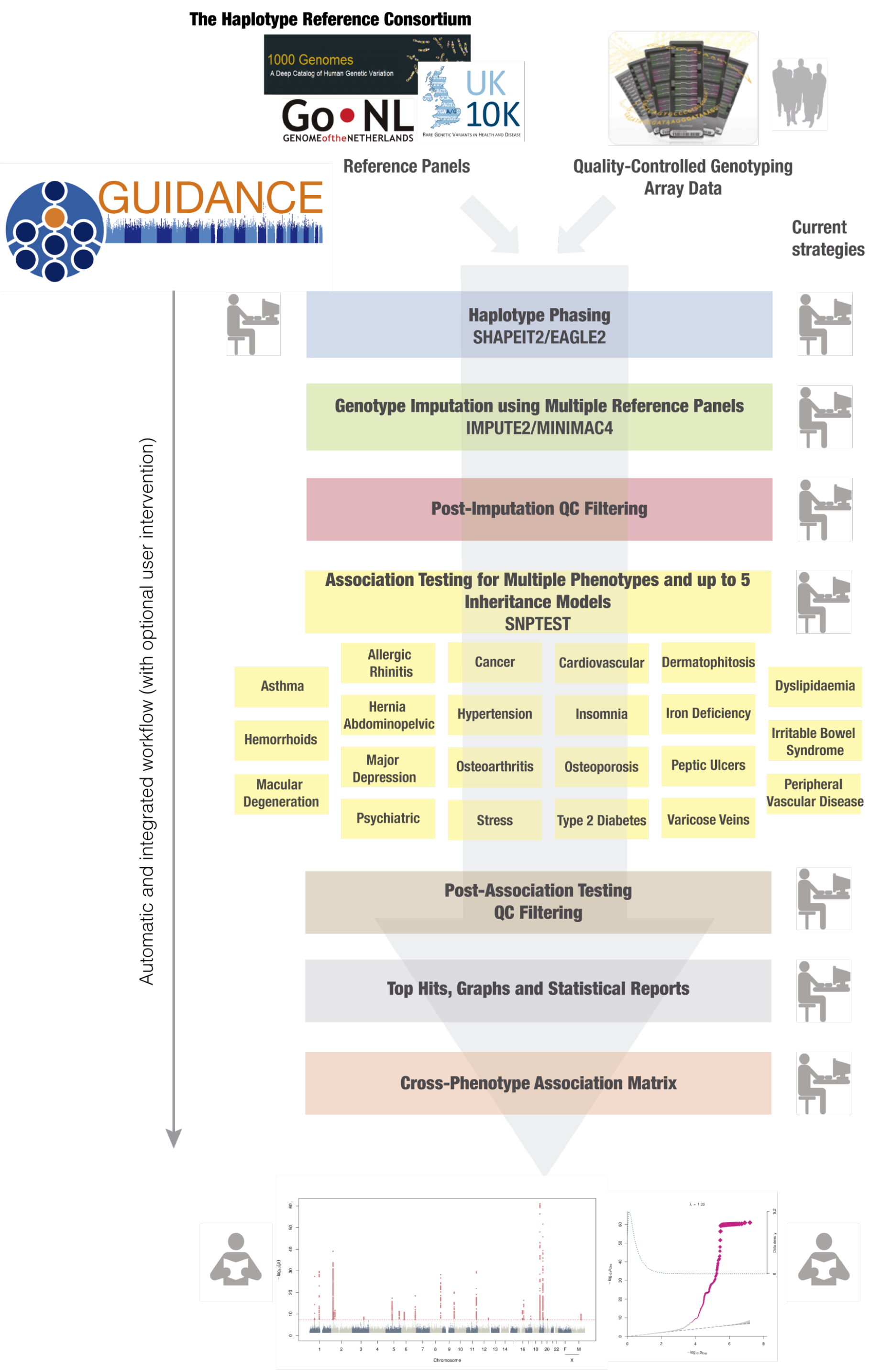
```
plink --noweb --bed mixed_GERA_ASTHMA_chr_23.bed --bim mixed_GERA_ASTHMA_chr_23.bim
--fam mixed/Chr_23/mixed_GERA_ASTHMA_chr_23.fam --filter-males --out
mixed_GERA_ASTHMA_chr_23_males --make-bed
```

split the X chromosome, encoded as “23” in the PLINK files, into males and females to allow the analysis of both males and females separately and together. From now on, X chromosome for both males and females, X chromosome for males and X chromosome for females, are treated as if they were three independent chromosomes through the pipeline.

In order to avoid strand orientation problems, the user can specify in the configuration file if C/G A/T SNPs need to be removed from genotyping data. The command

```
java createRsIdList mixed_GERA_chr_23.bim YES mixed_chr_23.pairs BED
```

creates a list of A/T and C/G SNPs from the .bim file using a method written in java.



The Haplotype Reference Consortium

Reference Panels

Quality-Controlled Genotyping Array Data

Current strategies

Automatic and integrated workflow (with optional user intervention)

Haplotype Phasing
SHAPEIT2/EAGLE2

Genotype Imputation using Multiple Reference Panels
IMPUTE2/MINIMAC4

Post-Imputation QC Filtering

Association Testing for Multiple Phenotypes and up to 5 Inheritance Models
SNPTEST

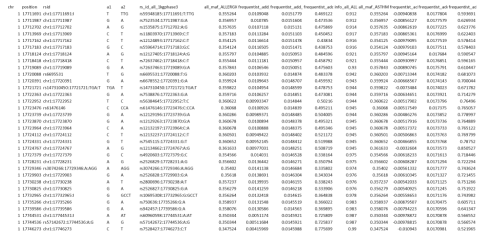
Asthma, Allergic Rhinitis, Cancer, Cardiovascular, Dermatophitosis, Dyslipidaemia, Hemorrhoids, Hernia Abdominopelvic, Hypertension, Insomnia, Iron Deficiency, Irritable Bowel Syndrome, Macular Degeneration, Major Depression, Osteoarthritis, Osteoporosis, Peptic Ulcers, Peripheral Vascular Disease, Psychiatric, Stress, Type 2 Diabetes, Varicose Veins

Post-Association Testing QC Filtering

Top Hits, Graphs and Statistical Reports

Cross-Phenotype Association Matrix

1000 Genomes
A Deep Catalog of Human Genetic Variation
Go•NL
GENOMEoftheNETHERLANDS
UK 10K
RARE GENETIC VARIANTS IN HEALTH AND DISEASE
GUIDANCE
−log10(p)
0
10
20
30
40
50
60
1
2
3
4
5
6
7
8
9
10
11
12
13
14
16
18
20
22
F
M
X
Chromosome
λ = 1.03
−log10 pObs
−log10 pExp
Data density
0
6.2



chr	position	rsid	a1	a2	rs_id_all_1kgphase3	all_maf_ALLERGI	frequentist_add	frequentist_add_	frequentist_adc	info_all_ALL	all_maf_ASTHM	frequentist_ac	frequentist_adi	frequentist_ac
1	17711691	chr1:17711691:I	T	TTG	rs59348185:17711691:T:TTG	0.355264	0.0109048	0.0151779	0.469122	0.912	0.355264	-0.00940838	0.0177804	0.593691
1	17711987	chr1:17711987	G	A	rs7523534:17711987:G:A	0.356957	0.010785	0.0151604	0.473536	0.912	0.356957	-0.00856127	0.0177579	0.626934
1	17712702	chr1:17712702	A	G	rs1535875:17712702:A:G	0.357635	0.0107118	0.015131	0.475869	0.914	0.357635	-0.00862619	0.0177225	0.623776
1	17713969	chr1:17713969	C	T	rs11803970:17713969:C:T	0.357183	0.0113284	0.0151103	0.450452	0.917	0.357183	-0.00865361	0.0176999	0.622403
1	17717162	chr1:17717162	C	T	rs12124893:17717162:C:T	0.354125	0.0116614	0.0151478	0.43834	0.916	0.354125	-0.00979095	0.0177519	0.578414
1	17717183	chr1:17717183	G	C	rs55964714:17717183:G:C	0.354124	0.0116505	0.0151471	0.438753	0.916	0.354124	-0.00979103	0.0177511	0.578403
1	17718124	chr1:17718124	A	G	rs12127405:17718124:A:G	0.355797	0.0104885	0.0150953	0.484596	0.921	0.355797	-0.00945164	0.017684	0.590547
1	17718418	chr1:17718418	C	T	rs72637462:17718418:C:T	0.355444	0.0111181	0.0150957	0.458792	0.921	0.355444	-0.00930997	0.0176851	0.596165
1	17719089	chr1:17719089	G	A	rs72637463:17719089:G:A	0.357843	0.0106546	0.0150051	0.475603	0.93	0.357843	-0.00890745	0.0175791	0.610447
1	17720088	rs6695531	T	G	rs6695531:17720088:T:G	0.360203	0.0103932	0.014874	0.483378	0.942	0.360203	-0.00713344	0.0174182	0.681073
1	17720391	chr1:17720391	G	A	rs6678552:17720391:G:A	0.359924	0.0109643	0.0148707	0.459592	0.943	0.359924	-0.00668567	0.0174143	0.700044
1	17721721	rs147310450:17721721:TGA:T	TGA	T	rs147310450:17721721:TGA:T	0.359822	0.0104954	0.0148599	0.478753	0.944	0.359822	-0.0073484	0.0174023	0.671782
1	17722363	chr1:17722363	G	A	rs7538876:17722363:G:A	0.359716	0.0106257	0.014851	0.473081	0.944	0.359716	-0.00634651	0.0173921	0.714279
1	17722952	chr1:17722952	T	C	rs56384645:17722952:T:C	0.360622	0.00993347	0.014844	0.50216	0.944	0.360622	-0.00517902	0.0173796	0.76496
1	17723476	rs61476146	C	CCA	rs61476146:17723476:C:CCA	0.36068	0.0100926	0.014839	0.495211	0.945	0.36068	-0.00517549	0.017375	0.765057
1	17723739	chr1:17723739	G	A	rs12129196:17723739:G:A	0.360286	0.00989371	0.0148485	0.504005	0.944	0.360286	-0.00486276	0.0173852	0.778997
1	17723870	chr1:17723870	G	A	rs12129263:17723870:G:A	0.360678	0.0100894	0.0148378	0.495321	0.945	0.360678	-0.00517916	0.0173736	0.764889
1	17723964	chr1:17723964	C	A	rs12132197:17723964:C:A	0.360678	0.0100888	0.0148375	0.495346	0.945	0.360678	-0.00517372	0.0173733	0.765122
1	17724112	chr1:17724112	C	T	rs12132237:17724112:C:T	0.360501	0.00949422	0.0148402	0.521172	0.945	0.360501	-0.00506863	0.0173763	0.769799
1	17724331	chr1:17724331	G	T	rs7545115:17724331:G:T	0.360652	0.00952145	0.0148412	0.519988	0.945	0.360652	-0.00466855	0.0173768	0.78752
1	17724767	chr1:17724767	A	G	rs12134662:17724767:A:G	0.361633	0.00977031	0.0148251	0.508719	0.945	0.361633	-0.0032604	0.0173573	0.850527
1	17727379	chr1:17727379	G	C	rs4920603:17727379:G:C	0.354566	0.014031	0.0146528	0.338164	0.975	0.354566	-0.00618233	0.0171613	0.718446
1	17728231	chr1:17728231	A	G	rs2526829:17728231:A:G	0.356602	0.0136442	0.0146271	0.350794	0.975	0.356602	-0.00608287	0.0171294	0.722294
1	17729346	rs3074266:17729346:A:AGG	A	AGG	rs3074266:17729346:A:AGG	0.35402	0.0141138	0.0146684	0.335802	0.974	0.35402	-0.00561332	0.0171777	0.743624
1	17729903	chr1:17729903	G	A	rs2526828:17729903:G:A	0.35618	0.0138691	0.0146304	0.343034	0.976	0.35618	-0.00610345	0.0171327	0.721455
1	17730238	chr1:17730238	A	T	rs2800696:17730238:A:T	0.357237	0.0139935	0.0146155	0.338243	0.976	0.357237	-0.00542033	0.0171125	0.751266
1	17730825	chr1:17730825	G	A	rs2526827:17730825:G:A	0.356279	0.0141259	0.0146218	0.333906	0.976	0.356279	-0.00540925	0.0171245	0.751922
1	17732965	chr1:17732965:I	G	GCCT	rs10695308:17732965:G:GCCT	0.356264	0.0132418	0.014615	0.364838	0.978	0.356264	-0.00558653	0.0171176	0.743982
1	17735266	chr1:17735266	G	A	rs750636:17735266:G:A	0.358937	0.0131548	0.0145519	0.366022	0.983	0.358937	-0.00879507	0.0170475	0.605711
1	17739586	chr1:17739586	G	A	rs942457:17739586:G:A	0.358076	0.0130586	0.014563	0.369895	0.983	0.358076	-0.00794223	0.0170596	0.641347
1	17744531	chr1:17744531:I	A	AT	rs60960598:17744531:A:AT	0.350344	0.0051174	0.0145921	0.725809	0.987	0.350344	-0.00978872	0.0170878	0.566552
1	17744536	rs57142672:17744536:A:G	A	G	rs57142672:17744536:A:G	0.350344	0.00511684	0.0145921	0.725837	0.987	0.350344	-0.00978815	0.0170878	0.566574
1	17746273	chr1:17746273	C	T	rs7528427:17746273:C:T	0.347524	0.00415969	0.0145988	0.775699	0.99	0.347524	-0.010943	0.0170981	0.521965

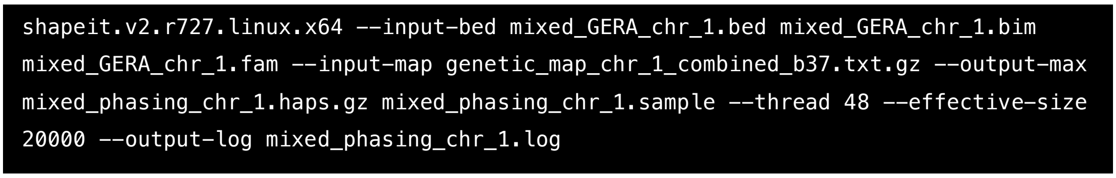
Figure 14. Schematic representation of GUIDANCE compared to current GWAS workflows.

At both sides of the workflow, the steps that typically require manual intervention in the case of current strategies are displayed (right) and compared with GUIDANCE requirements of user intervention (left), which allows an automatic execution. Starting with Quality Controlled genetic data (top), through phasing and imputation using multiple panels, and association testing considering multiple phenotypes and inheritance models. GUIDANCE finishes with summary statistics and graphical representation of the results (bottom). Multiple genotypes displayed correspond to those found in GERA.

Once data have been split, pre-phasing is conducted using SHAPEIT2 or Eagle as required by the user in the configuration file (Figure 15). Each chromosome is phased in independent nodes, and an extra level of parallelization is achieved using the --thread flag in the SHAPEIT2 command line and the --numThreads flag in the Eagle command line.

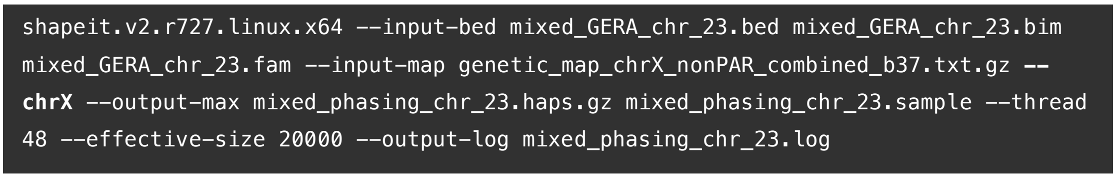
1.1.2 Phasing haplotypes using SHAPEIT2

The following command line corresponds to the autosomes when using SHAPEIT2:



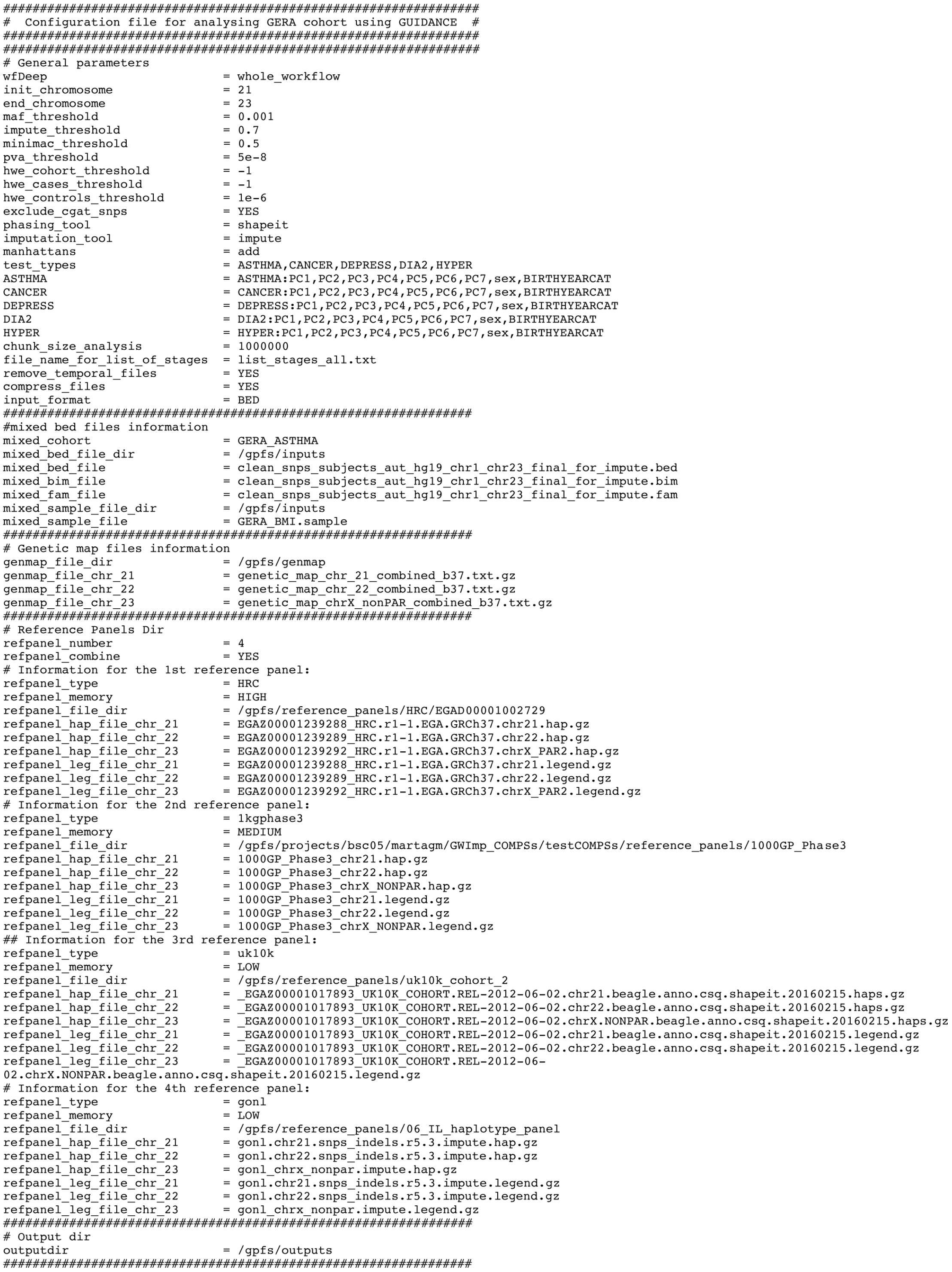
```
shapeit.v2.r727.linux.x64 --input-bed mixed_GERA_chr_1.bed mixed_GERA_chr_1.bim
mixed_GERA_chr_1.fam --input-map genetic_map_chr_1_combined_b37.txt.gz --output-max
mixed_phasing_chr_1.haps.gz mixed_phasing_chr_1.sample --thread 48 --effective-size
20000 --output-log mixed_phasing_chr_1.log
```

To set up SHAPEIT2 for the X chromosome, --chrX flag is required to only phase female samples and impute missing data in male samples as shown in the following command line:



```
shapeit.v2.r727.linux.x64 --input-bed mixed_GERA_chr_23.bed mixed_GERA_chr_23.bim
mixed_GERA_chr_23.fam --input-map genetic_map_chrX_nonPAR_combined_b37.txt.gz --
chrX --output-max mixed_phasing_chr_23.haps.gz mixed_phasing_chr_23.sample --thread
48 --effective-size 20000 --output-log mixed_phasing_chr_23.log
```

Although males do not require phasing the genotypes for the X chromosome, the commands



```
################################################################
#  Configuration file for analysing GERA cohort using GUIDANCE  #
################################################################
################################################################
# General parameters
wfDeep                         = whole_workflow
init_chromosome                = 21
end_chromosome                 = 23
maf_threshold                  = 0.001
impute_threshold               = 0.7
minimac_threshold              = 0.5
pva_threshold                  = 5e-8
hwe_cohort_threshold           = -1
hwe_cases_threshold            = -1
hwe_controls_threshold         = 1e-6
exclude_cgat_snps              = YES
phasing_tool                   = shapeit
imputation_tool                = impute
manhattans                     = add
test_types                     = ASTHMA,CANCER,DEPRESS,DIA2,HYPER
ASTHMA                         = ASTHMA:PC1,PC2,PC3,PC4,PC5,PC6,PC7,sex,BIRTHYEARCAT
CANCER                         = CANCER:PC1,PC2,PC3,PC4,PC5,PC6,PC7,sex,BIRTHYEARCAT
DEPRESS                        = DEPRESS:PC1,PC2,PC3,PC4,PC5,PC6,PC7,sex,BIRTHYEARCAT
DIA2                           = DIA2:PC1,PC2,PC3,PC4,PC5,PC6,PC7,sex,BIRTHYEARCAT
HYPER                          = HYPER:PC1,PC2,PC3,PC4,PC5,PC6,PC7,sex,BIRTHYEARCAT
chunk_size_analysis            = 1000000
file_name_for_list_of_stages   = list_stages_all.txt
remove_temporal_files          = YES
compress_files                 = YES
input_format                   = BED
###############################################################
#mixed bed files information
mixed_cohort                   = GERA_ASTHMA
mixed_bed_file_dir             = /gpfs/inputs
mixed_bed_file                 = clean_snps_subjects_aut_hg19_chr1_chr23_final_for_impute.bed
mixed_bim_file                 = clean_snps_subjects_aut_hg19_chr1_chr23_final_for_impute.bim
mixed_fam_file                 = clean_snps_subjects_aut_hg19_chr1_chr23_final_for_impute.fam
mixed_sample_file_dir          = /gpfs/inputs
mixed_sample_file              = GERA_BMI.sample
###############################################################
# Genetic map files information
genmap_file_dir                = /gpfs/genmap
genmap_file_chr_21             = genetic_map_chr_21_combined_b37.txt.gz
genmap_file_chr_22             = genetic_map_chr_22_combined_b37.txt.gz
genmap_file_chr_23             = genetic_map_chrX_nonPAR_combined_b37.txt.gz
###############################################################
# Reference Panels Dir
refpanel_number                = 4
refpanel_combine               = YES
# Information for the 1st reference panel:
refpanel_type                  = HRC
refpanel_memory                = HIGH
refpanel_file_dir              = /gpfs/reference_panels/HRC/EGAD00001002729
refpanel_hap_file_chr_21       = EGAZ00001239288_HRC.r1-1.EGA.GRCh37.chr21.hap.gz
refpanel_hap_file_chr_22       = EGAZ00001239289_HRC.r1-1.EGA.GRCh37.chr22.hap.gz
refpanel_hap_file_chr_23       = EGAZ00001239292_HRC.r1-1.EGA.GRCh37.chrX_PAR2.hap.gz
refpanel_leg_file_chr_21       = EGAZ00001239288_HRC.r1-1.EGA.GRCh37.chr21.legend.gz
refpanel_leg_file_chr_22       = EGAZ00001239289_HRC.r1-1.EGA.GRCh37.chr22.legend.gz
refpanel_leg_file_chr_23       = EGAZ00001239292_HRC.r1-1.EGA.GRCh37.chrX_PAR2.legend.gz
# Information for the 2nd reference panel:
refpanel_type                  = 1kgphase3
refpanel_memory                = MEDIUM
refpanel_file_dir              = /gpfs/projects/bsc05/martagm/GWImp_COMPSs/testCOMPSs/reference_panels/1000GP_Phase3
refpanel_hap_file_chr_21       = 1000GP_Phase3_chr21.hap.gz
refpanel_hap_file_chr_22       = 1000GP_Phase3_chr22.hap.gz
refpanel_hap_file_chr_23       = 1000GP_Phase3_chrX_NONPAR.hap.gz
refpanel_leg_file_chr_21       = 1000GP_Phase3_chr21.legend.gz
refpanel_leg_file_chr_22       = 1000GP_Phase3_chr22.legend.gz
refpanel_leg_file_chr_23       = 1000GP_Phase3_chrX_NONPAR.legend.gz
## Information for the 3rd reference panel:
refpanel_type                  = uk10k
refpanel_memory                = LOW
refpanel_file_dir              = /gpfs/reference_panels/uk10k_cohort_2
refpanel_hap_file_chr_21       = _EGAZ00001017893_UK10K_COHORT.REL-2012-06-02.chr21.beagle.anno.csq.shapeit.20160215.haps.gz
refpanel_hap_file_chr_22       = _EGAZ00001017893_UK10K_COHORT.REL-2012-06-02.chr22.beagle.anno.csq.shapeit.20160215.haps.gz
refpanel_hap_file_chr_23       = _EGAZ00001017893_UK10K_COHORT.REL-2012-06-02.chrX.NONPAR.beagle.anno.csq.shapeit.20160215.haps.gz
refpanel_leg_file_chr_21       = _EGAZ00001017893_UK10K_COHORT.REL-2012-06-02.chr21.beagle.anno.csq.shapeit.20160215.legend.gz
refpanel_leg_file_chr_22       = _EGAZ00001017893_UK10K_COHORT.REL-2012-06-02.chr22.beagle.anno.csq.shapeit.20160215.legend.gz
refpanel_leg_file_chr_23       = _EGAZ00001017893_UK10K_COHORT.REL-2012-06-
02.chrX.NONPAR.beagle.anno.csq.shapeit.20160215.legend.gz
# Information for the 4th reference panel:
refpanel_type                  = gonl
refpanel_memory                = LOW
refpanel_file_dir              = /gpfs/reference_panels/06_IL_haplotype_panel
refpanel_hap_file_chr_21       = gonl.chr21.snps_indels.r5.3.impute.hap.gz
refpanel_hap_file_chr_22       = gonl.chr22.snps_indels.r5.3.impute.hap.gz
refpanel_hap_file_chr_23       = gonl_chrx_nonpar.impute.hap.gz
refpanel_leg_file_chr_21       = gonl.chr21.snps_indels.r5.3.impute.legend.gz
refpanel_leg_file_chr_22       = gonl.chr22.snps_indels.r5.3.impute.legend.gz
refpanel_leg_file_chr_23       = gonl_chrx_nonpar.impute.legend.gz
###############################################################
# Output dir
outputdir                      = /gpfs/outputs
###############################################################
```

Figure 15. GUIDANCE configuration file for chromosome 21 to chromosome X (encoded 23) for GERA cohort.

All the thresholds and pathways are specified in advance to free the user from constant intervention. Of note, no additional handling is needed to include the analysis of the X chromosome.

```
shapeit.v2.r727.linux.x64 --input-bed mixed_GERA_chr_23_males.bed
mixed_GERA_chr_23_males.bim mixed_GERA_chr_23_males.fam --input-map
genetic_map_chrX_nonPAR_combined_b37.txt.gz --chrX --output-max
mixed_phasing_chr_23_males.haps.gz mixed_phasing_chr_23_males.sample --thread 48 --
effective-size 20000 --output-log mixed_phasing_chr_23_males.log
```

```
shapeit.v2.r727.linux.x64 --input-bed mixed_GERA_chr_23_females.bed
mixed_GERA_chr_23_females.bim mixed_GERA_chr_23_females.fam --input-map
genetic_map_chrX_nonPAR_combined_b37.txt.gz --chrX --output-max
mixed_phasing_chr_23_females.haps.gz mixed_phasing_chr_23_females.sample --thread
48 --effective-size 20000 --output-log mixed_phasing_chr_23_females.log
```

perform a separately pre-phasing for males and females to avoid format issues in further steps.

1.1.3 Generating a new sample file

After phasing the genotypes into haplotypes, a new sample file is generated since after splitting the X chromosome into males and females the number of individuals is different than in the original sample file. However, since it is not computational demanding as well to ensure consistency, a sample file is also generated for the autosomes and the X chromosome with males and females together. This is performed using a method written in java as shown in the following command lines:

```
java newSample.jar GERA.sample mixed_phasing_chr_23.sample
new_mixed_phasing_chr_23.sample PC1,PC2,PC3,PC4,PC5,PC6,PC7,sex,BIRTHYEARCAT
INSOMNIA,MACDEGEN,DERMATOPHYTOSIS,VARICOSE_VEINS,HEMORRHOIDS,PSYCHIATRIC,PEPTIC_ULC
ERS,HYPER,OSTIOA,CANCER,CARD,ALLERGIC_RHINITIS,IRON_DEFICIENCY,DEPRESS,HERNIA_ABDOM
INOPELVIC,DIA2,ASTHMA,STRESS,IRRITABLE_BOWEL,OSTIOP,DYSLIPID,PVD
```

```
java newSample.jar GERA.sample mixed_phasing_chr_23_males.sample
new_mixed_phasing_chr_23_males.sample PC1,PC2,PC3,PC4,PC5,PC6,PC7,sex,BIRTHYEARCAT
INSOMNIA,MACDEGEN,DERMATOPHYTOSIS,VARICOSE_VEINS,HEMORRHOIDS,PSYCHIATRIC,PEPTIC_ULC
ERS,HYPER,OSTIOA,CANCER,CARD,ALLERGIC_RHINITIS,IRON_DEFICIENCY,DEPRESS,HERNIA_ABDOM
INOPELVIC,DIA2,ASTHMA,STRESS,IRRITABLE_BOWEL,OSTIOP,DYSLIPID,PVD
```

```
java newSample GERA.sample mixed_phasing_chr_23_females.sample
new_mixed_phasing_chr_23_females.sample
PC1,PC2,PC3,PC4,PC5,PC6,PC7,sex,BIRTHYEARCAT
INSOMNIA,MACDEGEN,DERMATOPHYTOSIS,VARICOSE_VEINS,HEMORRHOIDS,PSYCHIATRIC,PEPTIC_ULC
ERS,HYPER,OSTIOA,CANCER,CARD,ALLERGIC_RHINITIS,IRON_DEFICIENCY,DEPRESS,HERNIA_ABDOM
INOPELVIC,DIA2,ASTHMA,STRESS,IRRITABLE_BOWEL,OSTIOP,DYSLIPID,PVD
```

1.2 Imputing genotypes using multiple reference panels

Genotype imputation is performed using IMPUTE2 or MINIMAC4 as required by the user in the configuration file. To maximize the parallelization without compromising the accuracy of the imputation, chromosomes are split into chunks of the size specified by the user in the configuration file. In the command line examples in the 1.2.1 section, chunks of 1,000,000 base pairs are analyzed.

Moreover, if the user has specified multiple reference panels in the configuration file, the imputation runs for each panel separately in parallel. As different panels required different memory constraints this can also be pre-set in the configuration file (e.g., “HIGH” for HRC, “MEDIUM” for 1000 G phase 3 and “LOW” for UK10K and GoNL) to ensure the proper utilization of the available resources. Memory and CPU usage for each step can be modified in an additional file that facilitates its manipulation by non-expert users.

1.2.1 Imputing genotypes using IMPUTE2

As stated in previous sections, no additional handling is required when using IMPUTE2 for genotype imputation, since both SHAPEIT2 and Eagle outputs are in the Oxford HAPS/SAMPLE format.

Hence, autosomes are imputed using the pre-phased haplotypes as input, excluding A/T and C/G SNPs from the phased data to avoid strand orientation issues, (therefore, A/T and C/G SNPs will be imputed) as shown in the following command line:

```
impute2 -use_prephased_g -m mixed_genetic_map_chr_22.txt -h
EGAZ00001017893_UK10K_COHORT.REL-2012-06-
02.chr22.beagle.anno.csq.shapeit.20160215.haps.gz -l
EGAZ00001017893_UK10K_COHORT.REL-2012-06-
02.chr22.beagle.anno.csq.shapeit.20160215.legend.gz -known_haps_g
mixed_phasing_chr_22.haps.gz -int 1 1000000 -exclude_snps_g mixed_chr_22.pairs -
impute_excluded -Ne 20000 -o chr_22_mixed_uk10k_1_1000000.impute.gz -i
chr_22_mixed_uk10k_1_1000000.impute_info -r
chr_22_mixed_uk10k_1_1000000.impute_summary -w
chr_22_mixed_uk10k_1_1000000.impute_warnings -no_sample_qc_info -o_gz
```

For the X chromosome, IMPUTE2 requires the -chrX flag as well as the sample file with the sex information.

```
impute2 -use_prephased_g -m genetic_map_chrX_nonPAR_combined_b37.txt.gz -h
EGAZ00001239292_HRC.r1-1.EGA.GRCh37.chrX_PAR2.hap.gz -l EGAZ00001239292_HRC.r1-
1.EGA.GRCh37.chrX_PAR2.legend.gz -known_haps_g mixed_phasing_chr_23.haps.gz -
sample_g new_mixed_phasing_chr_23.sample -int 1 1000000 -chrX -exclude_snps_g
mixed_chr_23.pairs -impute_excluded -Ne 20000 -o chr_23_mixed_HRC_1_1000000.impute
-i chr_23_mixed_HRC_1_1000000.impute_info -r
chr_23_mixed_HRC_1_1000000.impute_summary -w
chr_23_mixed_HRC_1_1000000.impute_warnings -no_sample_qc_info -o_gz
```

1.2.2 Imputing genotypes using MINIMAC4

The command

```
minimac4 --refHaps EGAZ00001017893_UK10K_COHORT.REL-2012-06-
02.chr22.beagle.anno.csq.shapeit.20160215.m3vcf.gz --haps
mixed_phasing_filtered_chr_22.vcf.gz --start 1 --end 1000000 --chr 22 --window
500000 --prefix chr_22_mixed_uk10k_1_1000000_minimac --log --allTypedSites --
noPhoneHome --format GT,DS,GP --nobgzip
```

performs the imputation using MINIMAC4. The outputs, in VCF format, will be in estimated most likely genotype (GT), estimated alternate allele dosage (DS) and estimated posterior genotype probabilities (GP). GP is the same estimate from IMPUTE2, thus giving comparable results.

Unfortunately, MINIMAC (neither 3 nor 4) cannot be used to impute the X chromosome since we detect an inconsistency in how males are coded. Therefore, even when selecting MINIMAC4 as the tool for imputation, if the X chromosome is in the analysis, this will be imputed using IMPUTE2 instead.

1.2.3 Filtering imputed variants according to the imputation accuracy

After imputation, variants ID from the genotyping array and each panel are substitute for a new ID based on the chromosome, the position and the alleles, using a bash command line.

Moreover, a list of variants that pass the imputation quality threshold (*info score* for IMPUTE or *Rsq* for MINIMAC4), as specified by the user in the configuration file, is created using a java method following the command line

```
java filterByInfo impute chr_23_mixed_HRC_1_1000000.impute_info
chr_23_mixed_HRC_1_1000000_filtered_rsid.txt 0.7
```

Thereafter, QCTOOL is used to keep these variants from the imputed genotypes alongside those that pass the MAF threshold specified in the configuration file.

```
qctool1.4 -g chr_23_mixed_HRC_1_1000000.impute.gz -og
chr_23_mixed_HRC_1_1000000_filtered.impute.gz -incl-rsids
chr_23_mixed_HRC_1_1000000_filtered_rsid.txt -omit-chromosome -force -log
chr_23_mixed_HRC_1_1000000_filtered.impute.log -maf 0.001 1
```

The same command can be used with MINIMAC4 results, as QCTOOL accepts VCF with genotype probabilities as input, specifying it with the "-vcf-genotype-field GP" flag in the command line.

The QCTOOL usefulness goes beyond filtering variants as it allows the homogenization of the different formats for later analysis using SNPTEST. After QCTOOL, both HAPS/SAMPLE from IMPUTE2 and VCF from MINIMAC4 are converted to GEN format.

1.3 The association testing including non-additive inheritance models

In GUIDANCE, the association test is performed using SNPTEST as it allows single-variant logistic regression adjusting for multiple covariates for both the autosomes and the X chromosome, and it allows multiple models of inheritance.

Hence, the command

```
snptest_v2.5 -data chr_1_mixed_1kgphase3_1_1000000_filtered.impute.gz
new_mixed_phasing_chr_1.sample -o
chr_1_ALLERGIC_RHINITIS_1kgphase3_1_1000000_snptest.out.gz -pheno ALLERGIC_RHINITIS
-cov_names PC1 PC2 PC3 PC4 PC5 PC6 PC7 sex BIRTHYEARCAT -hwe -log
chr_1_ALLERGIC_RHINITIS_1kgphase3_1_1000000_snptest.log -method em -frequentist 1 2
3 4 5
```

allows the analysis of five different models coded as 1=Additive, 2=Dominant, 3=Recessive, 4=General and 5=Heterozygote using an EM algorithm to estimate the parameters in the missing data likelihood for the model when analyzing the autosomes. The user specifies in the configuration file which inheritance models wants in the association test, and the command is consequently generated.

For the X chromosome, the command

```
snptest_v2.5 -data chr_23_mixed_HRC_1_1000000_males_filtered.impute.gz
new_mixed_phasing_chr_23_males.sample -o
chr_23_ALLERGIC_RHINITIS_HRC_1_1000000_males_snptest.out.gz -pheno
ALLERGIC_RHINITIS -cov_names PC1 PC2 PC3 PC4 PC5 PC6 PC7 sex BIRTHYEARCAT -hwe -log
chr_23_ALLERGIC_RHINITIS_HRC_1_1000000_males_snptest.log -method newml -
assume_chromosome X -stratify_on sex -frequentist 1
```

uses -method newml to ignore samples with missing sex or males encoded wrongly (males should be coded 0 / 1, as homozygote females), and to assume a model of full X inactivation. Hence, the logistic regression model assumes a complete inactivation of one allele in females and equal effect size between males and females.

Therefore, in order to allow for heterogeneity between males and females and to allow a complete inactivation of the X chromosome in females, the -stratify_on option is used to separate the effects and the baselines for males and females specifying the same variable (i.e., “sex”) as a covariate into the -cov_names flag.

1.3.1 Filtering the results from the association test

After running SNPTEST, an additional filtering step is applied by chunk using a java method. The command

```
java filterByAll minimac chr_22_ALLERGIC_RHINITIS_uk10k_1_1000000_summary.txt.gz
chr_22_ALLERGIC_RHINITIS_uk10k_1_1000000_summary_filtered.txt.gz 0.001 0.0 -1.0 -
1.0 1.0E-6 uk10k
```

filters by the thresholds specified by the user in the configuration file, such as MAF < 0.001 and HWE for controls ≤ 1 × 10^{-6} in this example.

1.3.2 Combining the results from different reference panels to get the final set of variants

One of the main features of GUIDANCE is the integration of the results from multiple reference panels after filtering the association testing results. To do so, when a variant is found in more than one reference panel, GUIDANCE selects the one from the reference panel with the best imputation accuracy (Figure 16).

To maximize the parallelization, this is done by chunk, and this method was primarily implemented in R, and finally in java, into the GUIDANCE framework as shown in the following command line:

```
java combinePanelsComplex
chr_22_ALLERGIC_RHINITIS_HRC_1_1000000_summary_filtered.txt.gz
chr_22_ALLERGIC_RHINITIS_1kgphase3_1_1000000_summary_filtered.txt.gz
chr_22_ALLERGIC_RHINITIS_uk10k_1_1000000_summary_filtered.txt.gz
chr_22_ALLERGIC_RHINITIS_gonl_1_1000000_summary_filtered.txt.gz
filteredByAll_results_ALLERGIC_RHINITIS_GERA_300_HRC_1kgphase3_uk10k_gonl_chr_22_1_
1000000_combined.txt.gz 1 1000000
```

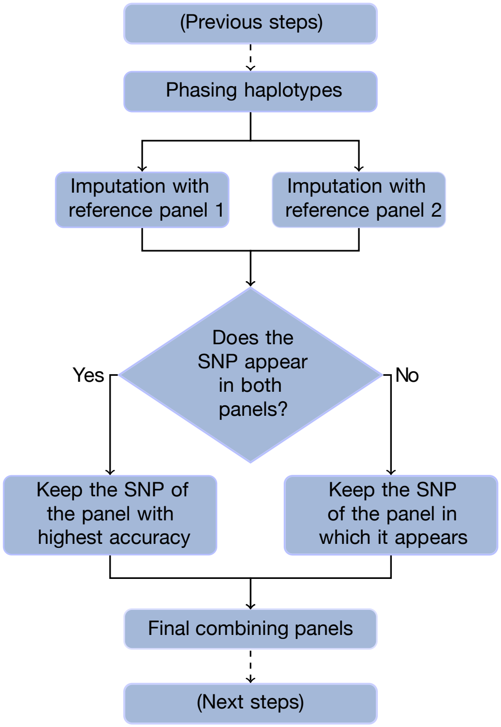

(Previous steps)
Phasing haplotypes
Imputation with reference panel 1
Imputation with reference panel 2
Does the SNP appear in both panels?
Yes
No
Keep the SNP of the panel with highest accuracy
Keep the SNP of the panel in which it appears
Final combining panels
(Next steps)

Figure 16. Flow-chart of how the results from several reference panels are combined.

In case a given variant is only imputed in one panel, the genotypes are selected from that panel. However, if a variant is present in more than one panel, the genotype from the reference panel with the best imputation score is selected. All the variants must have an imputation score higher than the threshold that is specified in the configuration file.

1.4 Summary statistics tables and graphical representation

After obtaining the filtered association testing results for each chunk from each panel, and the combined final set of variants if requested, chunks are merged into a final file containing all the chromosomes to facilitate the management of the results. For the same reason, an additional file containing only the top variants, including a list of the inheritance models for which those variants pass the p-value threshold specified in the configuration file, is provided.

Moreover, to simplify the interpretation of the results, Q-Q plots and Manhattan plots, both in TIFF and PDF format, are generated using R as shown in the following command line:

```
Rscript qqplot_manhattan_all_models.R
ALLERGIC_RHINITIS_uk10k_condensed_chr_22_to_23.txt.gz
QQplot_ALLERGIC_RHINITIS_GERA_300_uk10k_add.pdf
manhattan_ALLERGIC_RHINITIS_GERA_300_uk10k_add.pdf
QQplot_ALLERGIC_RHINITIS_GERA_300_uk10k_add.tiff
manhattan_ALLERGIC_RHINITIS_GERA_300_uk10k_add.tiff frequentist_add_pvalue 5e-8
```

1.5 Cross-phenotype association

As emerging cohorts are usually encompassing multiple phenotypes, we implemented a cross-phenotype association test in GUIDANCE to assess if any top variant for a particular disease is also associated with an additional one. This is implemented in R.

The following command,

```
Rscript crossphenotype_crossmodel.R
tophits_merge_ALLERGIC_RHINITIS.txt,tophits_merge_ASTHMA.txt,
tophits_merge_MACDEGEN.txt,tophits_merge_CARD.txt,tophits_merge_DIA2.txt
cross_pheno_all.txt 5e-8 add,rec
```

results in a summary table taking into account the most significant inheritance models used for each variant. The p-value (i.e., 5×10^{-8}) is used to select the significant variants. However, to avoid false positives due to multiple testing, a new p-value threshold is internally calculated taking into account the number of regions and diseases analyzed.

2 The impact of non-additive genetic variants on age-related diseases across 60,000 individuals

Marta Guindo-Martínez[1,*], Ramon Amela[1,*], Silvia Bonàs-Guarch[1], Montserrat Puiggròs[1], Cecilia Salvoro[1], Caitlin E. Carey[2,3], Joanne B. Cole[4,5], Friman Sanchez[1], Cristian Ramon-Cortés[1], Jorge Ejarque[1], Carlos Díaz[1], Enric Tejedor[1], Rosa M. Badia[1,6], Duncan S. Palmer[2,3], Jose C. Florez[4,5,7], Josep M. Mercader[4,5,1,#], David Torrents[1,8,#]

[1] Barcelona Supercomputing Center (BSC), Barcelona, Spain
[2] Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, MA, USA
[3] Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA
[4] Programs in Metabolism and Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA, USA
[5] Diabetes Unit and Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA
[6] Artificial Intelligence Research Institute (IIIA), Spanish Council for Scientific Research (CSIC), Barcelona, Spain
[7] Department of Medicine, Harvard Medical School, Boston, MA, USA
[8] Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

[*] Both authors contributed equally to this work.
[#] These authors jointly directed this work.

In this section, there are the results of a manuscript in preparation describing our findings after applying our methodology, described in the previous section, to analyze the GERA cohort, the largest public cohort for age-related diseases to date.

2.1 Imputation using multiple reference panels

After applying quality control, 56,637 individuals of European Ancestry from the GERA cohort were analyzed using GUIDANCE (see Methods).

We imputed GERA genotypes using IMPUTE2 and GoNL, UK10K, 1000G Phase 3 and HRC reference panels, obtaining 22,1 M, 25,8 M, 88,3 M and 41,5 M variants, respectively. After applying an info score ≥ 0.7 and a minor allele frequency (MAF) > 0.001 filters, 11,2 M, 11,4 M, 13,1 M, and 11,7 M good quality variants remained for each panel, respectively (Figure 17a).

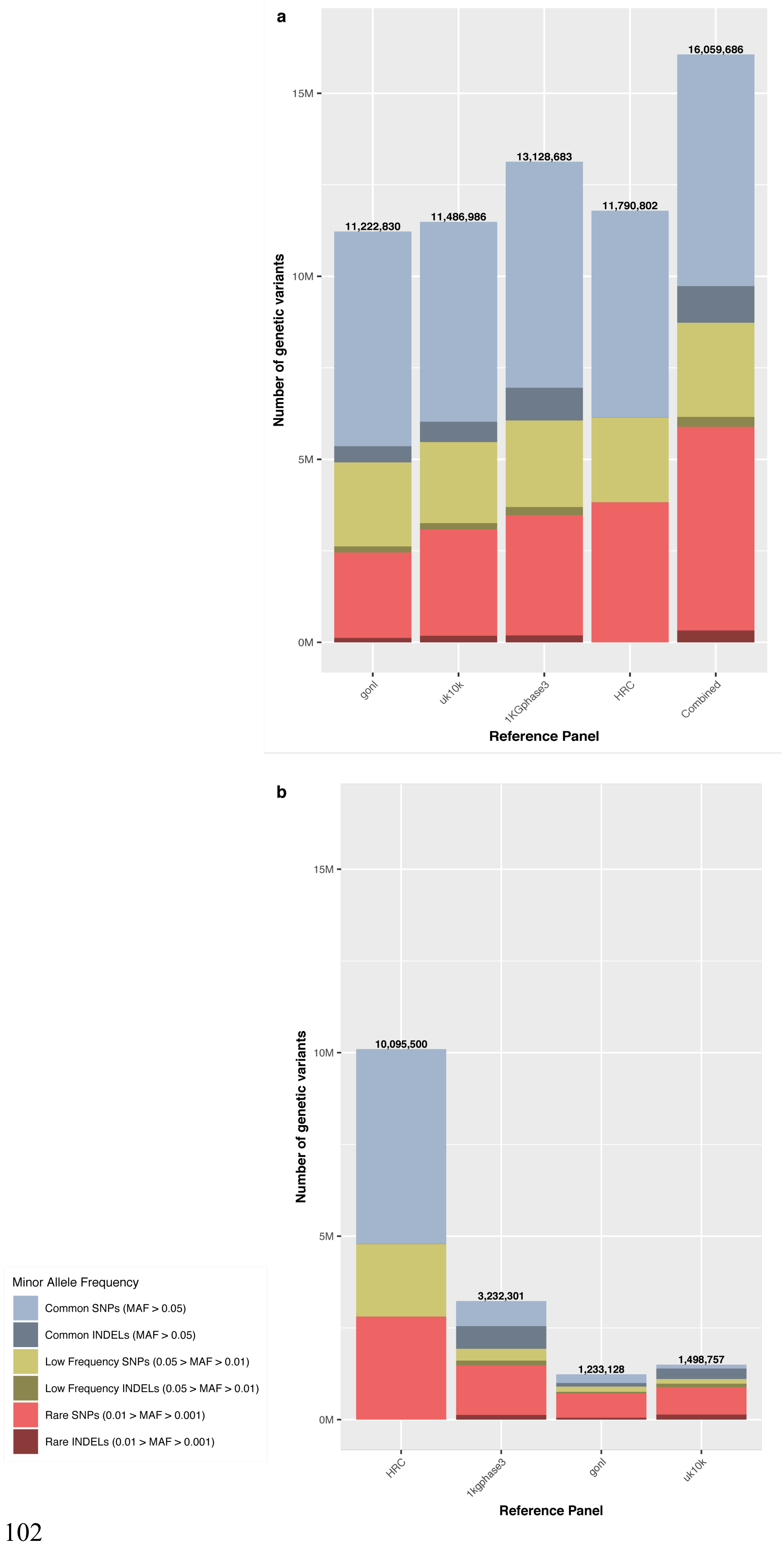
When combining the results from the four reference panels, we were able to test 16,059,686 variants for association, including 5,5 M of high quality rare variants (0.01 > MAF > 0.001), while only 2.3 M, 2.9 M, 3.2 M and 3.8 M rare variants were tested for association when using GoNL, UK10K, 1000G phase 3 and HRC alone (Figure 17a).

Among the four reference panels, HRC had higher imputation scores, as 10 M out of the 16 M final variants had the highest imputation accuracy when imputed with HRC (Figure 17b). However, more than 1.5 M variants from the 16 M obtained when combining the results were INDELs than could only be imputed by the other reference panels since HRC panel does not include INDELs.

2.2 Association testing for additive and non-additive inheritance models

By testing 16 M variants for association considering multiple inheritance models using GUIDANCE for 22 diseases, we found 94 associated loci at the genome-wide significance level ($p < 5 \times 10^{-8}$) (Supplementary Table 1, Supplementary Figure 1-22). Interestingly, the model with the most significant result for 37 out of these 94 variants was a non-additive model, and 20 loci were only genome-wide significant when non-additive models were tested (Figure 18 and Supplementary Table 1).

From these 94 associated loci, 68 loci had been previously reported. Of note, some of the well-known variants for certain diseases could be identified only by combining the results from several reference panels. In particular, only 67 of 94 loci were found with


a
Number of genetic variants
15M
10M
5M
0M
11,222,830
11,486,986
13,128,683
11,790,802
16,059,686
gonl
uk10k
1KGphase3
HRC
Combined
Reference Panel
b
Number of genetic variants
15M
10M
5M
0M
10,095,500
3,232,301
1,233,128
1,498,757
HRC
1kgphase3
gonl
uk10k
Reference Panel
Minor Allele Frequency
Common SNPs (MAF > 0.05)
Common INDELs (MAF > 0.05)
Low Frequency SNPs (0.05 > MAF > 0.01)
Low Frequency INDELs (0.05 > MAF > 0.01)
Rare SNPs (0.01 > MAF > 0.001)
Rare INDELs (0.01 > MAF > 0.001)

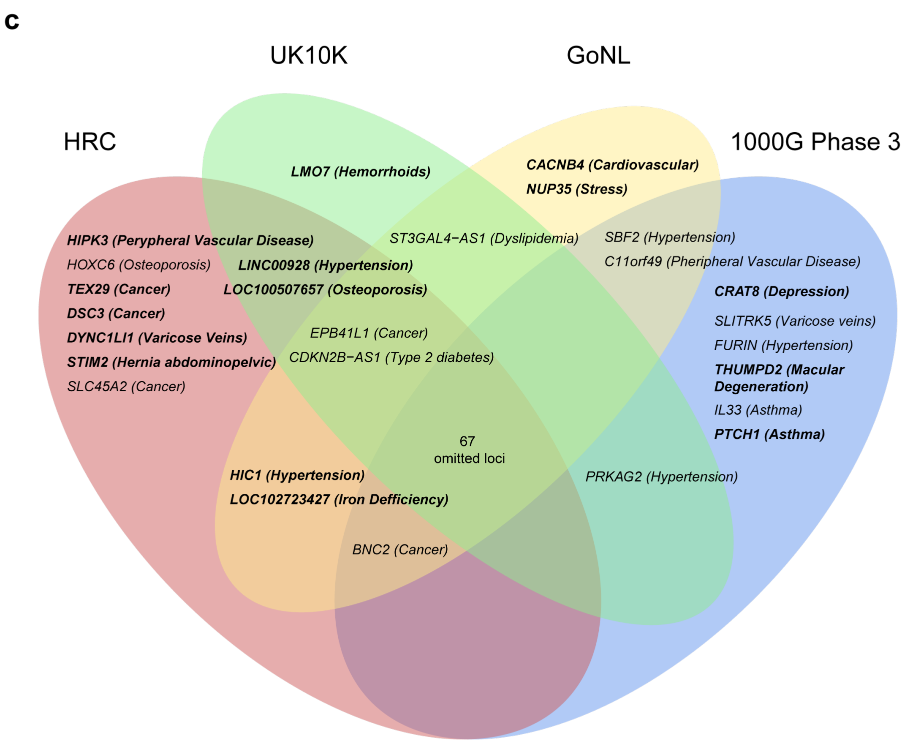


Figure 17. Graphical representation illustrating the benefits of combining the results from different reference panels.

a Comparison of the number of variants after the imputation with the four reference panels, covering common, low frequency and rare variants in different colors (info score ≥ 0.7). INDELs are also represented with the corresponding darker color depending on the MAF. As shown in the bar plot, the combination increases the final set of variants for association testing when compared with the results for each of the panels alone (GoNL, UK10K, 1000G Phase 3 or HRC), especially in the low and rare frequency spectrum. **b** Comparison of the contribution of each reference panel in the combined results. Each bar represents the number of variants that had the best imputation accuracy for a given reference panel. As seen in the figure, HRC shows the highest imputation accuracy for most of the variants. Nevertheless, all the reference panels contribute to the final result, especially for INDELs since HRC does not include them. All variants have an info score ≥ 0.7, MAF ≥ 0.001 and HWE for controls > 1 × 10^{-6}. **c** Venn Diagram illustrating the genome-wide significant loci that could be identified by each reference panel. Novel associations are depicted in bold. As shown in this figure, only 67 of the 94 GWAS significant loci were identified by the four reference panels, while 27 of them (28.7%) were only identified by one, two or three of the four panels.

all the four reference panels (71%) while 16 significant loci (17%) were identified only by one of the four reference panels (Figure 17c). In fact, only 81, 77, 74, and 77 loci would have been identified by using only HRC, 1000G Phase 3, UK10K or GoNL, respectively.

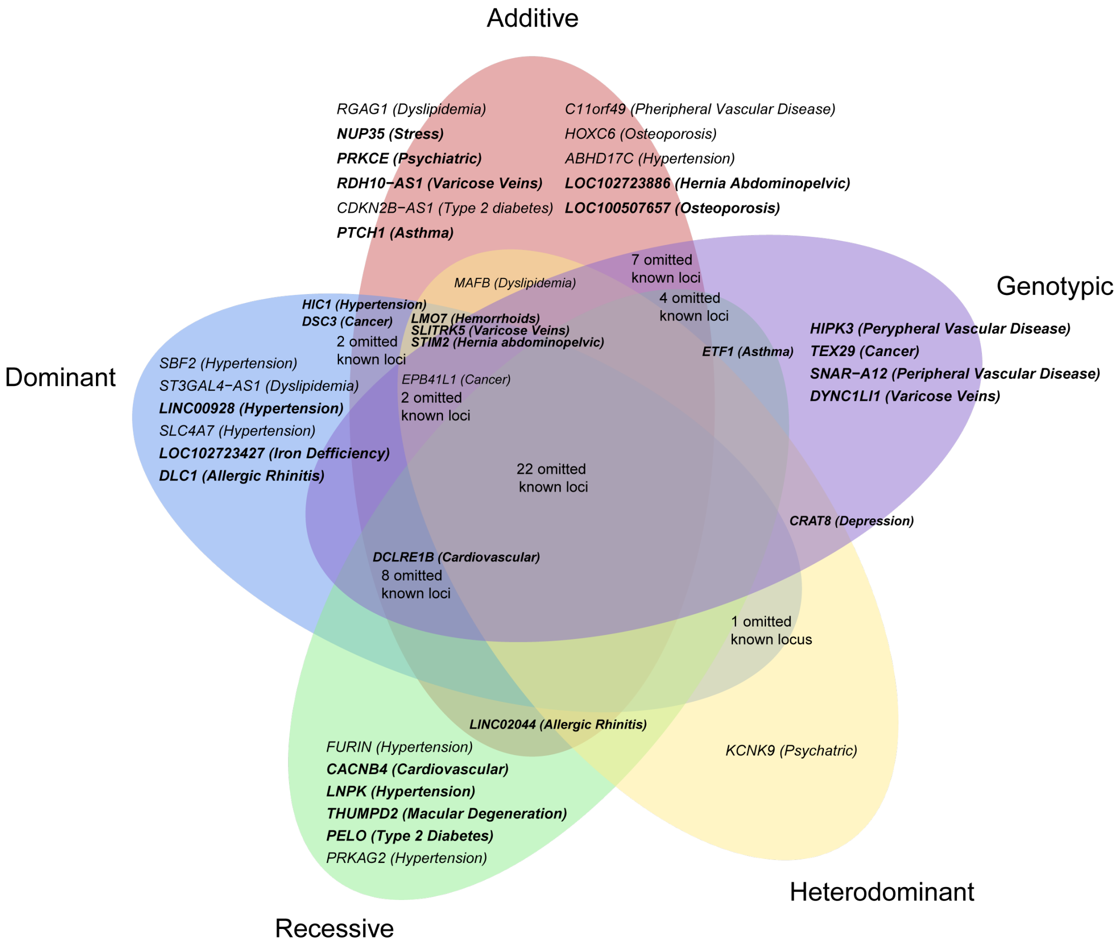


Figure 18. Venn Diagram showing the loci that could be identified after analyzing multiple inheritance models.

As seen in the Venn Diagram, the analysis of non-additive models was crucial for the identification of 13 novel (in bold) associated loci.

We identified 26 GWAS loci for 16 phenotypes that had not been reported before (Table 7). Among them, 15 would have not been identified by the four reference panels (Figure 17c). For example, the *CACNB4* loci was better imputed, and only genome-wide significant, by GoNL. Among the 26 novel loci, 11 (55%) out of 20 had a significant dominance deviation ($p < 0.05$) (Table 7).

In addition, three low-frequency and rare variants have large recessive effects (Table 7 and Supplementary Figure 23). We found an INDEL associated with cardiovascular disease in *CACNB4* (rs201654520, MAF = 0.017, OR [CI 95%] = 19.02 [5.50-65.84], $p = 4.32 \times 10^{-8}$), a gene previously associated with idiopathic dilated cardiomyopathy in African American (Xu et al., 2018), a variant near the *PELO* gene associated with

Table 7. New associations from the GERA cohort analysis

Phenotype (Cases/Controls)	CHR	Nearest Gene	Position	rsID	Alleles	MAF	Lowest P-value Model	Additive Model		Lowest P-value Model		Dominance Deviation
								OR (CI 95%)	P-value	OR (CI 95%)	P-value	P-value
Allergic Rhinitis (13,936/42,701)	3	*LINC02044*	112,911,615	rs2399472	C/T	0.073	Additive	1.17 (1.10-1.23)	1.55×10^{-8}	1.17 (1.10-1.23)	1.55×10^{-8}	6.66×10^{-1}
	8	*DLC1*	13,164,746	rs10112506	A/G	0.39	Dominant	0.94 (0.91-0.97)	8.61×10^{-6}	0.89 (0.86-0.93)	1.54×10^{-8}	2.86×10^{-4}
Asthma (9,209/47,428)	5	*ETF1*	137,858,067	rs154073	C/T	0.429	Recessive	1.09 (1.06-1.13)	6.06×10^{-8}	1.18 (1.12-1.25)	4.23×10^{-9}	9.28×10^{-3}
	9	*PTCH1*	98,344,866	rs67053006	C/G	0.139	Additive	0.87 (0.83-0.91)	4.14×10^{-8}	0.87 (0.83-0.91)	4.14×10^{-8}	8.10×10^{-1}
Cancer (17,131/39,506)	13	*TEX29*	112,115,591	rs138646839	C/T	0.005	Genotypic	1.68 (1.39-2.03)	1.45×10^{-7}	80.11 (0.00->100) / 0.02 (0.00->100)	3.54×10^{-8}	*NA*
	18	*DSC3*	28,442,343	rs2014497	A/G	0.008	Additive	1.50 (1.30-1.72)	2.44×10^{-8}	1.50 (1.30-1.72)	2.44×10^{-8}	6.00×10^{-1}
Cardiovascular (15,009/41,628)	1	*DCLRE1B*	114,448,752	rs10858023	C/T	0.35	Dominant	1.09 (1.06-1.12)	3.26×10^{-8}	1.14 (1.09-1.19)	2.11×10^{-9}	9.68×10^{-2}
	2	*CACNB4*	152,912,244	rs201654520	CT/C	0.017	Recessive	1.10 (0.98-1.22)	1.10×10^{-1}	19.02 (5.50-65.84)	4.32×10^{-8}	1.94×10^{-2}
Depression (7,264/49,373)	12	*CRAT8*	128,551,715	rs374090272	GT/G	0.281	Heterodominant	0.94 (0.90-0.98)	3.00×10^{-3}	1.18 (1.12-1.25)	3.15×10^{-9}	4.36×10^{-6}
Type 2 Diabetes (6,967/49,670)	5	*PELO*	52,080,909	rs77704739	T/C	0.036	Recessive	1.15 (1.05-1.26)	2.80×10^{-3}	4.32 (2.70-6.92)	1.75×10^{-8}	1.10×10^{-6}
Hemorrhoids (9,129/47,508)	13	*LMO7*	76,281,808	rs186102686	C/T	0.004	Heterodominant	1.98 (1.58-2.48)	2.18×10^{-8}	1.99 (1.59-2.49)	2.03×10^{-8}	*NA*
Hernia Abdominopelvic (6,291/50,346)	1	*LOC102723886*	219,762,581	rs2494196	C/A	0.274	Additive	1.13 (1.08-1.18)	2.03×10^{-8}	1.13 (1.08-1.18)	2.03×10^{-8}	6.87×10^{-1}
	4	*STIM2*	27,019,359	rs113180595	T/C	0.004	Heterodominant	2.17 (1.69-2.78)	1.59×10^{-8}	2.18 (1.70-2.8)	1.27×10^{-8}	*NA*
Hypertension (28,391/28,246)	2	*LNPK*	176,532,019	rs1446802	A/G	0.5	Recessive	1.07 (1.04-1.09)	1.66×10^{-6}	1.13 (1.08-1.17)	4.42×10^{-8}	6.85×10^{-3}
	15	*LINC00928*	90,081,905	rs28792763	G/A	0.462	Dominant	0.94 (0.91-0.96)	4.14×10^{-6}	0.88 (0.84-0.92)	4.42×10^{-8}	4.80×10^{-3}
	17	*HIC1*	1,959,826	rs112963849	C/A	0.082	Additive	1.15 (1.10-1.21)	1.71×10^{-8}	1.15 (1.10-1.21)	1.71×10^{-8}	8.01×10^{-1}
Iron Deficiency (2,439/54,198)	7	*LOC102723427*	67,292,424	rs79798837	C/T	0.118	Dominant	0.77 (0.70-0.85)	1.69×10^{-7}	0.74 (0.66-0.83)	3.80×10^{-8}	8.92×10^{-2}
Macular Degeneration (3,685/52,952)	2	*THUMPD2*	40,010,523	rs557998486	T/TG	0.009	Recessive	1.07 (0.81-1.41)	6.28×10^{-1}	10.5*	2.75×10^{-8}	*NA*
Osteoporosis (5,399/51,238)	22	*LOC100507657*	27,772,054	rs139959245	C/T	0.007	Additive	1.91 (1.53-2.37)	4.79×10^{-8}	1.91 (1.53-2.37)	4.79×10^{-8}	*NA*
Psychiatric (8,624/48,013)	2	*PRKCE*	46,278,720	rs12712961	T/A	0.452	Additive	1.10 (1.06-1.14)	1.66×10^{-8}	1.10 (1.06-1.14)	1.66×10^{-8}	6.32×10^{-6}
Peripheral Vascular Disease (4,301/52,336)	11	*HIPK3*	33,391,655	rs80274406	A/G	0.091	Genotypic	1.06 (0.98-1.15)	1.76×10^{-1}	0.51 (0.36-0.73) / 2.26 (1.58-3.24)	4.26×10^{-8}	1.35×10^{-8}
	19	*SNAR-A12*	48,403,215	rs2932761	A/G	0.289	Genotypic	0.97 (0.93-1.02)	3.04×10^{-1}	0.87 (0.81-0.93) / 1.27 (1.17-1.38)	3.55×10^{-8}	2.57×10^{-1}
Stress (4,314/52,323)	2	*NUP35*	184,407,101	rs577242570	T/G	0.004	Additive	2.33 (1.77-3.08)	4.56×10^{-8}	2.33 (1.77-3.08)	4.56×10^{-8}	*NA*
Varicose Veins (2,483/54,154)	3	*DYNC1LI1*	32,652,184	rs62250779	G/A	0.073	Genotypic	1.17 (1.05-1.3)	5.60×10^{-3}	0.36 (0.17-0.78) / 3.61 (1.65-7.89)	2.13×10^{-8}	1.92×10^{-7}
	8	*RDH10-AS1*	74,284,818	rs2383896	A/G	0.479	Additive	1.17 (1.11-1.24)	5.00×10^{-8}	1.17 (1.11-1.24)	5.00×10^{-8}	9.58×10^{-4}
	13	*SLITRK5*	88,346,617	rs117798068	T/C	0.011	Heterodominant	2.03 (1.63-2.53)	1.59×10^{-8}	2.07 (1.66-2.59)	8.41×10^{-9}	9.88×10^{-1}

CHR = Chromosome, Position = Position Hg19, Alleles = Non-effect Allele / Effect Allele, MAF=Minor Allele Frequency, OR= Odds Rstio, CI= Confidence Interval

* Obtained through RAFT

type 2 diabetes with the greatest odds ratio for type 2 diabetes in Europeans reported to date (rs77704739, MAF=0.036, OR [CI 95%] = 4.32 [2.70-6.92], p = 1.75 × 10^{-8}), and a rare INDEL associated with age-related macular degeneration near *THUMPD2* (rs557998486, MAF= 0.009, OR = 10.5, p = 2.75 × 10^{-8}).

2.3 Replication with UK Biobank and supporting evidence

We sought replication of previously unreported loci in UK Biobank, a prospective cohort of ~500,000 individuals aged between 40 to 69 years when recruited in 2006-2010, and genotyped with a high-density array (Bycroft et al., 2018).

We noted that the available phenotypes for the GERA cohort are collapsed by groups of diseases. For example, the condition *cancer* includes a variety of cancer conditions. Therefore, an association can be driven by a specific type of cancer, which we may be able to disentangle by analyzing various related phenotypes in UK Biobank. Likewise, some of the conditions may not be ascertained or have later age at onset than the average age at ascertainment in UK Biobank (56,52 years) (Hewitt et al., 2016), which could affect the replication success.

Despite these differeces between UK Bioank and GERA, among these 26 new associated loci, we replicated 4 of the novel associations for which we had an equivalent phenotype in UK Biobank (Table 8). Among them, 2 recessive associations in GERA cohort replicated in UK Biobank, including the loci associated with type 2 diabetes near *PELO* (rs77704739, meta-analysis OR [CI 95%] = 2.46 [1.88 - 3.21], meta-analysis p = 4.68 × 10^{-11}), and the INDEL associated with age- related macular degeneration near *THUMPD2* (rs557998486, mega-analysis OR [CI 95%] = 26.51 [7.57-92.85], meta-analysis p = 3.29 × 10^{-8}). This variant could not be tagged by the HRC reference panel, which does not have INDELs, since the most correlated SNP was in very weak linkage disequilibrium with the lead INDEL (rs724682, LD r^2 = 0.37).

Two autosomal variants were also found associated using the additive model and further replicated with UK Biobank, including a variant associated with hernia abdominopelvic (rs2494196, MAF=0.28, meta-analysis with ICD10 code K42 [Umbilical hernia] OR [CI 95%] = 1.19 [1.15 - 1.22], meta-analysis p = 2.94 × 10^{-22})

Table 8. Replication results with UK Biobank

				Stage 1. Discovery					Stage 2. Replication				Stage 1 + Stage 2. Meta-analysis			
				Additive		Best Model			Additive		Lowest p-value model		Additive		Lowest p- value model	
CHR	rsID (Alleles) (MAF)	Best Model	Phenotype (Cases/Controls)	OR (CI 95%)	P-value	OR (CI 95%)	P-value	Field (Cases/Controls or Sample Size)	OR (CI 95%)	P-value	OR (CI 95%)	P-value	OR (CI 95%)	P-value	OR (CI 95%)	P-value
18	rs2014497 (A/G) (0.008)	Additive	Cancer (17,131/39,506)	1.50 (1.30-1.72)	2.44×10^{-8}	1.50 (1.30-1.72)	2.44×10^{-8}	Self-reported: chronic lymphocytic (237/360,904)	2.13 (1.14-3.97)	3.50×10^{-2}	2.13 (1.14-3.97)	3.50×10^{-2}	1.52 (1.33-1.74)	1.60×10^{-9}	1.52 (1.33-1.74)	1.60×10^{-9}
								Self-reported: kidney/renal cell cancer (473/360,668)	1.75 (1.07-2.86)	4.25×10^{-2}	1.75 (1.07-2.86)	4.25×10^{-2}	1.51 (1.32-1.73)	1.49×10^{-9}	1.51 (1.32-1.73)	1.49×10^{-9}
								C69 Malignant neoplasm of eye and adnexa (146/361,048)	2.51 (1.19-5.3)	3.56×10^{-2}	2.51 (1.19-5.3)	3.56×10^{-2}	1.52 (1.33-1.75)	1.95×10^{-9}	1.52 (1.33-1.75)	1.95×10^{-9}
1	rs2494196 (C/A) (0.274)	Additive	Hernia Abdominopelvic (6,291/50,346)	1.13 (1.08-1.18)	2.03×10^{-8}	1.13 (1.08-1.18)	2.03×10^{-8}	Self-reported: umbilical hernia (328/360,813)	1.42 (1.21-1.67)	2.31×10^{-5}	1.42 (1.21-1.67)	2.31×10^{-5}	1.15 (1.10-1.19)	5.35×10^{-11}	1.15 (1.10-1.19)	5.35×10^{-11}
								K40 Inguinal hernia (13,365/347,829)	1.09 (1.06-1.12)	3.95×10^{-10}	1.09 (1.06-1.12)	3.95×10^{-10}	1.10 (1.08-1.12)	7.78×10^{-17}	1.10 (1.08-1.12)	7.78×10^{-17}
								K41 Femoral hernia (475/360,719)	1.44 (1.26-1.64)	1.24×10^{-7}	1.44 (1.26-1.64)	1.24×10^{-7}	1.16 (1.11-1.21)	2.26×10^{-12}	1.16 (1.11-1.21)	2.26×10^{-12}
								K42 Umbilical hernia (2,623/358,571)	1.29 (1.22-1.37)	1.14×10^{-17}	1.29 (1.22-1.37)	1.14×10^{-17}	1.19 (1.15-1.22)	2.94×10^{-22}	1.19 (1.15-1.22)	2.94×10^{-22}
								K43 Ventral hernia (2,470/358,724)	1.18 (1.11-1.25)	1.77×10^{-7}	1.18 (1.11-1.25)	1.77×10^{-7}	1.15 (1.11-1.19)	1.99×10^{-14}	1.15 (1.11-1.19)	1.99×10^{-14}
2	rs557998486 (T/TG) (0.009)	Recessive	Macular Degeneration (3,685/52,952)	1.07 (0.81-1.41)	6.28×10^{-1}	10.5*	2.75×10^{-8}	Eye problems/disorders: Macular degeneration	0.98 (0.72-1.32)	8.81×10^{-1}	7.58 (1.54-37.32)	4.1×10^{-2}	1.01(0.82-1.24)**	7.91×10^{-1}***	26.51(7.57-92.85)**	3.29×10^{-8}***
5	rs77704739 (T/C) (0.036)	Recessive	Type 2 Diabetes (6,967/49,670)	1.15 (1.05-1.26)	2.80×10^{-3}	4.32 (2.70-6.92)	1.75×10^{-8}	Self-reported: diabetes (14,114/347,027)	1.03 (0.97-1.09)	3.87×10^{-1}	1.88 (1.35-2.6)	4.95×10^{-4}	1.06 (1.01-1.12)	1.78×10^{-2}	2.46 (1.88-3.21)	4.68×10^{-11}

ΩCHR = Chromosome, Position = Position Hg19, Alleles = Non-effect Allele / Effect Allele, MAF= Minor Allele Frequency, OR= Odds Ratio

* Obtained through RAFT

** Obtained through a mega-analysis with UK Biobank using the “expected” method from SNPTEST

*** Method sample size

and a rare variant associated with cancer near *DSC3* (rs2014497, MAF=0.008, meta-analysis with self-reported kidney/renal cell cancer OR [CI 95%] = 1.51 [1.32 - 1.73], meta-analysis p = 1.49 × 10^{-9}).

For the majority of novel loci, including those that did not replicate with an equivalent phenotype in UK Biobank, we found additional evidences from related conditions, treatments or biomarkers in UK Biobank (Supplementary Table 2). As examples, the rs10858023 variant associated with cardiovascular disease with a dominant effect was associated with insulin treatment (OR [CI 95%]= 1.15 [1.05-1.25], p= 1.88 × 10^{-3}) and with hypothyroidism/myxoedema (OR [CI 95%]= 1.19 [1.16-1.23], p= 9.64 × 10^{-28}) as well as levothyroxine sodium treatment (beta [CI 95%]=1.18 [1.14-1.22], p= 2.67 × 10^{-20}). Both type 2 diabetes and hypothyroidism are well-established risk factors for cardiovascular disease (Biondi and Klein, 2004; Martin-Timon et al., 2014). The recessive variant associated with hypertension, rs1446802 (MAF=0.5, OR [CI 95%]= 1.13 [1.08-1.17], p= 4.42 × 10^{-8}), was also associated with kidney failure (MAF=0.5, OR [CI 95%]= 1.50 [1.09-2.06], p= 0.015, cases= 171), as hypertension is one of the leading causes of chronic kidney disease (Jha et al., 2013). A rare additive variant, rs139959245, associated with osteoporosis (MAF= 0.007, OR [CI 95%]= 1.91 [1.53-2.37], p= 4.79 × 10^{-8}) was also associated with spine fracture (OR [CI 95%]= 1.72 [1.09-2.71], p= 0.034, cases= 812), and vertebrae fractures have been long regarded as osteoporotic (Cummings and Melton, 2002). In addition, this variant was also associated with arthrotec tablet (OR [CI 95%]= 2.06 [1.23-3.45], p=0.015, cases= 518), a treatment for osteoarthritis and rheumatoid arthritis, and calcichew forte treatment (OR [CI 95%]= 2.37 [1.34-4.21], p= 9.51 × 10^{-3}, cases = 357), and adjunct to specific osteoporosis treatment of patients with calcium deficiency. The variant rs77704739, near *PELO*, which we replicated with type 2 diabetes in UK Biobank, was also associated with metformin, a well-known treatment for type 2 diabetes (Knowler et al., 2002), and only for the recessive model in UK Biobank as well (OR [95% CI] = 2.34 [1.63 - 3.36], p = 3.77 × 10^{-5}). Further evidence are also found for the recessive INDEL associated with age-related macular degeneration and replicated in UK Biobank, as it was also associated with eye surgery only for the recessive model (beta [CI 95%] = 1.60 [1.83-13.42], p = 1.17 × 10^{-3}).

We also sought additional evidence among the data for biomarkers available at UK Biobank (Supplementary Table 2). Remarkable associations include the recessive variant rs154073 associated with asthma in the GERA cohort (OR [CI 95%] = 1.18 [1.12-1.25], p = 4.23 × 10^{-9}) that was also associated with testosterone for the recessive model in UK Biobank (beta [CI 95%] = -0.01 [-0.01- -0.01], p = 4.42 × 10^{-7}). Testosterone is known to attenuate group 2 innate lymphoid cells (ILC2) that are increased in asthmatic patients thus contributing to the sexual dimorphism observed for asthma (Cephus et al., 2017). Finally, the novel recessive rare INDEL associated with age-related macular degeneration, rs557998486, was found associated with C-reactive protein only when testing the recessive model (beta [CI 95%] =1.11 [0.70-1.53], p = 1.15 × 10^{-4}). C-reactive protein is a known biomarker for macular degeneration (Molins et al., 2018).

2.4 New variants effects in gene expression and their functional characterization

We tested the effect of all novel variants on expression to identify their possible effector genes by interrogating eQTLGen Consortium (Võsa et al., 2018) and GTEx (GTEx Consortium, 2013; GTEx Consortium et al., 2017) data. Through this analysis, 18 out of 26 variants were found to be associated with the expression of a gene (Supplementary Table 3). Of note, 3 out of 4 replicated loci in UK Biobank are among these associations, including the two recessive variants for *PELO* and *THUMPD2*.

In particular, for the rs77704739 SNP, associated with type 2 diabetes, we found an association with the expression of *PELO* in blood in the eQTLGen Consortium and across multiple tissues in GTEx, including pancreas (p = 1.00 × 10^{-6}) (Supplementary Table 3).

Through this exploratory analysis, we also found a variant in moderate linkage disequilibrium (LD) with the rare recessive INDEL rs557998486 associated with age-related macular degeneration (rs116649730, LD r^2= 0.32) associated with reduced expression of its nearest gene, *THUMPD2* (Z-score = -4.85, p = 1.25 × 10^{-6}), according to eQTLGen Consortium data.

We further interrogated the *PELO* and *THUMPD2* loci analyzing available epigenomic data sets (Roadmap Epigenomics et al., 2015), since they displayed an exclusive association under the recessive model with large effects. Through this exploratory analysis, we found that the variants in the credible set of *PELO* locus associated with type 2 diabetes are in a regulatory element in pancreatic islets, including active enhancers and promoters bounded by pancreatic islet specific transcription factors (Figure 19). For *THUMPD2*, we found DNAse I signals in retinal and iris tissues, suggesting an open chromatin state in the *THUMPD2* variant (Figure 20). However, there was not available data for H3K27Ac marks in eye tissues, neither in ENCODE data (Encode Project Consortium, 2012) nor in the Epigenome Roadmap, to confirm the possible regulatory effect of the rs557998486 variant in *THUMPD2*.

2.5 Cross-phenotype analysis

After analyzing all the phenotypes available at GERA cohort independently, GWAS significant variants were selected for a cross-phenotype association analysis in GUIDANCE based on the most significant model of inheritance according to the association results. In this analysis, 8 genome-wide associated loci were found significant ($p \leq 2.58 \times 10^{-5}$) for at least one additional disease (Supplementary Table 4).

Of note, *HLA* regions were associated with many complex diseases, including asthma, type 2 diabetes, dyslipidemia, cancer and age-related macular degeneration. Additionally to the known comorbidity at *TSLP/WDR36* locus (rs252716) that was associated with an increased risk of asthma (OR [CI 95%]= 1.10 [1.07-1.14], p-value= 3.53×10^{-9}) and allergic rhinitis (OR [CI 95%] = 1.07 [1.04-1.1], p-value = 2.79×10^{-6}) (Ferreira et al., 2014), the new rs154073 variant at *ETF1* locus associated with an increased risk of asthma (OR [CI 95%] = 1.18 [1.12-1.25], p-value = 4.23×10^{-9}) was also associated with an increased risk of irritable bowel following the recessive model (OR [CI 95%] = 1.22 [1.12-1.34], p-value = 1.30×10^{-5}).

Moreover, 5 up to 8 loci, including *HLA*, associated dyslipidemia with cancer, asthma, age-related macular degeneration, hypertension, peripheral vascular disease, osteoarthritis, type 2 diabetes and cardiovascular disease.

The *TRIB1* locus was associated with a protective effect for both dyslipidemia (rs2954038, OR [CI 95%] = 0.88 [0.86-0.91], p-value = 2.53 × 10^{-19}) and hypertension (rs2954038, OR [CI 95%]= 0.94 [0.91-0.97], p-value = 2.40 × 10^{-5}), and *ABO* locus at chromosome 9 is associated with an increased risk of dyslipidemia (rs532436, OR [CI 95%] = 1.16 [1.13-1.2], p-value = 8.10 × 10^{-21}), peripheral vascular disease (rs8176685, OR [CI 95%] = 1.20 [1.13-1.27], p-value= 4.91 × 10^{-10}) and osteoarthritis (rs9650778, OR [CI 95%] = 1.11 [1.06-1.16], p-value = 2.34 × 10^{-5}).

Furthermore, the *NECTIN2*/*PVRL2* locus was associated with an increased risk of dyslipidemia (rs66626994, OR [CI 95%] = 1.26 [1.21-1.31], p-value = 4.55 × 10^{-31}) and cardiovascular disease (rs66626994, OR [CI 95%] = 1.11 [1.06-1.16], p-value = 1.44 × 10^{-5}), but with a protective effect in type 2 diabetes (rs34342646, OR [CI 95%] = 0.89 [0.84-0.94], p-value = 1.03 × 10^{-5}).

In addition, the rs4722756 variant at *JAZF1* locus was associated both type 2 diabetes (p-value = 4.36 × 10^{-8}) and asthma (p-value= 6.87 × 10^{-6}) for the genotypic model.

Finally, the rs3764261 variant at *CETP* locus at chromosome 16 was associated with a protective effect for dyslipidemia (OR [CI 95%] = 0.91 [0.88-0.93], p-value = 2.85 × 10^{-12}) but an increased risk of age-related macular degeneration (OR [CI 95%] = 1.12 [1.07-1.19], p-value = 1.94 × 10^{-5}).

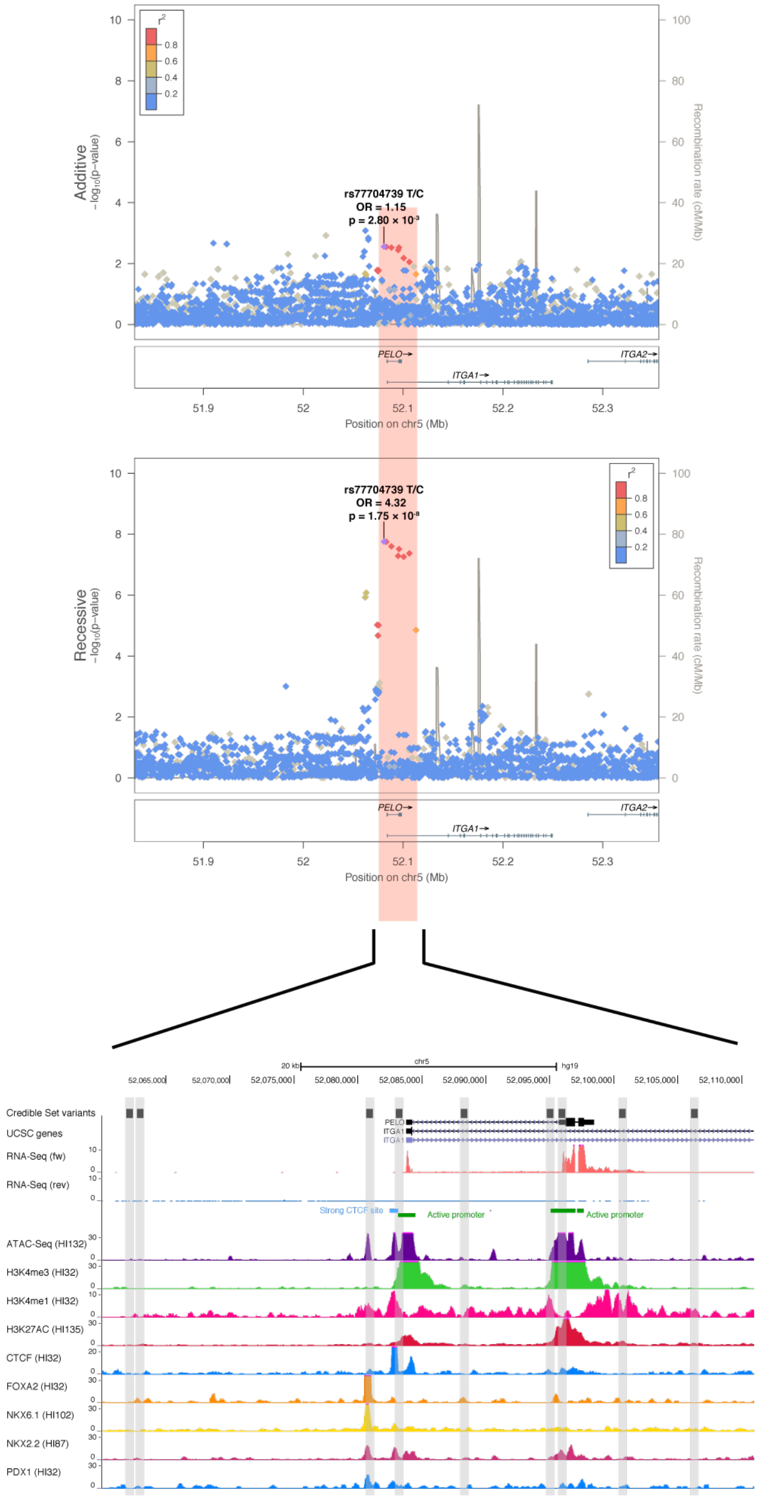


Figure 19. Functional characterization of rs77704739.

Signal plot for chromosome 5 region surrounding rs77704739. Each point represents a variant, with its p-value from the discovery stage on a −log10 scale in the y-axis. The x-axis represents the genomic position (hg19). For the 5 SNPs in the credible set, the tracks show open chromatin sites in pancreatic islets, two of them classified as active promoters and one bounded by pancreatic islet specific transcription factors, such as PDX, NKX and FOXA2.

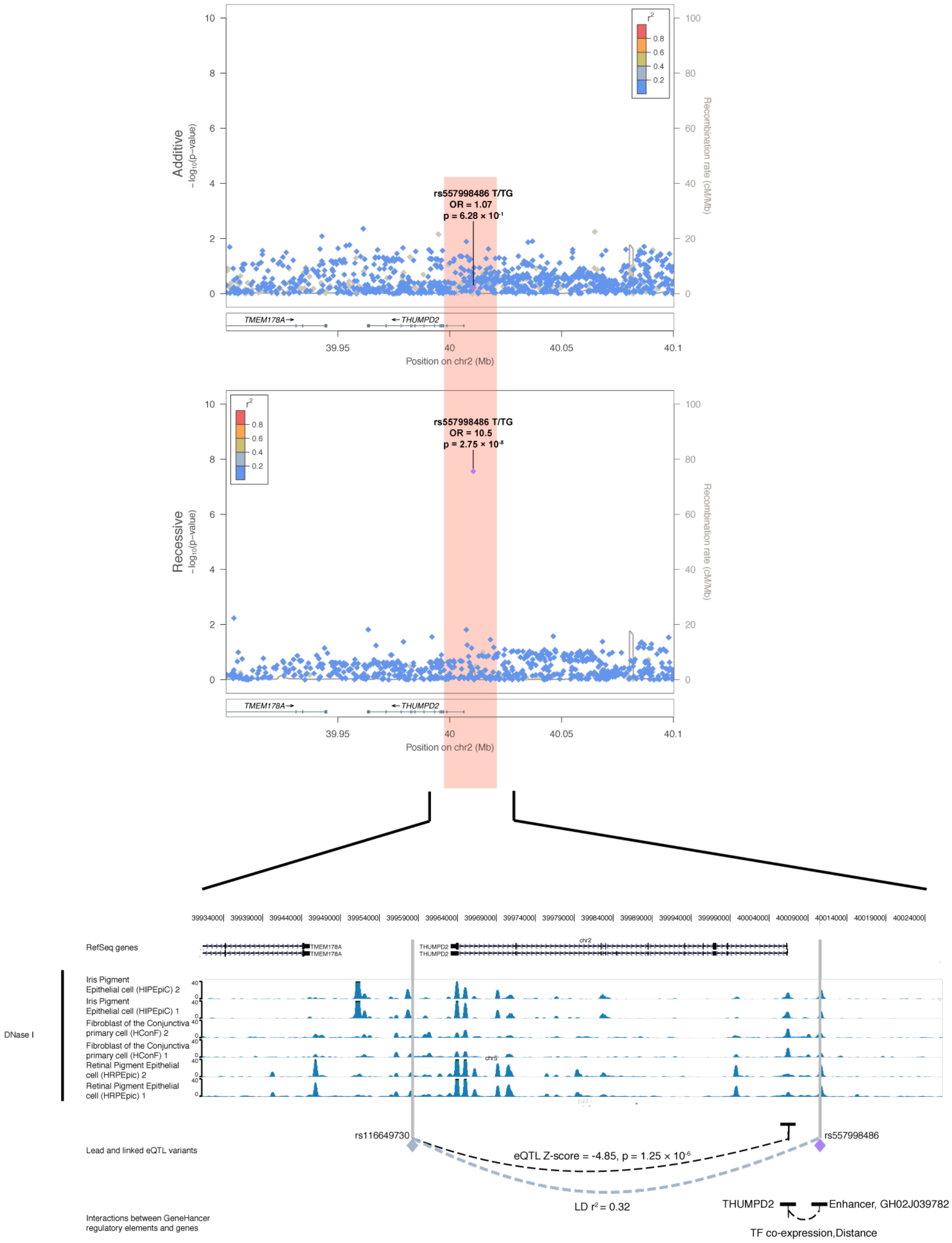


Figure 20. Functional characterization rs557998486.

Signal plot for chromosome 2 region surrounding rs557998486. Each point represents a variant, with its p-value from the discovery stage on a −log10 scale in the y-axis. The x-axis represents the genomic position (hg19). DNAse I signals in eye tissues suggest an open chromatin site in the rs557998486 locus.

Discussion

After almost 15 years of GWAS, thousands of loci have been identified for a vast number of complex diseases, including type 2 diabetes, coronary artery disease, schizophrenia, major depressive disorder and subtypes of cancer, among others (Tam et al., 2019). However, initial studies did not have the necessary statistical power to find common associated variants with small effects, which are responsible for a large fraction of the heritability (i.e., the variance explained by genetic factors) (Yang et al., 2010). This fact pointed at the sample size as the first limiting factor for finding regions associated with complex diseases. Another explanation that limits the statistical power of GWAS is the incomplete LD that can exist between the causal variant and genotyped variants, especially for variants with low MAFs (Yang et al., 2010). Therefore, low frequency and rare variants, and their possible associations, remain elusive. Hence, the lack of enough sample sizes and the inability to detect loci-trait associations lead by low-frequency and rare variants or common variants with small effects, contribute to the missing heritability (Wainschtein et al., 2019).

With the need for increasing sample sizes, next-generation sequencing is still prohibitive as a common strategy, mostly due to its elevated cost. Although moving from genotyping data to next-generation sequencing is perceived as the natural evolution of GWAS, still the cost of sequencing far exceeds the cost of genotyping. Some examples of initiatives that have sequenced thousands of individuals are of an extensive utility for the community, such as UK10K (UK10K Consortium et al., 2015) or TopMED (Taliun et al., 2019). However, these initiatives represent titanic efforts and cannot constitute the default procedure yet to study the genetics behind complex diseases. Hence, even though sequencing would resolve the limitation of including low-frequency and rare variants into the analysis, the affordable small sample sizes would be unpowered to detect new associations with small effects.

For that reason, GWAS are still mainly based on genotyping array data, and GWAS have remarkably enlarged the number of known associated loci by 1) increasing the sample sizes and 2) including genotype imputation to increment the number of variants to analyze. However, as some authors argued (Tam et al., 2019), GWAS findings to date only represent the tip of the iceberg, and many opportunities to improve and increment novel findings are still ahead. Beyond increasing the sample size, expanding the range of phenotypes, the populations studied and including different analyses may lead to additional findings and may build a better picture of the genetic landscape behind complex diseases (Tam et al., 2019).

In brief, this thesis sought to widen and deepen the analysis of the genetics behind complex diseases by improving current GWAS strategies.

In the next pages, the results from this thesis that we devoted to boost GWAS are discussed. To improve GWAS, we implemented genotype imputation using multiple reference panels to cover more variants, including variants with a MAF as low as 0.1% and INDELs. In addition, we significantly increased the number of discoveries from GWAS by including the X chromosome and all the possible inheritance models in the association test. To bring a complete GWAS tool to a broader community, we integrate this methodology into an easy-to-use GWAS workflow called GUIDANCE.

Finally, during the development of GUIDANCE, we have applied its methodology to different cohorts contributing to multiple studies (Appendix 1-5), with particular focus on the GERA dataset, a large cohort for age-related diseases. After the analysis of GERA cohort using GUIDANCE, the identification of novel genetic variants that modify risk of complex diseases, including rare variants, INDELs and non-additive associations, we demonstrated the importance of performing a comprehensive analysis of the data to better understand the genetic architecture behind complex diseases, as well as a way to take full advantage of the existing and newly generated GWAS data.

1 Comparison of GUIDANCE with previous efforts to collect and organize GWAS steps into single applications

Current workflows to perform GWAS are complex and time-consuming. The large sample sizes partly explain this. However, GWAS steps have to manage tasks dependencies and different configuration parameters, computing requirements, levels of parallelism, disk and memory usage as well as different file formats.

Some efforts have been done to collect and organize some GWAS steps into single applications, such as GRIMP (Estrada et al., 2009), a web-based interface and application to run the association testing, or GWASpi (Muniz-Fernandez et al., 2011), a tool that integrates a QC pipeline and the association testing. Of note, the authors of these applications already expressed the need for these automatized pipelines and the complexity of manually handling a GWAS, even without taking into account genotype imputation. Integrated today by default in every single GWAS pipeline, genotype imputation takes GWAS complexity and its computational requirements one step further.

During the development of GUIDANCE, online imputation servers (Michigan Imputation Server, https://imputationserver.sph.umich.edu/index.html, and Sanger Imputation Server, https://imputation.sanger.ac.uk/) that host the largest reference panels to date, both the HRC and TOPMed, were released. These servers allow accurate and efficient imputation with these sequence-based reference panels. Nevertheless, they still do not constitute a solution for many current analysis scenarios, as users are required to upload their individual-level genotyped data, cannot use and integrate their in-house reference panels, and cannot perform medium or large-scale analysis due to intrinsic limitations of the resources. In addition, even when using these servers is possible, there are still several steps needed beyond the imputation, such as post-imputation quality filtering and association testing, which are not yet covered by these servers, as it would require uploading sensitive phenotype data.

Recently, during this thesis, the Pan-African bioinformatics network, H3ABioNet, has developed a workflow for association and imputation on top of NextFlow (Baichoo et al., 2018), which has direct support for Amazon EC2 and Google Cloud. H3ABioNet pipelines include Eagle 2 and MINIMAC4, as well as a linear mixed-model association testing with GEMMA (Zhou and Stephens, 2012), BOLT-LMM (Loh et al., 2015) or FaST-LMM (Lippert et al., 2011).

In addition, RICOPILI, which stands for “Rapid Imputation for COnsortias PIpeLIne”, has been developed (Lam et al., 2019). RICOPILI includes QC and genotype imputation through Michigan or Sanger imputation servers as well as a post-imputation module, including association using PLINK, graphical representation of the results, meta-analysis using METAL, and conditional analysis among others.

Moreover, Hail (Hail Team, 2016), a Python-based genomic data analysis tool that can be executed in Google Cloud, has been developed. This multi-purpose Python library, implemented in Scala, Spark (Zaharia et al., 2012), and C++, includes QC and GWAS among its features, and have been widely adopted in academia. However, it still not constitute an easy-to-use integrated solution for a complete GWAS, and some studies using Hail have also used additional tools for GWAS, such as SNPTEST, R, PLINK (Roselli et al., 2018) or BOLT-LMM (Kerminen et al., 2019).

Compared to existing pipelines, we wanted an integrated pipeline that executes in a single run all the main steps in any current GWAS workflow, and not running separated scripts that can be integrated, even though GUIDANCE can also run in modules if needed. For that reason, we did not include QC as part of the workflow since we consider that manually checking the QC results before any analysis is essential to avoid dragging errors throughout the pipeline, thus reducing the possibility of spurious associations. Besides, separating QC from the actual pipeline is important because pre-phasing and genotype imputation are highly computational demanding, and checking the data before executing those steps can prevent unnecessary waste of resources.

Due to its nature, the X chromosome requires additional handling, and autosomal pipelines are not accurate methods to analyze it. Some pipelines, such as XWAS (Gao et al., 2015), have been developed for analyzing the X chromosome in GWAS. However, conceiving the analysis of the X chromosome and the analysis of autosomes as pipelines that run separately had not helped to integrate the X chromosome analysis in GWAS by default (Wise et al., 2013).

H3ABioNet, RICOPILI and Hail pipelines constitute an enormous effort to make current GWAS steps more feasible. We include GUIDANCE as an additional pipeline for GWAS that covers in a single run all the steps included in any current GWAS workflow without user intervention, from QCed files to graphical representation (i.e., haplotypes phasing, genotype imputation, post-imputation filters, association testing, post-association filtering, and summary reports and graphical representation). Besides, GUIDANCE offers several methodological advantages over the other pipelines. Those advantages, which will be fully discussed in further sections, include the imputation of genotypes using multiple reference panels, the integration of the X chromosome's analysis, and the association testing with non-additive inheritance models.

2 Developing GUIDANCE

2.1 The programming framework behind GUIDANCE

GUIDANCE has been developed on top of COMPSs (Lordan et al., 2014), a programming framework that, similarly to NextFlow, aims to make the development of parallel workflows in computing platforms such as clusters or clouds, easier (see Methods, section 1.1, for a detailed description of COMPSs). Other options of different Big Data frameworks are available, such as Hadoop (D. Cutting, 2006), used by the Michigan Imputation Server, or Spark (Zaharia et al., 2012). However, these frameworks require their own defined operators and specific data types, while GUIDANCE is based on the use of external binaries and the exchange of files, which requires the flexibility of a more generic programming model, like the one offered by COMPSs. By combining and integrating state-of-the-art GWAS analysis tools employing COMPSs, GUIDANCE frees users from the responsibility of dealing with the computational complexity of the whole process.

However, all the advantages and plasticity offered by COMPSs have a counterpoint. As COMPSs is still in constant development, the lack of a stable version of COMPSs might limits GUIDANCE utility. Nevertheless, working together has offered us the possibility of improving both COMPSs itself and, therefore, also the capabilities of GUIDANCE, making it more efficient and portable. We are already working on running GUIDANCE on Google Cloud, and we were able to run GUIDANCE outside the MareNostrum III, with LSF queue system, and Marenostrum IV, with Slurm queue system. Specifically, we run GUIDANCE using SuperMUC, a supercomputer from the Leibniz Supercomputing Centre with a Loadleveler queue system, thus demonstrating the possibilities of GUIDANCE's portability.

2.2 State-of-the-art GWAS tools integrated in GUIDANCE

GUIDANCE includes SHAPEIT2 and Eagle2, and IMPUTE2 and MINIMAC4 for pre-phasing and genotype imputation, respectively, to suit users' demands.

However, during the development of GUIDANCE, SHAPEIT3 an IMPUTE4 have been developed and applied to large cohorts. Concretely, IMPUTE4 uses the same method behind IMPUTE2; thus the results from both methods are identical but reducing memory usage and increasing speed using compact data structures (Bycroft et al., 2018). Hence, the integration of IMPUTE4 in GUIDANCE will improve its performance without compromising accuracy. Recently, IMPUTE5 was also released (Rubinacci et al., 2019).

For SHAPEIT3, the developers recommend to use it when the sample size is larger than 20,000 since SHAPEIT2 is more accurate at least for that sample size (the authors did not run SHAPEIT2 in larger sample sizes due to computational limitations) (O'Connell et al., 2016).

Lately, SHAPEIT4 has also been released (Delaneau et al., 2018). Applied to UK Biobank, it demonstrated to be faster than SHAPEIT3, Eagle2 and Beagle5 (Browning et al., 2018), with an improved memory usage compared to SHAPEIT3 and similar performance compared to Eagle2 and Beagle5 (Delaneau et al., 2018). In terms of accuracy, all methods showed low error rates that decrease as the sample size increase (Delaneau et al., 2018). No accuracy comparison between SHAPEIT2 and SHAPEIT4 has been made. Nevertheless, even if the accuracy were better when using SHAPEIT2, phasing increasingly large sample sizes would not be feasible with SHAPEIT2 at some point. Hence, the integration of SHAPEIT4 in GUIDANCE would be necessary to deal with the continuously increasing sample sizes.

To date, GUIDANCE uses SNPTEST for the association test. A limitation of SNPTEST is that it cannot handle related individuals, thus limiting the analysis to case-control or unrelated population samples. For case-control studies with unrelated individuals, population stratification can be mitigated with an adequate QC before GWAS and adding PCA as covariates for the associatin test. However, the new trend of collecting thousands of data from volunteers in Electronic Health Records and biobanks, make cryptic relatedness and population stratification harder to account for. To that end, methods based on linear mixed models (LMM), such as GEMMA (Zhou and Stephens, 2012), FaST-LMM (Lippert et al., 2011) and BOLT-LMM (Loh et al., 2015), have received an increasing attention since they account for relatedness and population structure. In a comparison of 1) excluding related samples in UK Biobank and performing linear regression, which implies losing around 30% of the sample size, or 2) using LMM with minimal sample removal, the authors demonstrated that the second option doubles the statistical power (Loh et al., 2018). However, these methods also have limitations since LMM can be applied to case-control studies but it has not been designed to test binary traits. For example, methods such as BOLT-LMM can produce a loss of power and inflate false positive rates in unbalanced case-control cohorts, especially for low frequency and rare variants (Loh et al., 2018). SAIGE, developed for binary traits with a logistic mixed model, copes with unbalanced case-control ratios for any MAF (Zhou et al., 2018). However, SAIGE comes too with its own limitations. SAIGE may experience a loss of power since it assumes an infinitesimal model (the effect sizes are normally distributed, i.e., all variants are causal with small effect sizes). Hence, it may not have power in non-infinitesimal genetic architectures (Zhou et al., 2018), and it has been estimated that complex diseases have, in fact, a limited number of causal loci (Stahl et al., 2012). Conversely, BOLT-LMM models non-infinitesimal genetic architectures (Loh et al., 2015), but does not produce accurate estimates of effect sizes (i.e., odds ratios) for binary traits.

There is no single solution for all possible scenarios, and the quantity of high-quality tools that are continually emerging makes it difficult to cover them all. Nevertheless, the actual version of GUIDANCE is limited to SNPTEST and the analysis of cases-controls cohorts, and does not have the most efficient tools that are currently available for the analysis of current and ongoing GWAS datasets such as UK Biobank. Adding new versions of the already included tools, such as SHAPEIT4 or IMPUTE5, adapting the pipeline to handle quantitative traits using SNPTEST or including LMM tools for the association, have to be considered as next development steps for GUIDANCE.

Finally, none of the methods described above except SNPTEST allow the association using different models of inheritance.

2.3 Unique features included in GUIDANCE

GUIDANCE integrates unique features that are not available when using other GWAS workflows describe above, which will be described in the following sections.

2.3.1 The benefits of genotype imputation using multiple reference panels

A common practice today is to impute genotypes using HRC, a set of large reference panels of nearly ~30,000 individuals predominantly from European populations. In contrast, GUIDANCE allows the combination of different reference panels.

The greatest benefit of HRC is its sample size, which allows imputing low-frequency and rare variants (MAFs as low as 0.001) with high accuracy (McCarthy et al., 2016). However, HRC does not include INDELs since INDELs have been inconsistently called through the studies that compose HRC. Hence, the inclusion of additional reference panels provides INDELs into the analysis, which are more likely to have a causal effect.

Besides, the sample size is not the only factor to take into account to gain genotype imputation accuracy. A similar ancestry for both the study samples and the reference panel increases haplotype matches, thus improving the imputation accuracy of rare variants. Hence, population-based panels outperformed 1000G and HRC, even though 1000G includes more variant and HRC includes more samples (Mitt et al., 2017).

Due to these facts, and to get full advantage of existing panels, we designed GUIDANCE to run genotype imputation with an unlimited number of reference panels simultaneously. GUIDANCE ultimately combines the results into a final set of variants, selecting for each variant the imputation result from the panel with the highest accuracy, without the need of merging the reference panels in advance. A different approach to combine reference panels is already included in IMPUTE2. However, this method is based on cross-imputing two reference panels to obtain a merged reference panel, and the evaluation of this method did not demonstrate its usefulness, giving similar results for the combined panel (i.e., UK10K + 1000G phase 1) in the presence of a large population-based panel (i.e., UK10K) (Huang et al., 2015).

In contrast, we demonstrated the usefulness of our method since imputing with different reference panels was determinant to identify 28.7% of all the GWAS significant associations in the GERA analysis. An additional study has also benefited from this feature during GUIDANCE development; the 70KforT2D project, a reanalysis and meta-analysis of ~70,000 type 2 diabetes cases and controls from European ancestry, gained up to 41% of INDELs after combining the imputation results from 1000G phase3 and UK10K (Bonas-Guarch et al., 2018).

The availability of computational resources is a limiting factor when deciding which and how many panels will be used for imputation. We have been able to perform an analysis on such a scale due to the privileged computational environment at the Barcelona Supercomputing Center (BSC). Imputing with four reference panels simultaneously, as we did, is highly computational demanding, and possibly more than imputing with a large cross-imputed reference panel. Nevertheless, a proper comparison of both approaches would be necessary to determine their advantages and disadvantages.

2.3.2 Including non-additive inheritance models in the association testing

Although most of the genetic variance is expected to be additive (Zhu et al., 2015), non-additive genetic associations might be missed in the conventional only-additive approach. Among the advantages of using SNPTEST to test for single-variant associations, there is the possibility of testing different models of inheritance. Using GUIDANCE, our analysis of GERA demonstrated that the analysis of non-additive models of inheritance was determinant to identify 21.3% of associated loci at a genome-wide significance level, and 50% of the novel variants.

The models of inheritance that can be analyzed using SNPTEST are additive, dominant, recessive, heterozygote and general, as they are defined by SNPTEST developers.

A technical limitation when including non-additive association tests using SNPTEST is the required computational time, as it increases with the number of models to be analyzed. An additional limitation is that users have to be aware that correction by multiple testing has to be adjusted properly when including multiple inheritance models into the analysis. This can be counterproductive, diminishing the power to find new associations. To adjust the code to users' preferences and needs, the user can specify which inheritance models want to analyze in the configuration file, selecting from one to all of them. This represents a unique feature in GUIDANCE, which accommodates downstream analysis to the selected inheritance models, thus facilitating the interpretation and management of the results.

2.3.3 Summary statistics, graphical presentation, and cross-phenotype analysis

After the association test, GUIDANCE generates easy-to-read tables from the summary statistics and its graphical representation, including Q-Q plot and Manhattan plots for each disease and inheritance model analyzed (see Supplementary Figures 1-22 to see the plots from the analysis of GERA). Therefore, the user can quickly focus on the results, detect possible errors in the data and consider future analyzes efficiently. Among the future improvements that could be of interest in GUIDANCE, it would be including the generation of regional plots for significant loci, since it is probably the next step that the user will perform, and it is a relatively easy step to integrate in GUIDANCE using tools such as LocusZoom.

GUIDANCE also includes a cross-phenotype analysis when more than one disease is available in the study to interrogate the association of multiple phenotypes for the same variant. The cross-phenotype analysis is performed only for the inheritance model with the lowest p-value. However, a variant can be found associated with a significant p-value following both additive and an additional non-additive model. In this case, if the non-additive p-value is lower than the additive p-value, the non-additive model is used for cross-association with the other phenotypes. Nevertheless, non-additivity is not defined by the model that shows the lowest p-value, but by the deviation from additivity (Wood et al., 2016; Zhu et al., 2015), a test that can be performed using PLINK. Including the dominance deviation test into GUIDANCE will further help in the interpretation of the results, properly differentiating truly non-additive variants from those that are additive but are found non-additive by a matter of statistical power.

A limitation of the cross-phenotype analysis is that it does not account for overlapping samples between diseases. Therefore, in cohorts including multiple diseases, such as GERA or large biobanks, where thousands of individuals might be classified as cases for multiple phenotypes (e.g., cardiovascular, hypertension and dyslipidemia), it may be difficult to distinguish the truly cross-phenotype associations from the spurious ones due to sample overlap.

2.3.4 Including the X chromosome analysis alongside the autosomes

As explained in previous sections, one of the main gaps in GWAS is the systematic omission of the X chromosome. The X chromosome analysis is promising because it has been understudied, is a relatively large chromosome, and many diseases show different prevalence between males and females. However, it requires additional handling compared to autosomes for a proper analysis. Even in the existence of efforts to collect the tools required for its analysis, such as XWAS (Gao et al., 2015), its exclusion is still systematic. This could be partially explained because, from a user point of view, it still represents a separate analysis from the autosomes that will require additional handling. Besides, the analysis of the autosomes had been enough for high impact publications. Nevertheless, the X chromosomes deserved more attention, and there is no reason to keep on its exclusion from GWAS.

In order to fill this gap, we integrated the X chromosome analysis in GUIDANCE. Thus, the user can effortlessly analyze the X chromosome. Hence, even though it has its own sequence of tasks accounting for its particularities, the analysis itself remains agnostic for the user, being as easy as to run any of the autosomes. That is, while for the user the analysis of the X chromosome does not differ from the analysis of the autosomes specified in the configuration file, GUIDANCE has an alternative and specific sequence of tasks for chromosome X along the code. Hence, the analysis of X chromosome occurs in parallel with the autosomes through its alternative route, from the starting point (i.e., PLINK files) until the end (i.e., summary statistics and graphical representation), without any additional handling from the researcher.

We hypothesize that the availability of public data containing the X chromosome that has not been analyzed, combined with GUIDANCE, which fully integrates its analysis, is a promising way to mine GWAS' existing data. In addition, the integration of the chromosome X analysis as it has been done in GUIDANCE will prevent this situation from happening again, and we expect that future GWAS will all include the analysis of the X chromosome by default.

In summary, for a large proportion of this thesis we focused our efforts on building a unique integrated framework that not only facilitates GWAS and its interpretation but also promotes a complete an comprehensive analysis of genotyped data.

3 The importance of analyzing the wide spectrum of allele frequencies and non-additive inheritance models; GERA results

3.1 Combining panels to cover low-frequency, and rare variants as well as INDELs

Using GUIDANCE to analyze the GERA cohort, the largest publicly available cohort for age-related diseases to date, and combining the results of 1000G phase 3, UK10K, GoNL and HRC, we were able to test for association 16,059,686 variants with high imputation accuracy (info score > 0.7).

Through this study, we showed that imputing with different reference panels was critical to identify 28.7% (27 out of 94 loci) of all the GWAS significant associations since different panels have a different power to properly impute certain variants.

As a known example, the well-known *IL33* locus associated with asthma (Bonnelykke et al., 2014; Ferreira et al., 2014; Ferreira et al., 2017; Pickrell et al., 2016) was found associated only when imputing with 1000G phase 3 as the reference panel. However, the full locus was excluded when imputing with the other reference panels.

Besides, we also demonstrated that genotype imputation using multiple reference panels allowed us to identify new associated regions led by low-frequency (0.05 > MAF > 0.01) and rare variants (0.01 > MAF > 0.001). In our analysis of the GERA cohort, including reference panel with large sample sizes or large number of variants (i.e., HRC and 1000G phase 3, respectively) as well as population-specific reference panels (i.e., UK10K and GoNL) allowed the identification of 15 out of 26 (57.7%) novel findings led by low-frequency and rare variants (Figure 17c).

In addition, three out of 26 novel associations were led by INDELs (Table 7). INDELs, the second most common type of variants after SNPs in the human genome (1000 Genomes Project Consortium et al., 2015), have been largely associated with human disease in other studies (Dai et al., 2019). This demonstrates that using only one reference panel that does not include them, i.e., the HRC, will hinder our understanding of the genetic architecture of complex diseases. For example, the novel recessive INDEL associated with age-related macular degeneration (rs557998486) could not have been well tagged by the HRC reference panel, as the SNP with the highest LD shows a modest linkage disequilibrium and significance way beyond the INDEL (rs724682, LD $r^2 = 0.3728$, p = 0.0057).

In addition to that, when looking for the causal variant in order to design future functional studies, using only one reference panel such as HRC will be misleading since the most likely causal variant according to the credible sets from HRC will ignore INDELS, which can result in useless follow-up functional experiments based on a non-functional variant.

In summary, the combination of multiple reference panels using our approach has demonstrated to cover a wider spectrum of variant frequencies and different forms of genetic polymorphisms, which best characterizes the genetic architecture of complex diseases, also increasing the association findings.

3.2 Obtaining a better profile of the genetic architecture of complex diseases by analyzing all inheritance models

To further improved our understanding of the genetics behind complex diseases, the inclusion of non-additive inheritance models has been proven to be crucial through our analysis of the GERA cohort.

After the association test of 16,059,686 variants for the 22 diseases in GERA, 13 out of 26 novel associations (50%) were only found when non-additive models were tested (Figure 18), which emphasizes the importance of including non-additive association tests to find new associations in GWAS. Besides, the fact that 13 top variants deviate from the additivity significantly (Table 7) suggests that analyzing the additive model alone, not only reduces the number of new associations but also limits what we know about the genetic architecture of complex diseases, which also obey non-additive models of inheritance.

As a key step to validate our findings, we sought for replication in UK Biobank. However, the differences between UK Biobank and the GERA cohort limited our possibilities of replication. For example, we could not find an equivalent phenotype in UK Biobank for most of the novel findings.

One of the reasons for this mismatch is that the GERA cohort has some general phenotypes encompassing multiple diseases, such as *cancer*, *cardiovascular* or *psychiatric*, which makes the interpretation of the results challenging.

Furthermore, the GERA cohort is based on age-related conditions, and participants have an average age of 63, while UK Biobank cohort is a prospective cohort with substantially healthier and younger participants, with an average age of 56.52 years (Fry et al., 2017; Hewitt et al., 2016). For instance, while 12.3% of participants in the GERA cohort have type 2 diabetes (6,967 from 56,637 samples), only 3.9% of UK Biobank participants do (14,114 from 361,141 samples). Besides, individuals considered as controls in UK Biobank may develop age-related conditions once they reach the age of onset. Even for diseases like asthma, which can occur at any age, it has also been suggested that the etiology of childhood and adult-onset differ, despite the existence of some shared genetic risk factors (Ferreira et al., 2019; Pividori et al., 2019). Therefore, the differences between the GERA cohort and UK Biobank make the use of UK Biobank for replication of binary traits limited, and highlight the importance and utility of aging cohorts like GERA.

Replication is necessary to confirm novel findings, but many reasons could explain the lack of replication when using UK Biobank. Exploring additional cohorts with similar demographic characteristics to GERA, or properly adjusting the replication analysis in UK Biobank, for example, filtering controls younger than the age at onset for each disease, as it has been successfully proven in other studies (Bonas-Guarch et al., 2018), or using a better definition of the phenotypes for UK Biobank, are promising further steps to confirm our novel associated loci.

Despite the utility of using non-additive models to discover new associations, the correction for multiple testing results in an experiment-wide significant p-value cutoff of $p < 9 \times 10^{-10}$ after the Bonferroni correction for 22 diseases and five inheritance models (2.5 effective tests). With this experiment-wide p-value cutoff, two out of four novel replicated loci are still significant. Applying this threshold, which may be too conservative since the 22 diseases are not fully independent, only the additive rare SNP rs2014497 (MAF = 0.008), associated with cancer, and the recessive INDEL rs557998486 (MAF = 0.009), associated with age-related macular degeneration, did not pass the p-value threshold accounting for multiple testing. The replication of rs2014497 may have required a larger sample size to have enough statistical power to confirm a rare variant association since all cancer types that replicate in the UK Biobank have a number of cases below 500. In addition, the replication of the rare INDEL rs557998486 associated with macular degeneration may have required a larger sample size as well since there are only 2,726 cases for age-related macular degeneration in UK Biobank compared to the 3,685 cases in the GERA cohort, which implies that GERA has a larger effective sample size than UK Biobank (13,780 and 10,652, respectively). Moreover, the lower average age in the UK Biobank participants compared to that in GERA can also impair the statistical power to identify this association with age-related macular degeneration.

However, additional findings provides support for the majority of our novel disease-associated loci, including the association with biomarkers or related traits in UK Biobank and further lines of evidence in previous studies.

To give an example, three loci were associated in our study with *cardiovascular disease* (“any” cardiovascular disease, according to GERA cohort description). Two out of three associated loci have never been associated with related traits. The known locus, leaded in our study by the common additive variant rs2466455 (MAF = 0.213, p = 9.44×10^{-9}, OR = 0.90), was previously associated with ischemic stroke (LD $r^2 > 0.90$) (Malik et al., 2018; Ninds Stroke Genetics Network and International Stroke Genetics Consortium, 2016; Traylor et al., 2012) and atrial fibrillation (LD $r^2 > 0.90$) (Christophersen et al., 2017; Ellinor et al., 2010; Low et al., 2017; Nielsen et al., 2018; Roselli et al., 2018), which points out that, despite the vague definition of the phenotype in GERA, our study is still powerful for finding new associations.

In UK Biobank, the most similar phenotype with the largest number of cases is “Vascular/heart problems diagnosed by doctor: Stroke” with 5,587 cases. Hence, UK Biobank might be unpowered to replicate our novel findings if they are also related to ischemic stroke or atrial fibrillation.

Nevertheless, we found additional evidence for our new associations with cardiovascular disease. Among the two new loci, one is led by the intronic rs10858023 common variant (MAF = 0.349) at the *DCLRE1B* gene. This variant it is significant under the additive model (p = 3.26×10^{-8}, OR [CI 95%] = 1.09 [1.06-1.12]) but showed a highest significant p-value and a higher effect size when the dominant model was tested (p = 2.11×10^{-9}, OR [CI 95%] = 1.14 [1.09 – 1.19]) and, in addition, it deviates from the additivity according to the dominance deviation test (p = 0.019). Interestingly, this variant is associated in UK Biobank with hypothyroidism/myxoedema (OR [CI 95%]= 1.19 [1.16-1.23], p= 9.64×10^{-28}) as well as levothyroxine sodium treatment (beta [CI 95%]=1.18 [1.14-1.22], p= 2.67×10^{-20}). Hypothyroidism is a well-established risk factor for cardiovascular disease (Biondi and Klein, 2004). In addition, this variant is associated with insulin-like growth factor-1 (IGF-1) (beta [CI 95%]= -0.01 [-0.02--0.01], p = 3.95×10^{-4}) and triglycerides (beta [CI 95%]= -0.01 [-0.02--0.005], p = 4.95 × 10^{-4}), both related to hypothyroidism as well as cardiovascular disease. Interestingly, low levels of IGF-1 have been associated with a higher risk of atrial fibrillation (Busch et al., 2019) and ischemic stroke (Saber et al., 2017). For triglycerides, a high level of triglycerides in plasma is associated with an increased risk of cardiovascular disease (Miller et al., 2011; Nordestgaard and Varbo, 2014), and the coexistence of dyslipidemia and hypothyroidism is linked to the development of coronary heart disease (Duntas and Brenta, 2018).

The second new locus associated with cardiovascular disease in the GERA cohort has a low-frequency deletion as the top variant, rs201654520 (MAF = 0.017), an intronic recessive INDEL at *CACNB4*. Due to its low MAF, the replication of this variant in the UK Biobank is even harder than for rs10858023 at *DCLRE1B*. However, additional evidence is also found for this association. An intronic variant in *CACNB4* has been previously associated with idiopathic dilated cardiomyopathy in African Americans in a recent GWAS (Xu et al., 2018). However, the variant described is in linkage equilibrium with our low-frequency INDEL (rs150793926, LD $r^2 = 0.0016$), and this association has never been reported in European populations. In addition to its low MAF, this variant was only found genome-wide significant when the recessive model was tested (OR [CI 95%] = 19.02 [5.50-65.84], p = 4.32 × 10^{-8}) and could not have been found with the additive model (OR [CI 95%] = 1.10 [0.98-1.22], p = 1.10 × 10^{-1}, dominance deviation p = 4.36 × 10^{-6}). More evidence is found at the Target Validation Platform (https://www.targetvalidation.org/, accession in September 2019), where two vasodilator drugs in phase IV targeting *CACNB4*, Suloctidil and Bepridil, are linked to cardiovascular disease. Interestingly, among the 20 adverse effects for Bepridil, it is included cardiac failure (second position), *torsades de pointes* (a polymorphic ventricular tachycardia) (third position), and atrial fibrillation (fourth position).

Nevertheless, despite the limitations of replication for GERA findings using UK Biobank, the exclusively recessive variant rs77704739 near *PELO* replicates with UK Biobank (OR meta-analysis = 2.46 [1.88 – 3.21], p meta-analysis = 4.68 × 10^{-11}). Further lines of evidence for this association are also found in previous studies, where an independent region near *PELO* was previously associated to type 2 diabetes in Greenlandic population with a large recessive impact, but in a variant that is not in LD with our discovery top variant (rs870992, p = 1.8 × 10^{-8}, OR = 2.79, LD r^2 = 0.0009) (Grarup et al., 2018).

Also, the recessive INDEL rs557998486 associated with age-related macular degeneration replicated with UK Biobank. However, the meta-analysis p-value does not pass the multiple testing threshold correction when considering several diseases and models of inheritance (p meta-analysis = 3.29× 10^{-8}). In summary, further analysis, such as interrogating additional cohorts that are more similar to GERA than UK Biobank, or filtering young controls in UK Biobank, are needed to confirm its association.

Our exploratory functional annotation analysis for these two variants with a recessive effect, rs77704739 and rs557998486, suggested that rs77704739 locus contains active promoters and an active enhancer bounded by pancreatic islet-specific transcription factors, and rs557998486 might be a regulatory region in an open chromatin site in eye tissues. Hence, these results highlight the importance that non-additive variants may have in the etiology of complex diseases, such as type 2 diabetes and age-related macular degeneration.

Overall, after considering all the supporting evidence illustrated with many examples for the novel associations findings in this study, we consider that all our novel findings deserve future validations and follow-up analyses, and demonstrate the importance of a comprehensive analysis including non-additive models when performing GWAS. Hence, our findings suggest that future analyses, including those involving large scale biobanks and large meta-analyses consortia, should consider analyzing, or re-analyzing existing data, using different models of inheritance and multiple reference panels.

In our commitment to share our results with the scientific community, a public searchable database including non-additive effects for 16 M of variants and 22 phenotypes is available at the Type 2 Diabetes Knowledge Portal (http://www.type2diabetesgenetics.org).

Conclusions

I. Imputing genotypes using 1000G phase 3, UK10K, GoNL and HRC as reference panels, instead of using a single panel, was crucial to identify 28.7% (27 out of 94) of all the GWAS significant associations, and 57.7% (15 out of 26) of the novel GWAS findings.

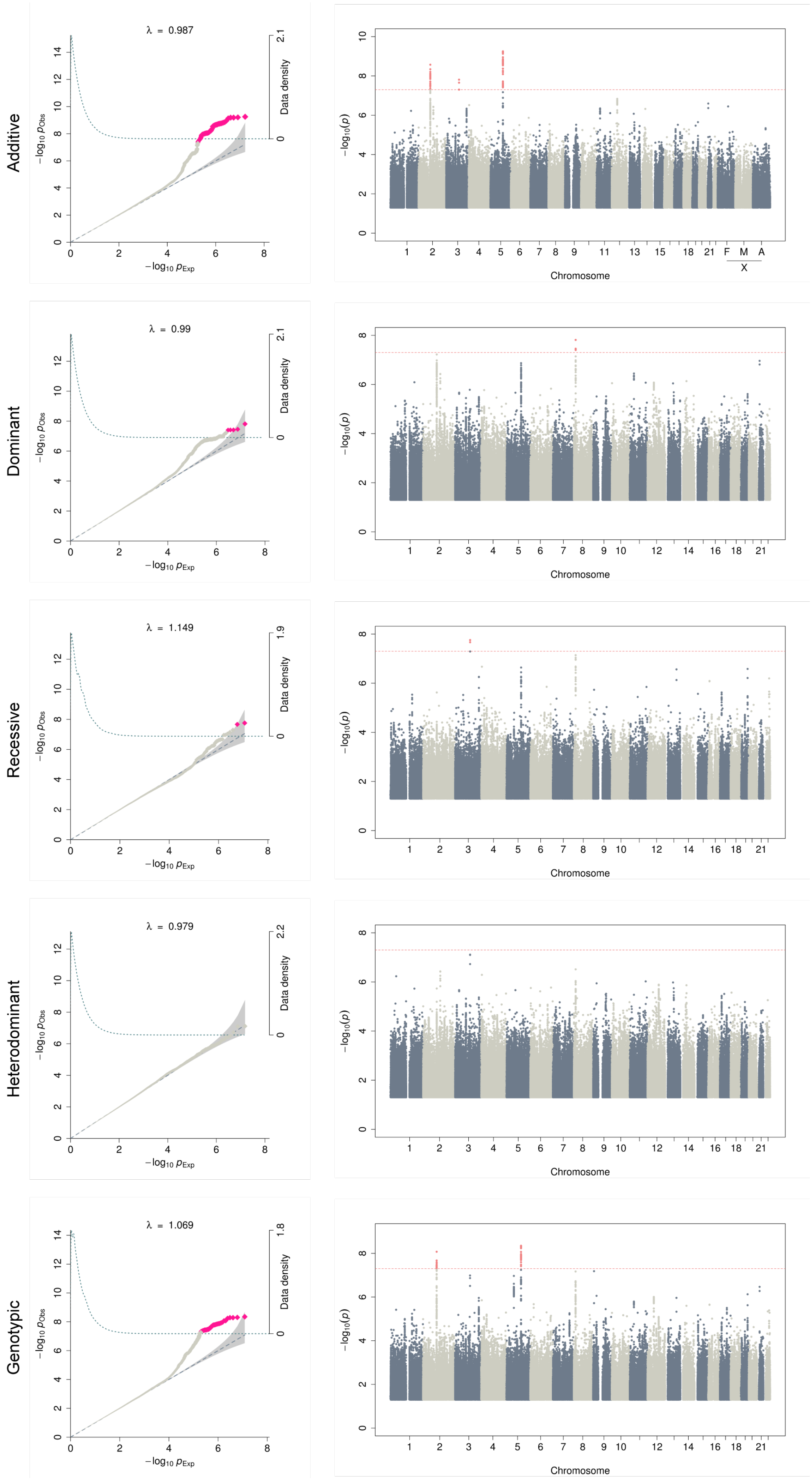
II. The integration of the results from multiple reference panels allows a better characterization of a wide spectrum of allelic frequencies as well as INDELs since HRC does not include them.

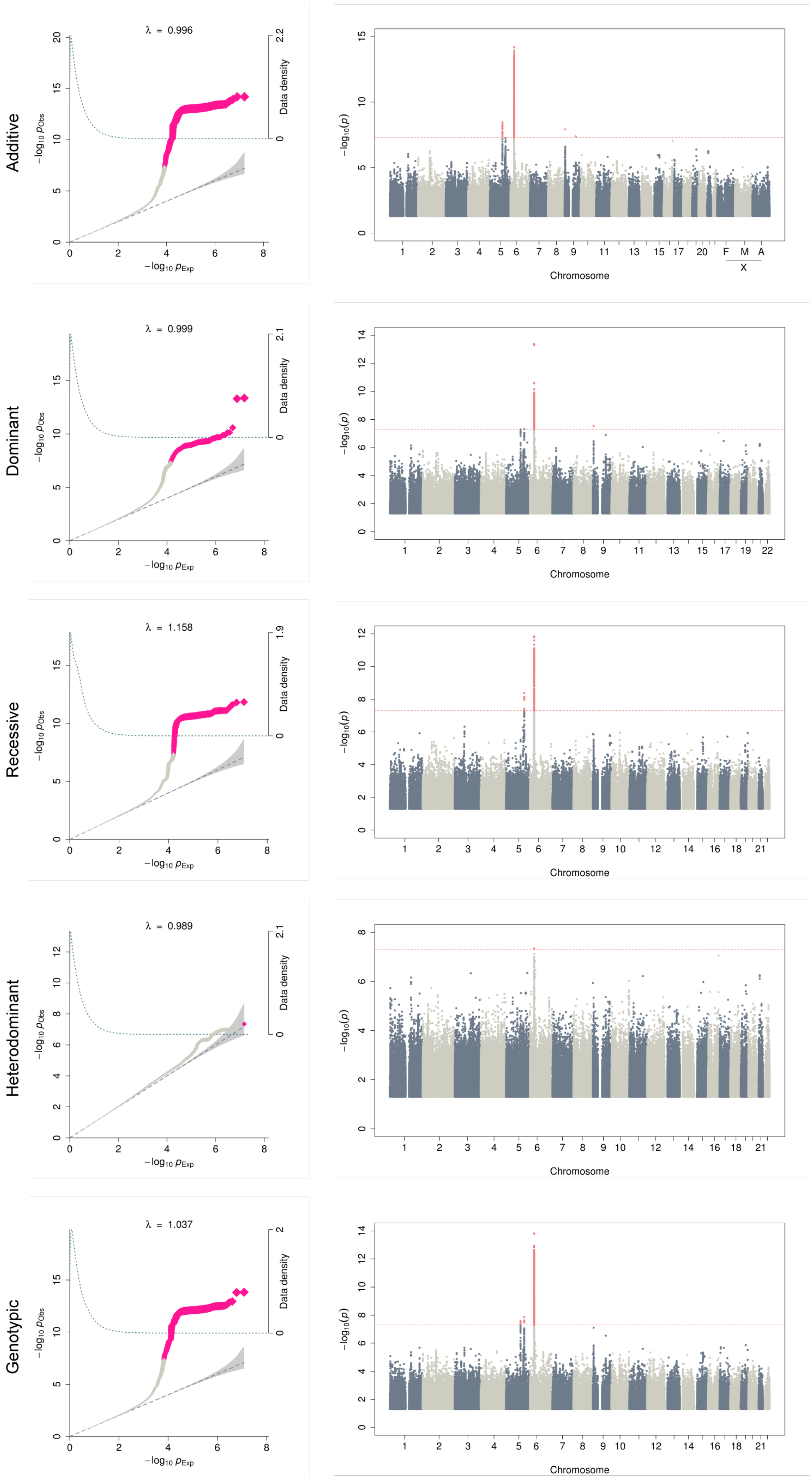
III. By analyzing non-additive inheritance models, we identified 20 loci out of 94 (21.3%) that reached the genome-wide significant threshold only when non-additive models were tested, and 13 out of the 26 (50%) novel loci were only identified by non-additive models. This highlights the importance of incorporating non-additive models into the association test.

IV. The development of GUIDANCE represents a step forward to promote comprehensive GWAS including the use of multiple panels for genotype imputation, non-additive models in the association testing and the X chromosome analysis.

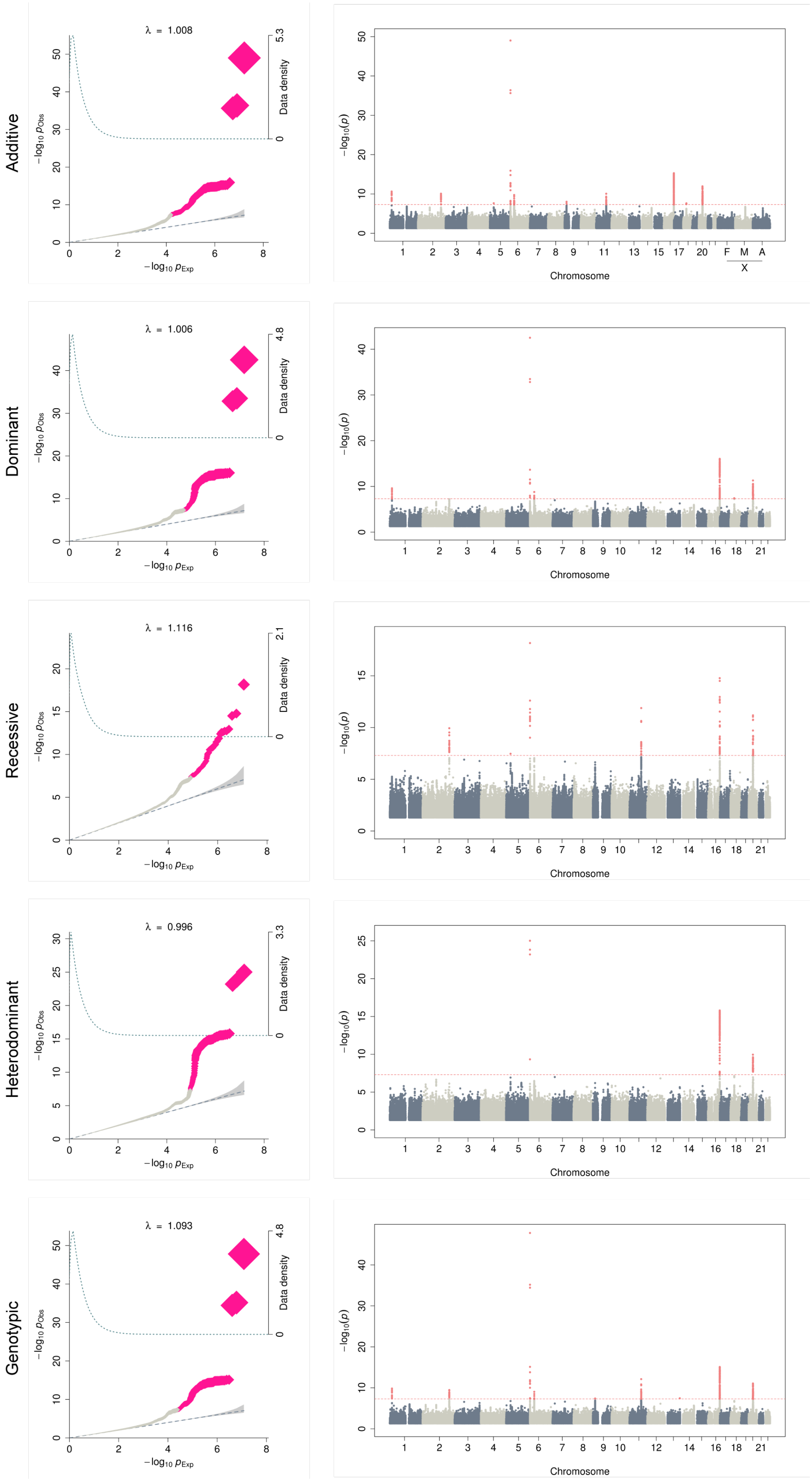
V. There is a need for making data and results publicly available. As it has been demonstrated in this study, the use of public individual-level genotype data, tools, and sequence-based reference panels is a cost-effective manner to exploit GWAS data.

Supplementary Material

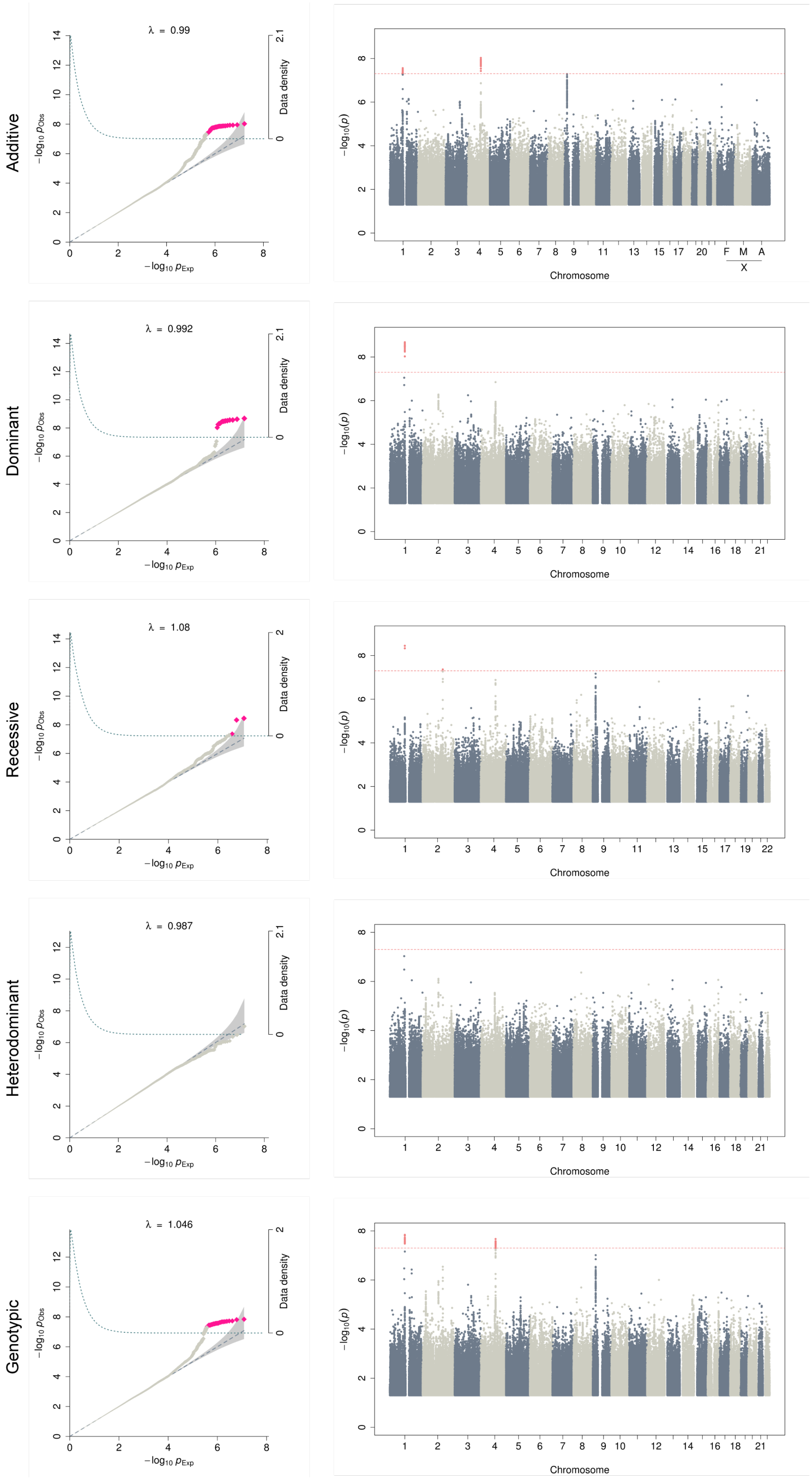

Additive
λ = 0.987
−log10 pObs
−log10 pExp
Data density
2.1
0
Chromosome
1 2 3 4 5 6 7 8 9 11 13 15 18 21 F M A
X
−log10(p)
Dominant
λ = 0.99
2.1
Recessive
λ = 1.149
1.9
Heterodominant
λ = 0.979
2.2
Genotypic
λ = 1.069
1.8
1 2 3 4 5 6 7 8 9 10 12 14 16 18 21



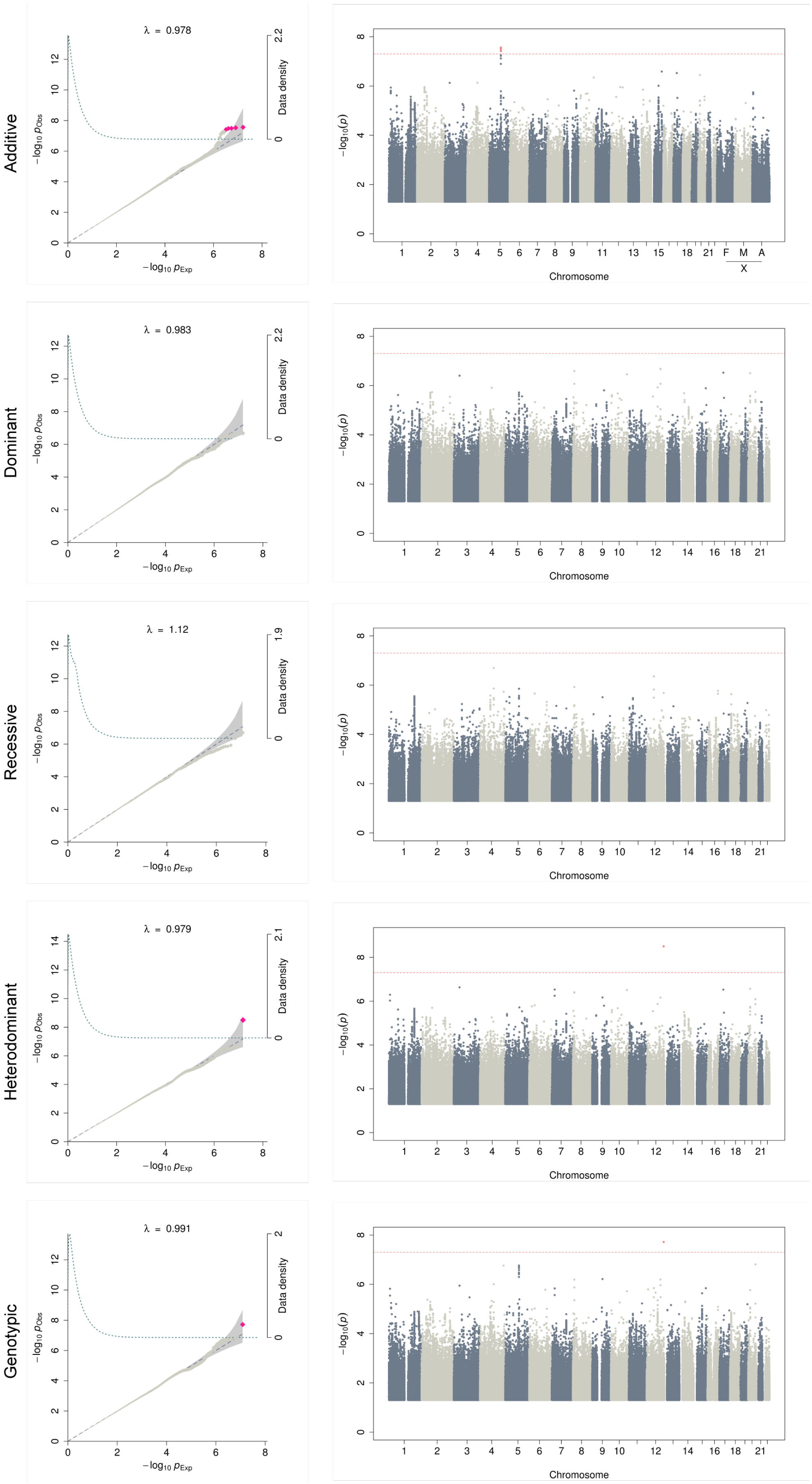
Supplementary Figure 2. Q-Q plots and Manhattan plots for asthma.



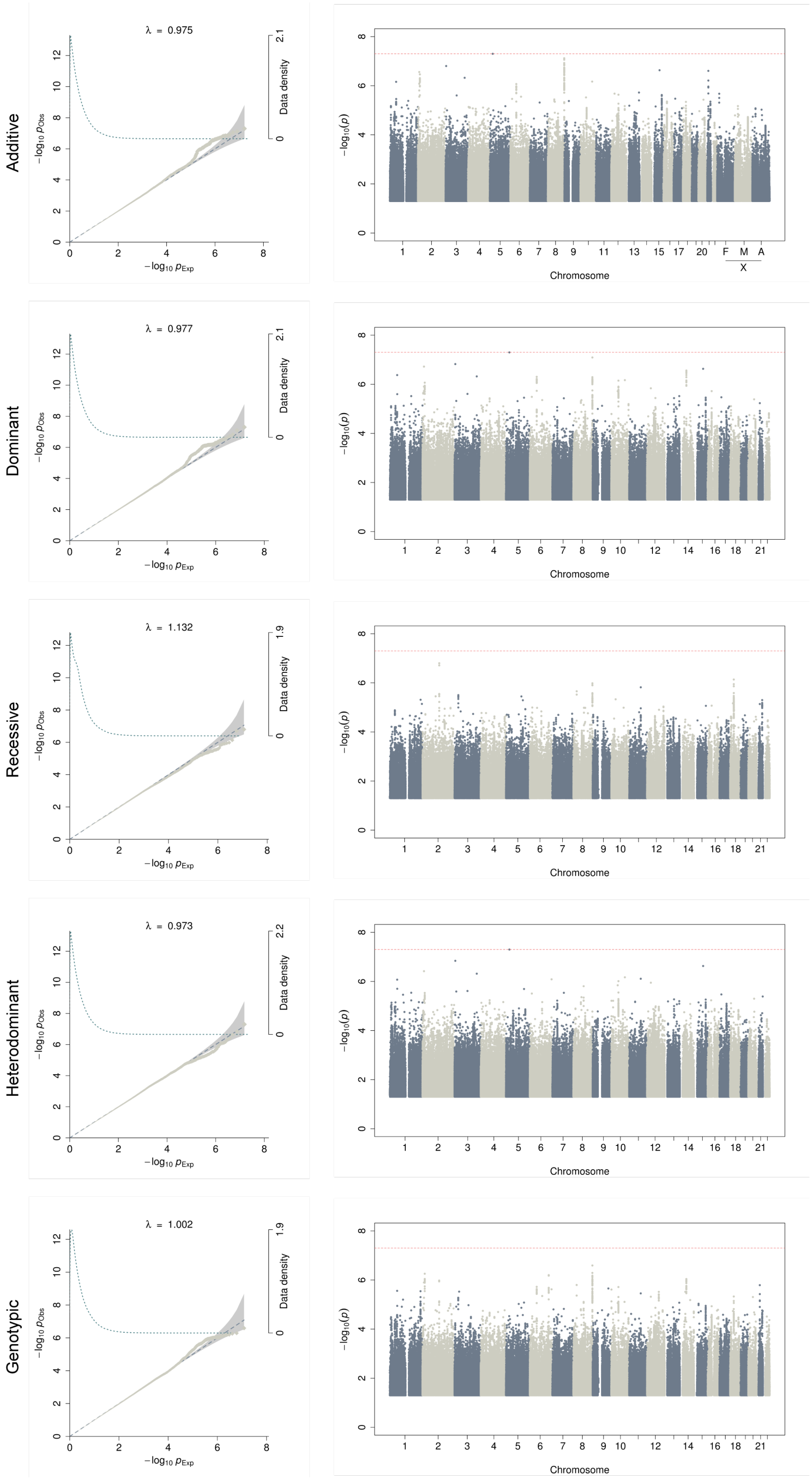
Supplementary Figure 3. Q-Q plots and Manhattan plots for cancer.



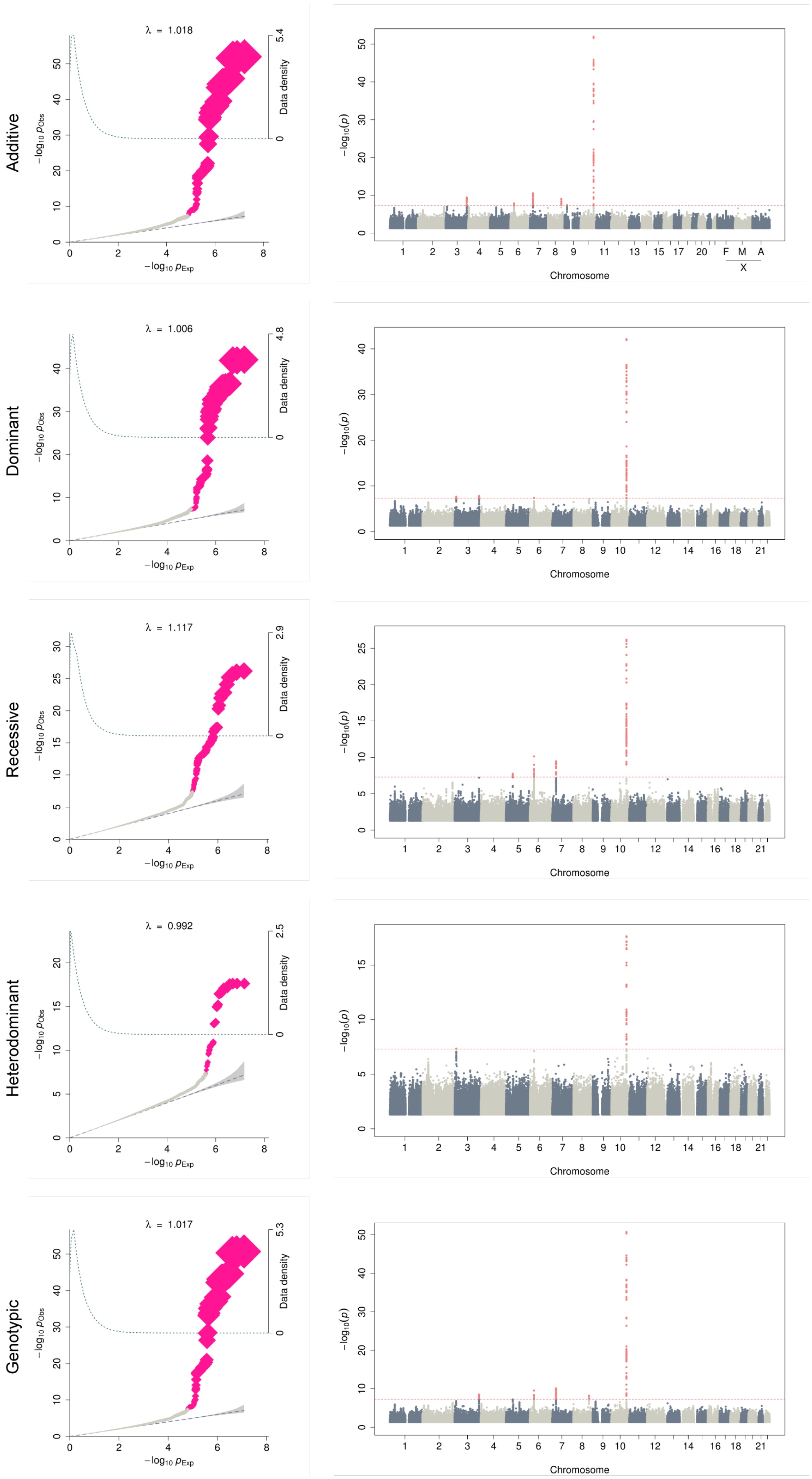
Supplementary Figure 4. Q-Q plots and Manhattan plots for cardiovascular disease.



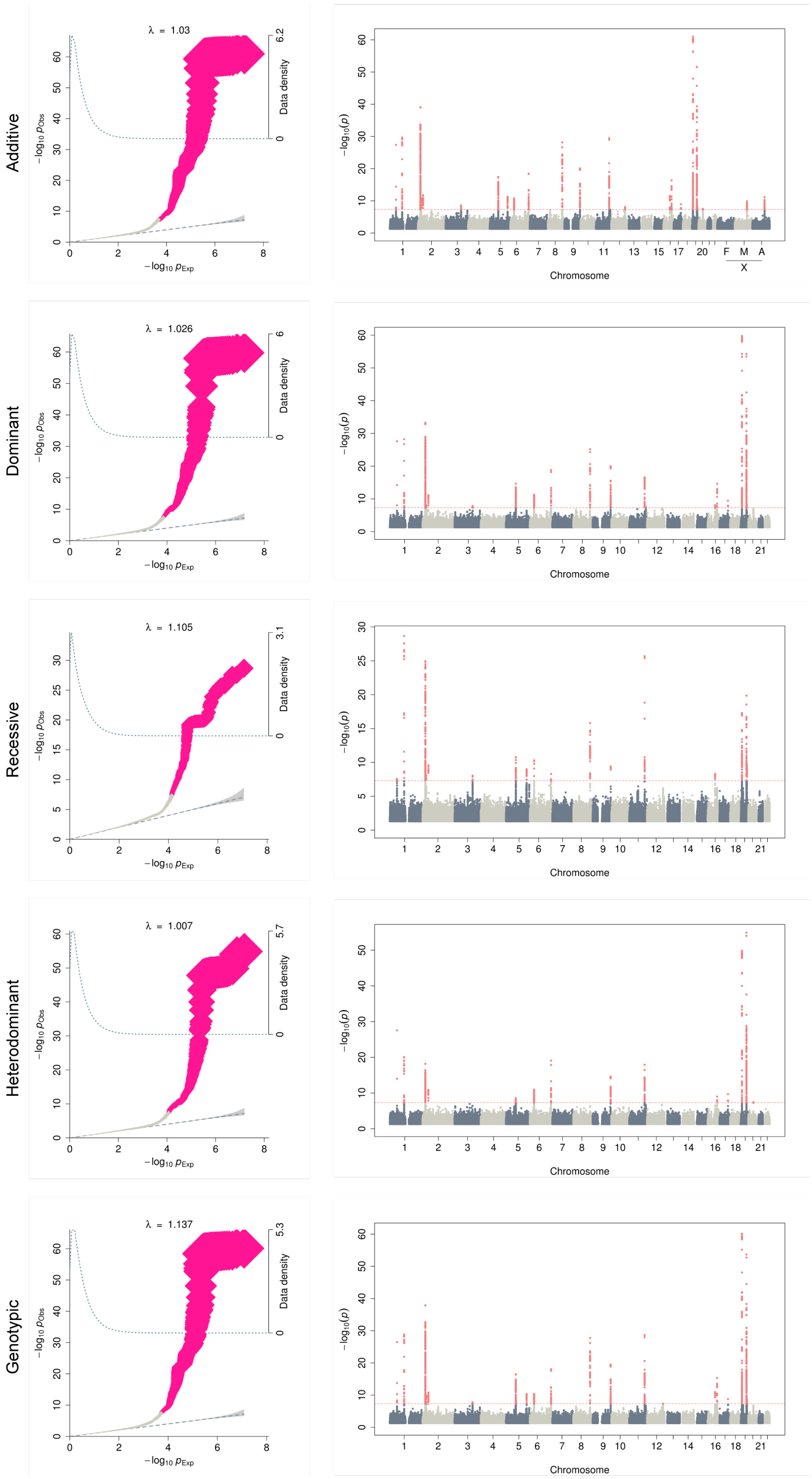
Supplementary Figure 5. Q-Q plots and Manhattan plots for major depressive disorder.



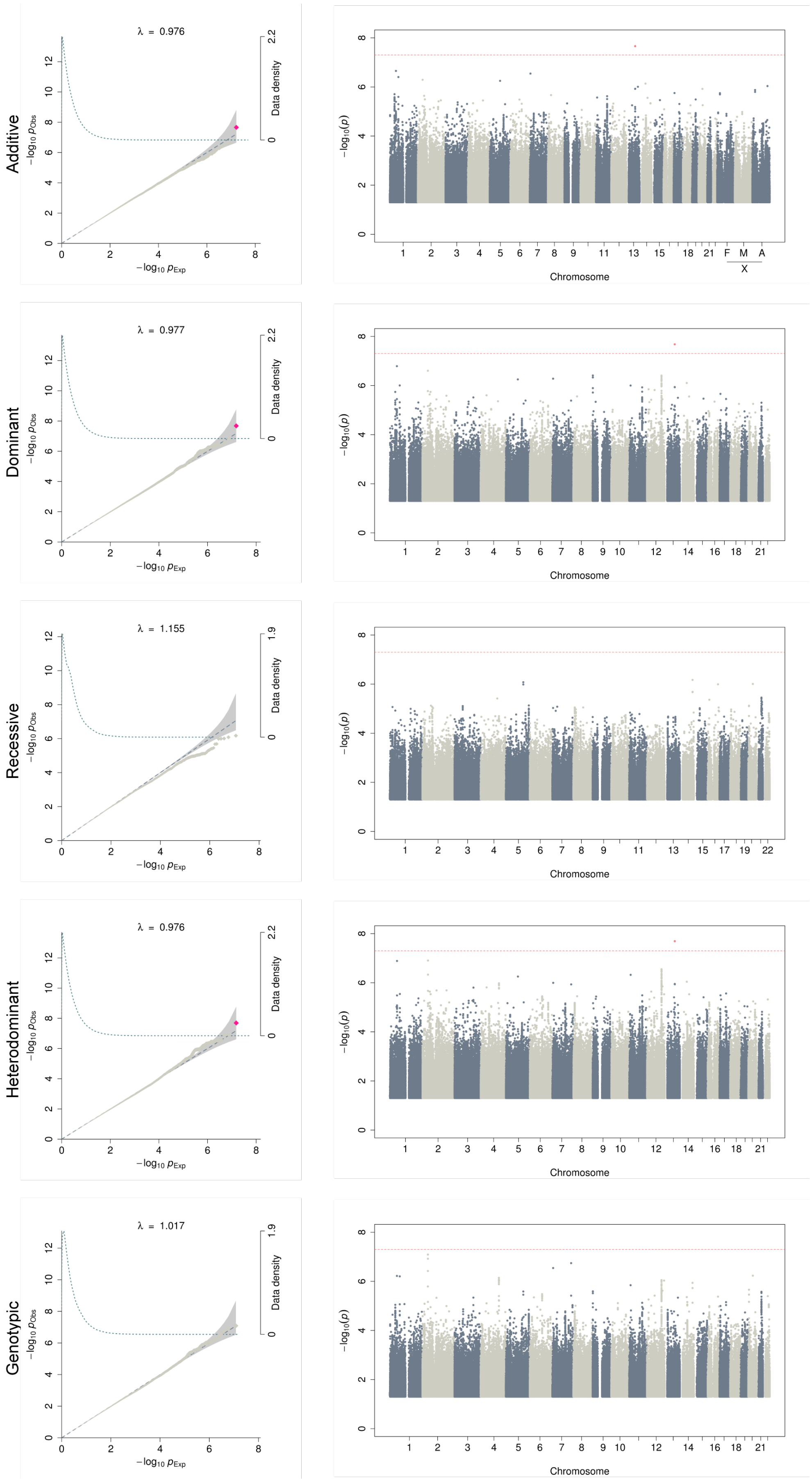
Supplementary Figure 6. Q-Q plots and Manhattan plots for dermatophytosis.



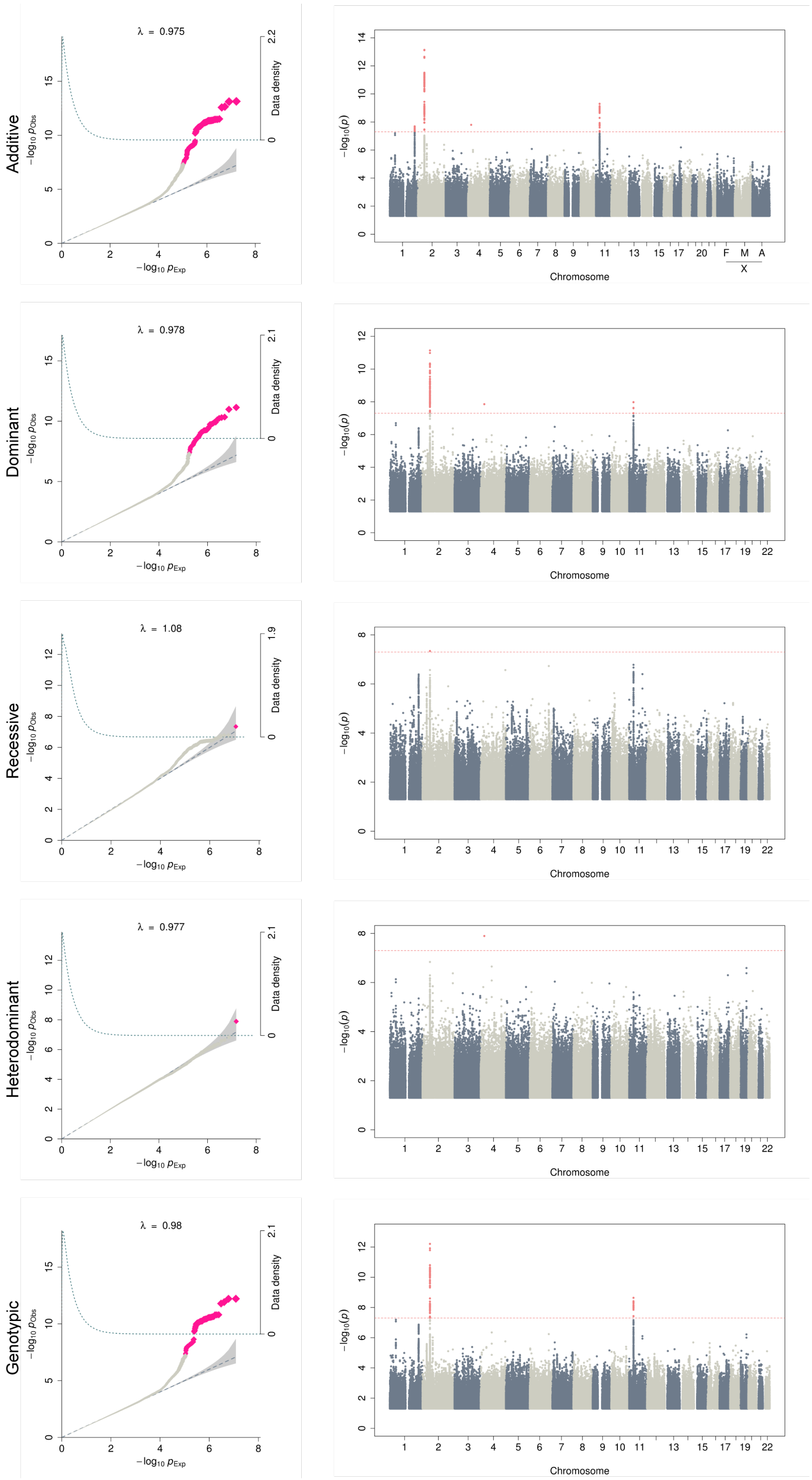
Supplementary Figure 7. Q-Q plots and Manhattan plots for type 2 diabetes.



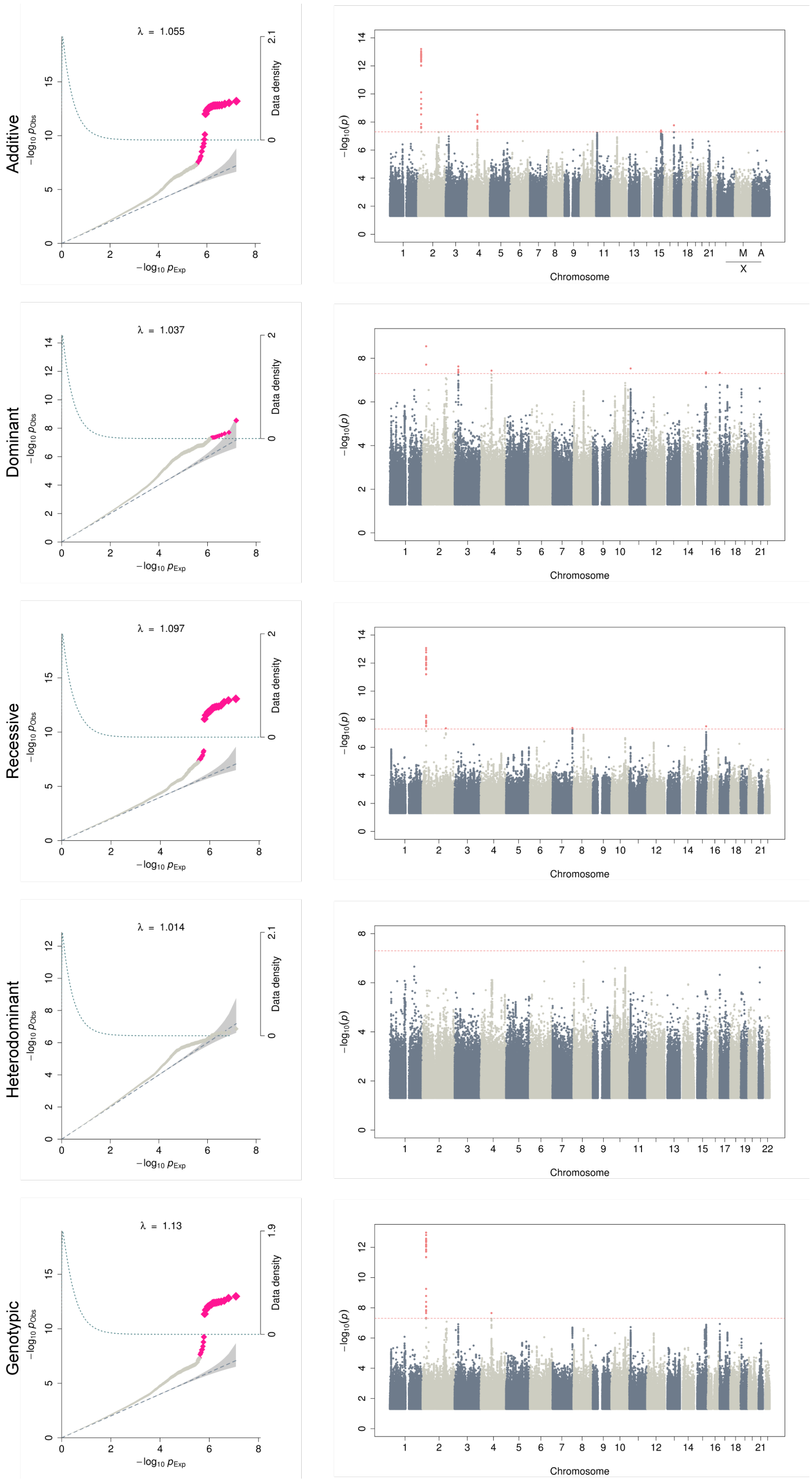
Supplementary Figure 8. Q-Q plots and Manhattan plots for dyslipidemia.



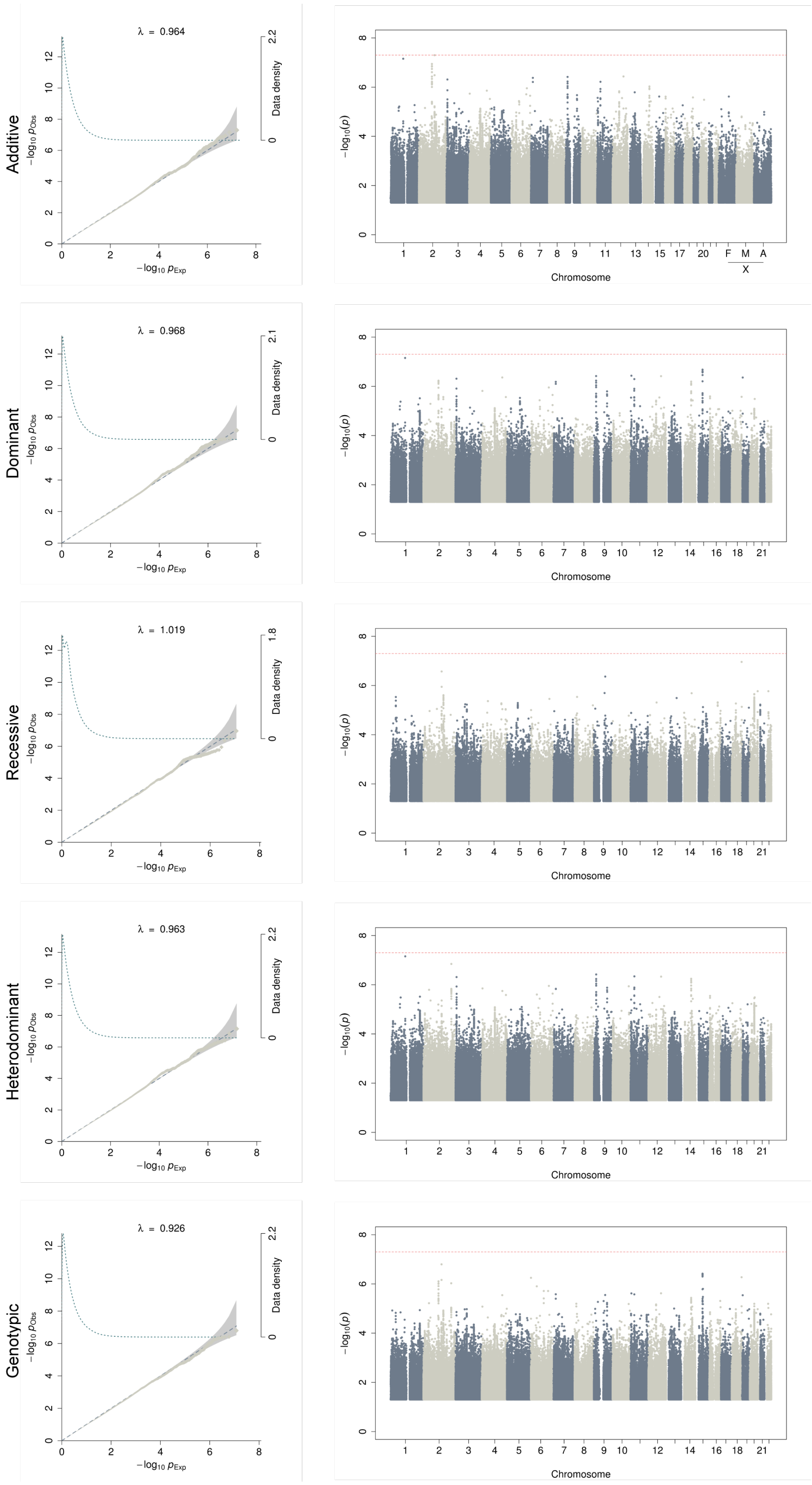
Supplementary Figure 9. Q-Q plots and Manhattan plots for hemorrhoids.



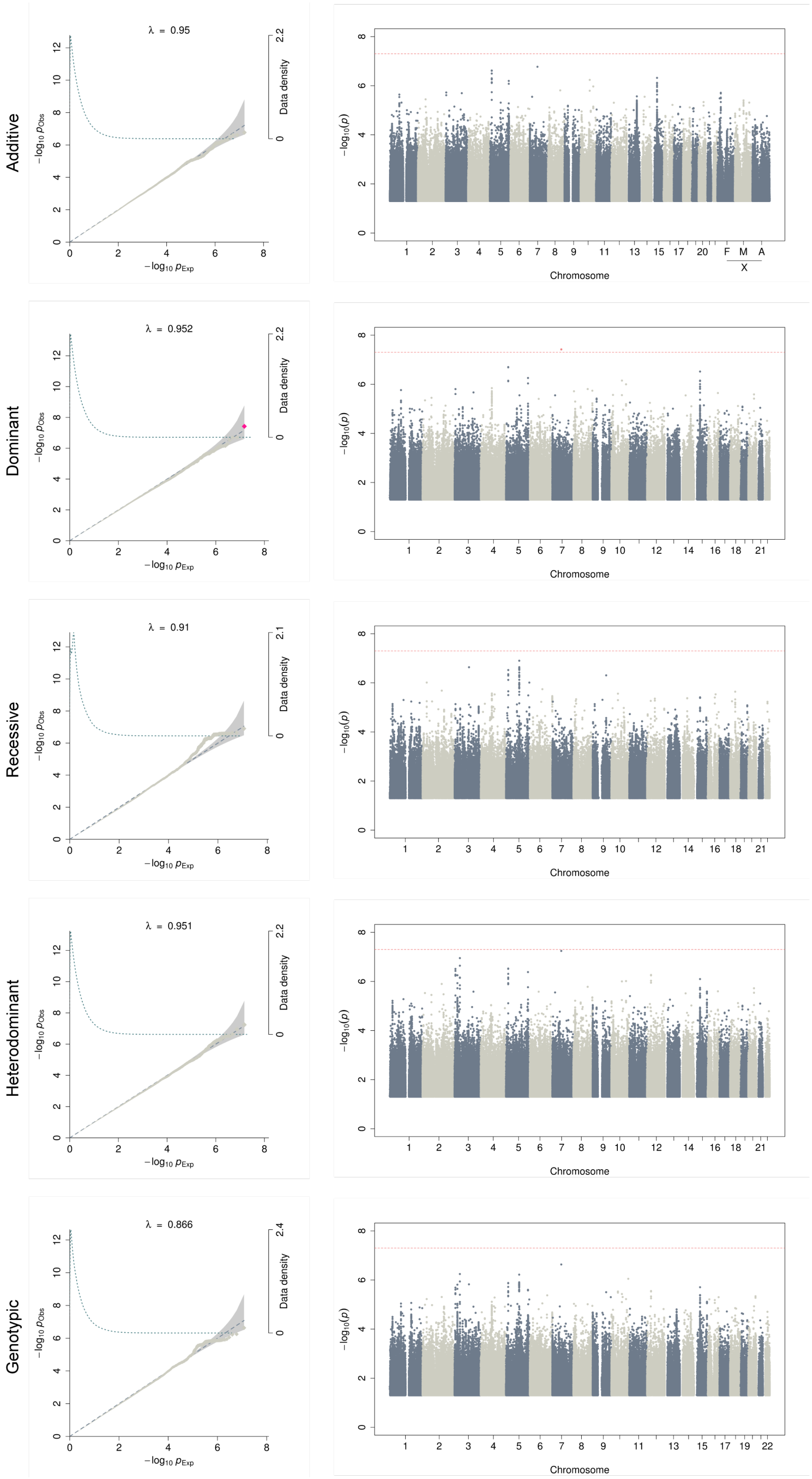
Supplementary Figure 10. Q-Q plots and Manhattan plots for hernia abdominopelvic cavity.



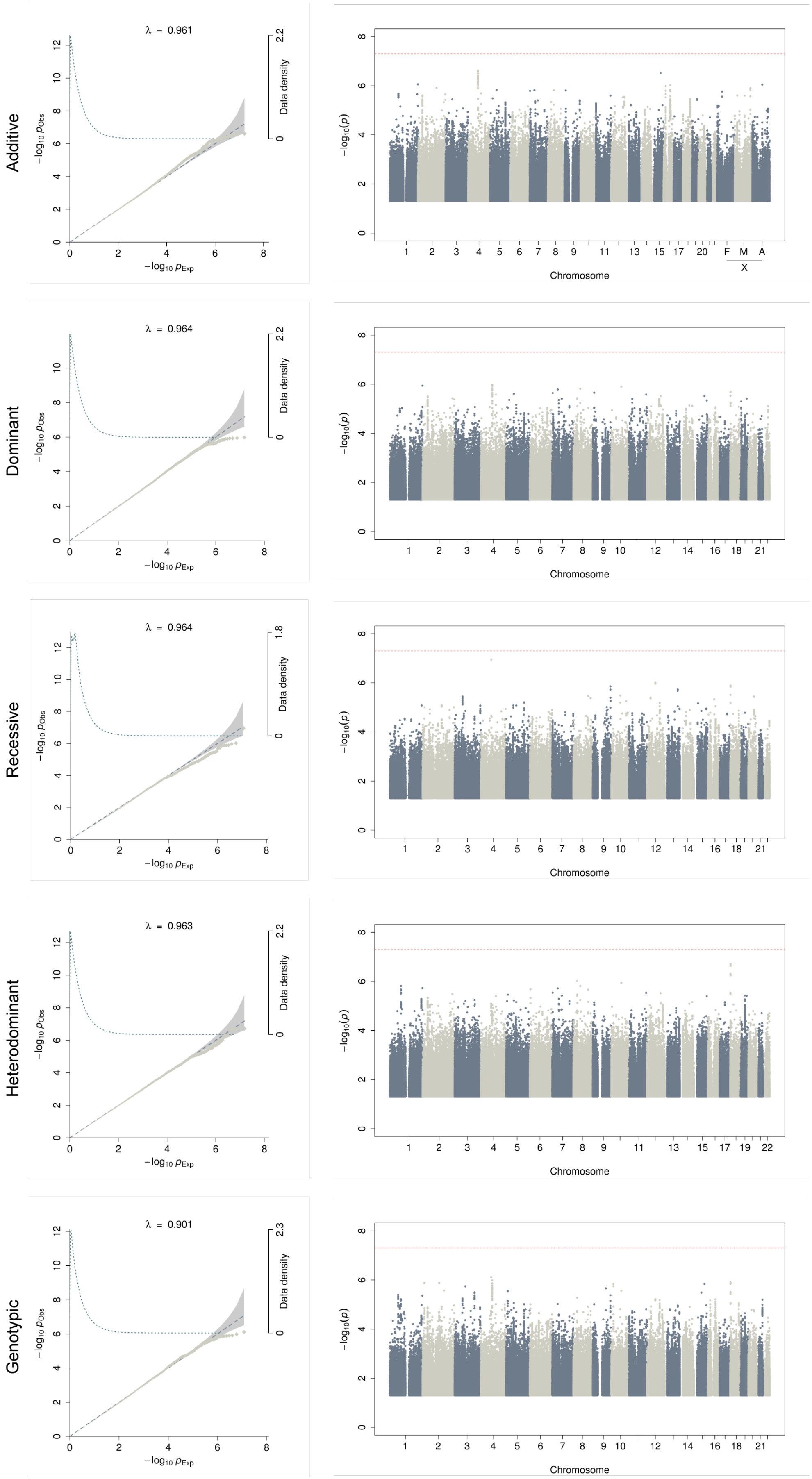
Supplementary Figure 11. Q-Q plots and Manhattan plots for hypertension.



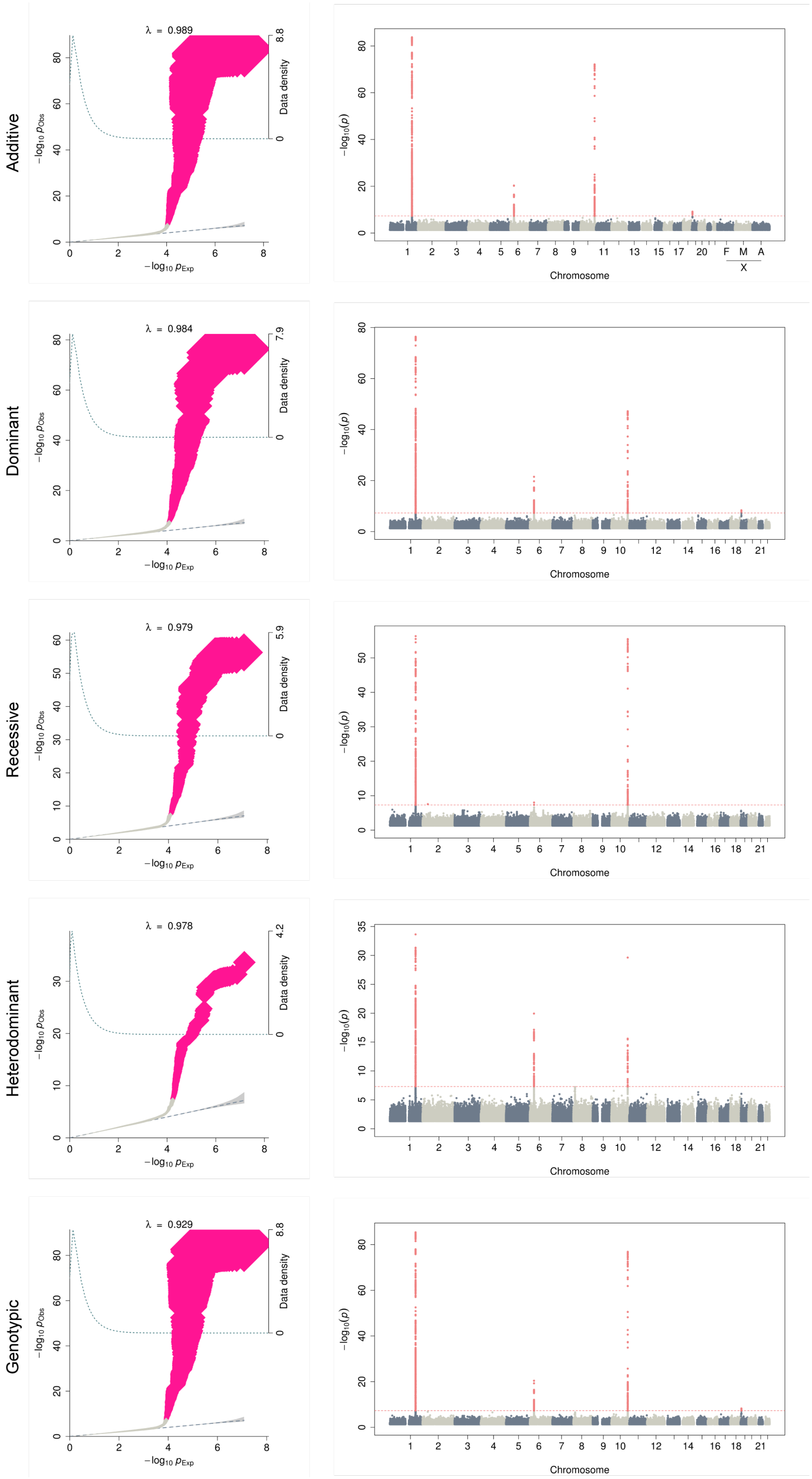
Supplementary Figure 12. Q-Q plots and Manhattan plots for insomnia.



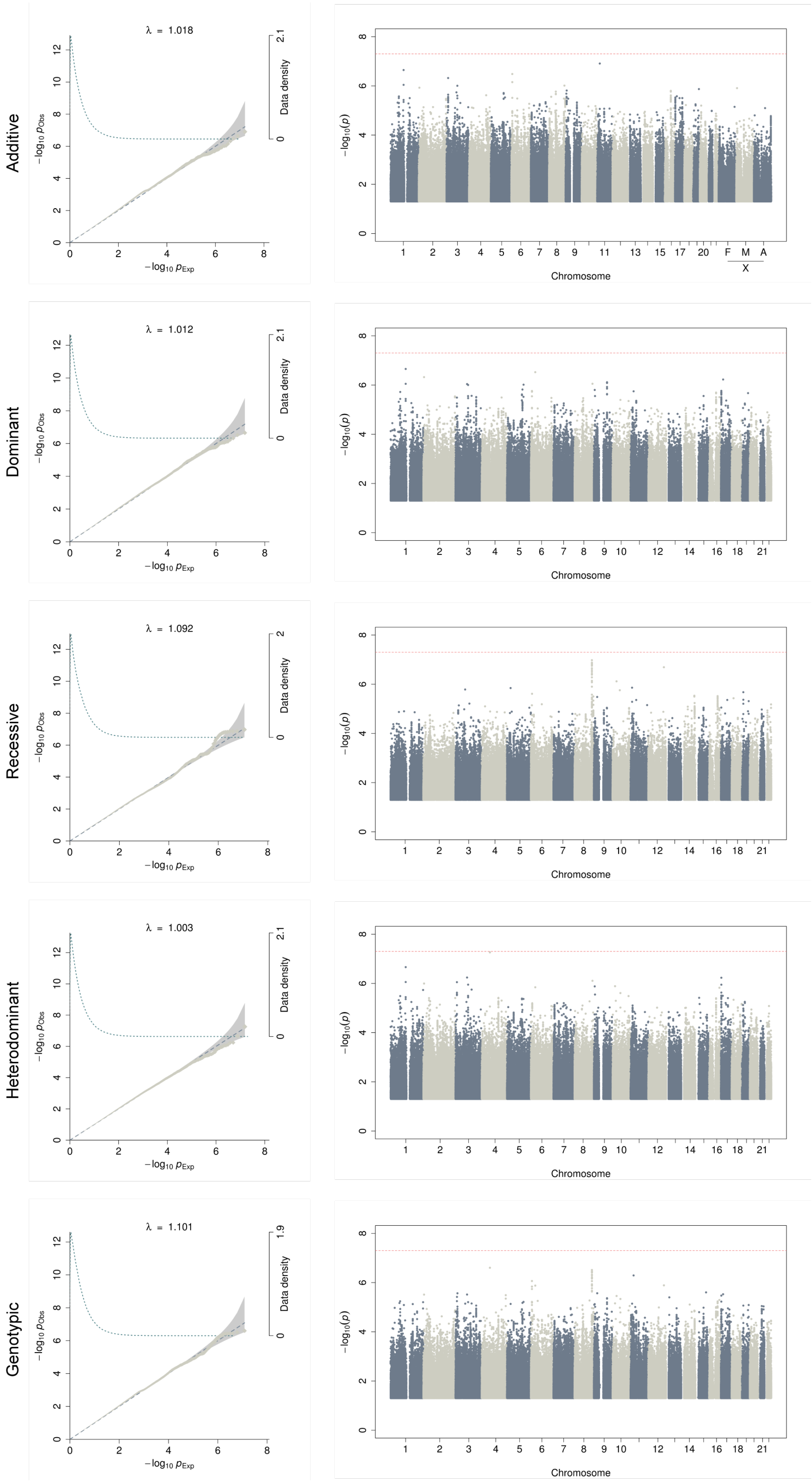
Supplementary Figure 13. Q-Q plots and Manhattan plots for iron deficiency



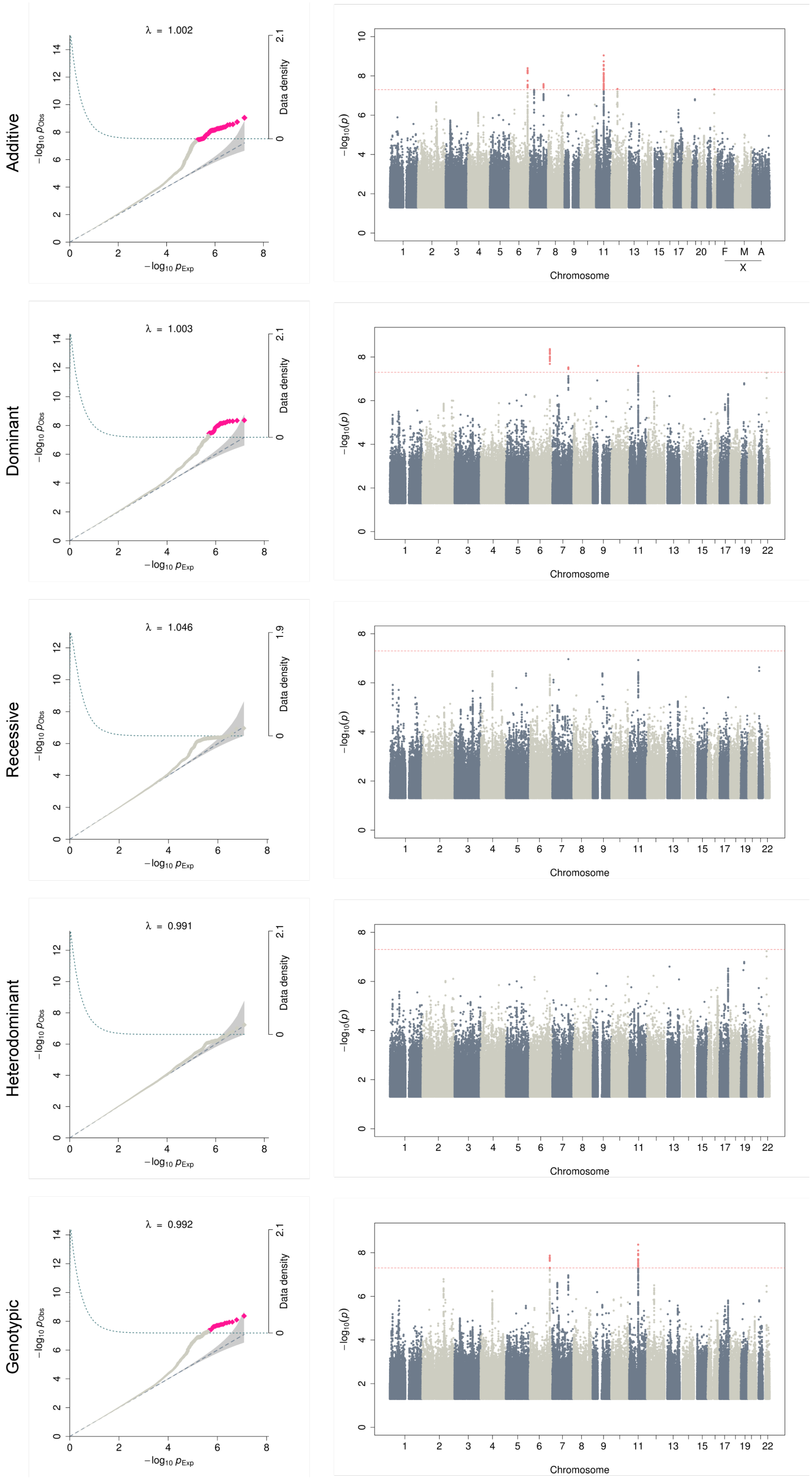
Supplementary Figure 14. Q-Q plots and Manhattan plots for irritable bowel.



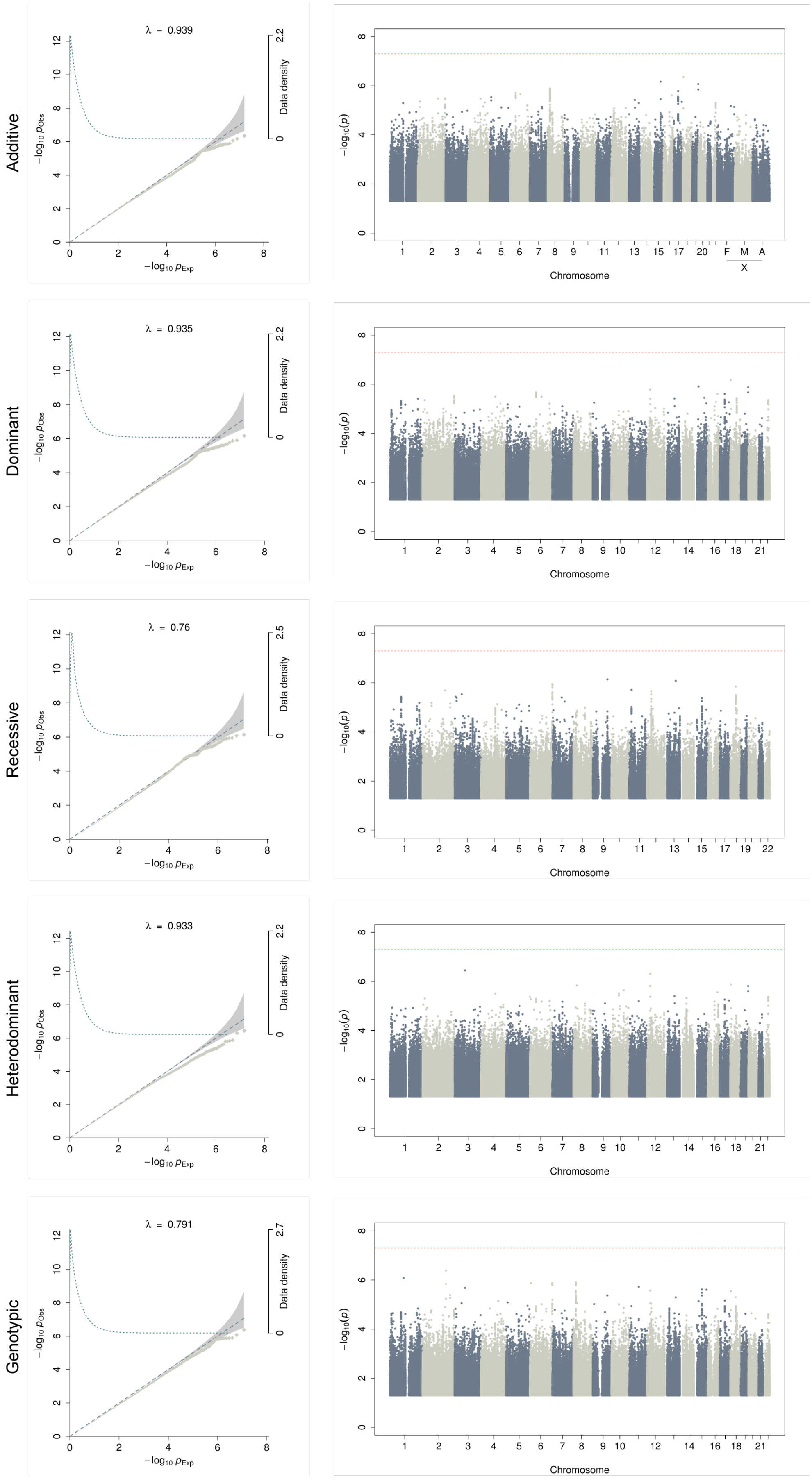
Supplementary Figure 15. Q-Q plots and Manhattan plots for macular degeneration.



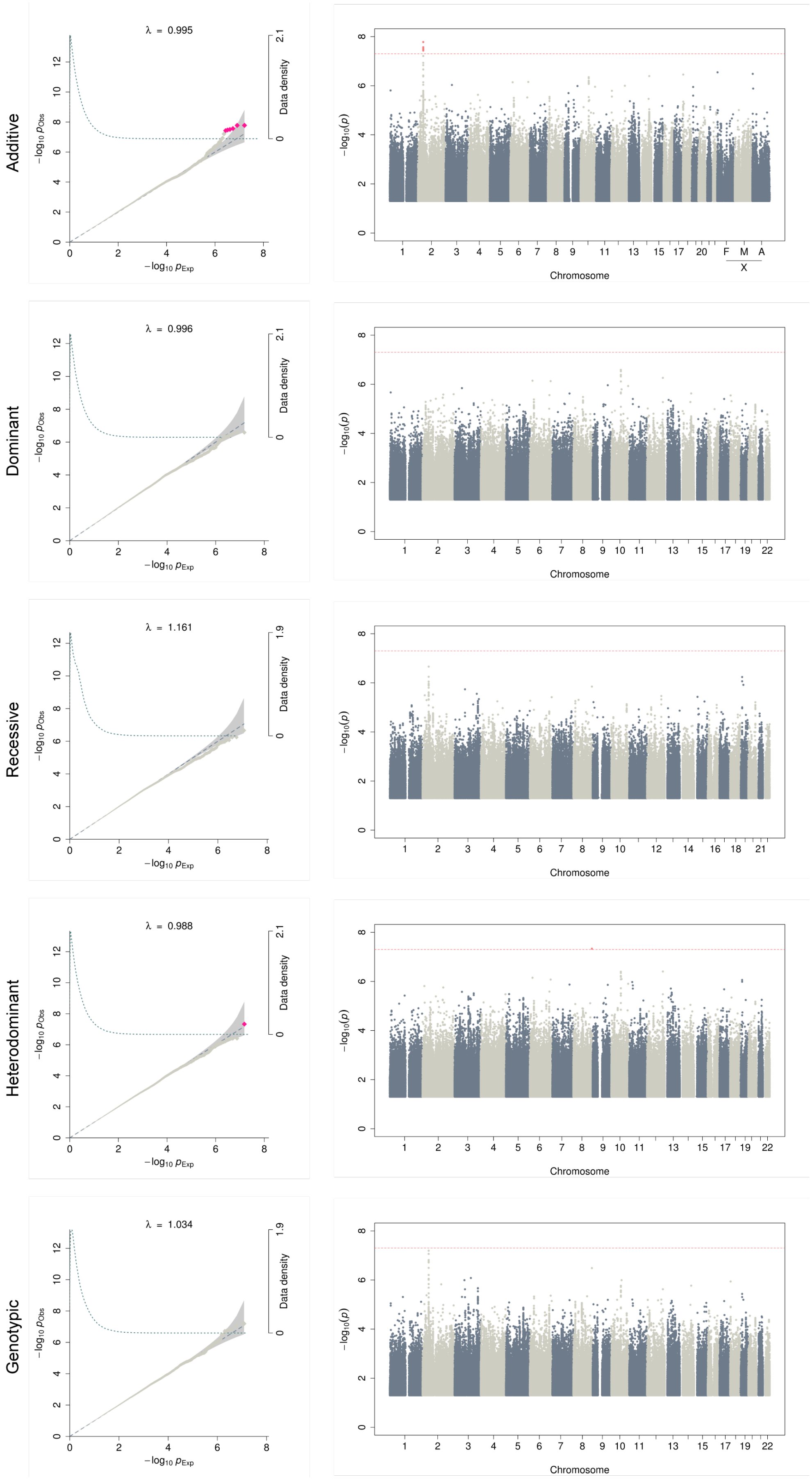
Supplementary Figure 16. Q-Q plots and Manhattan plots for osteoarthritis.



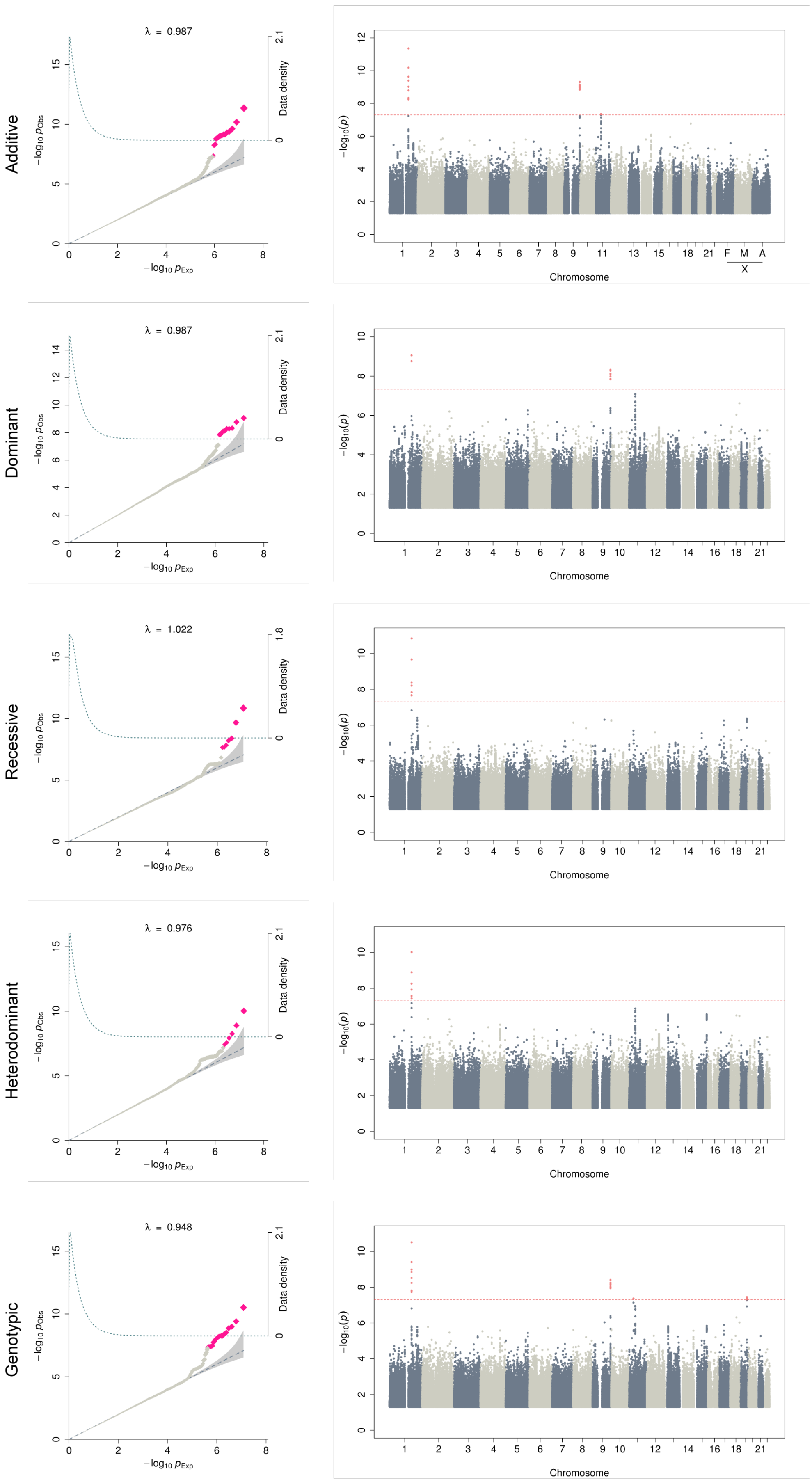
Supplementary Figure 17. Q-Q plots and Manhattan plots for osteoporosis.



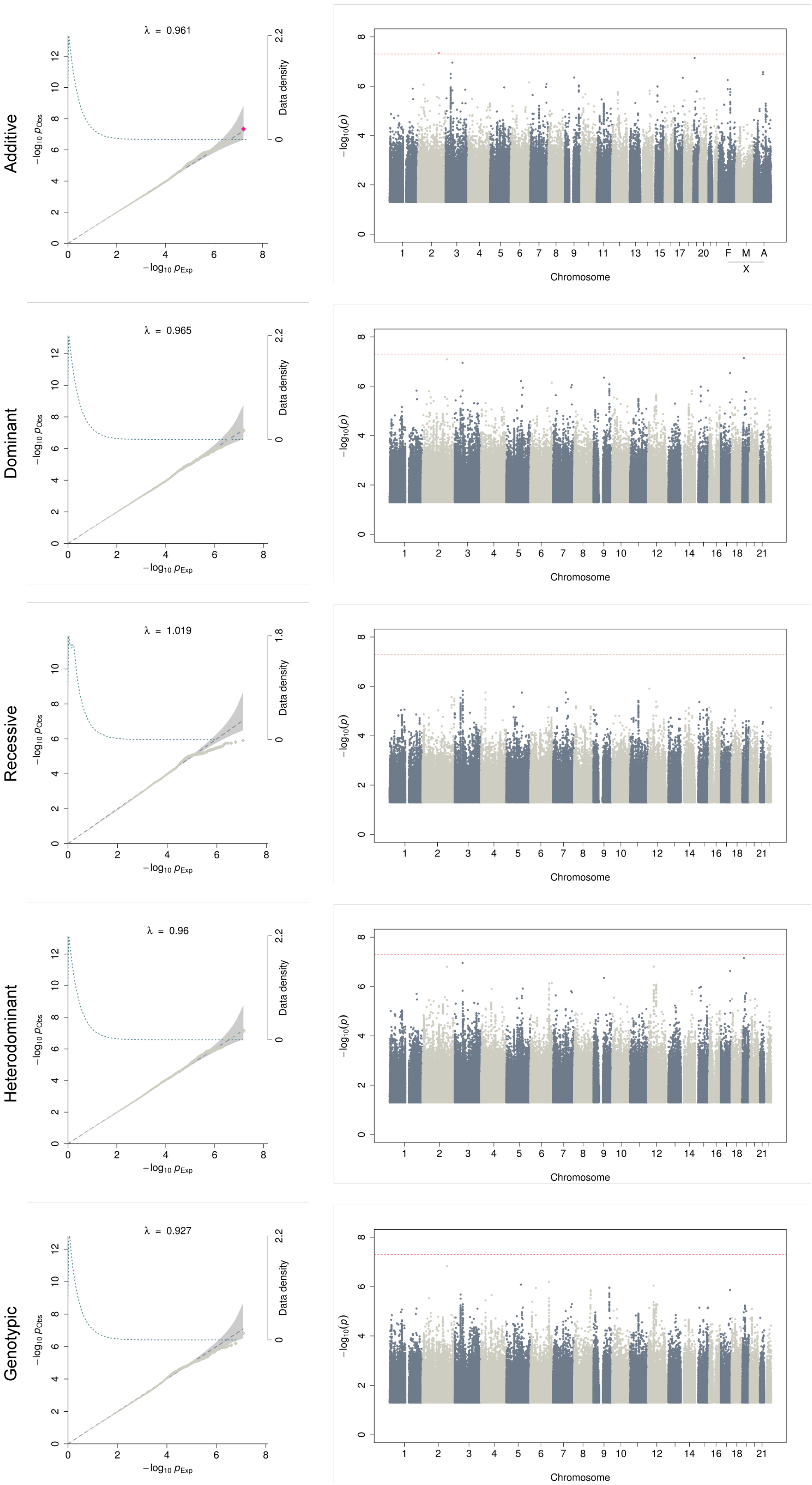
Supplementary Figure 18. Q-Q plots and Manhattan plots for peptic ulcers.



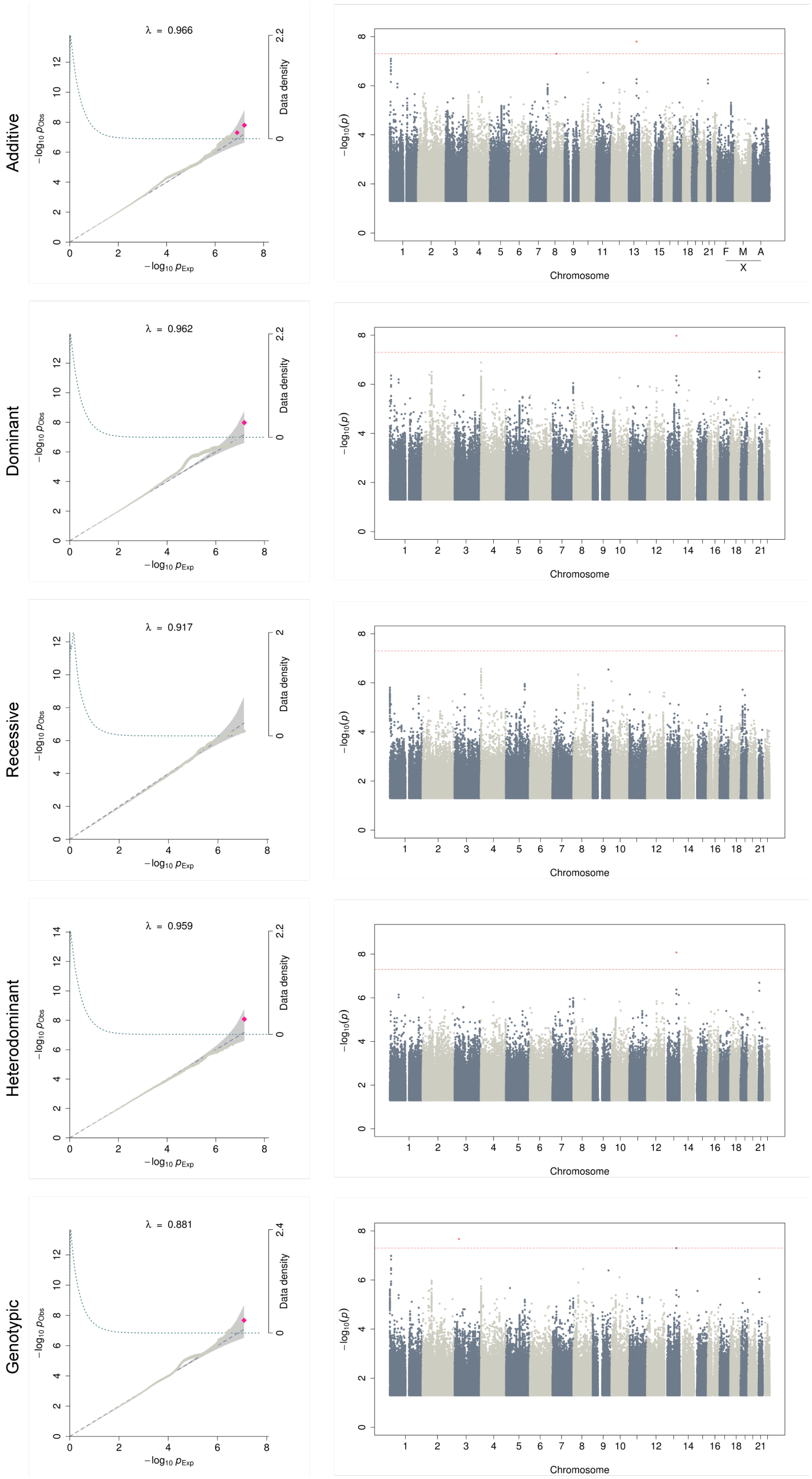
Supplementary Figure 19. Q-Q plots and Manhattan plots for psychiatric disorders.



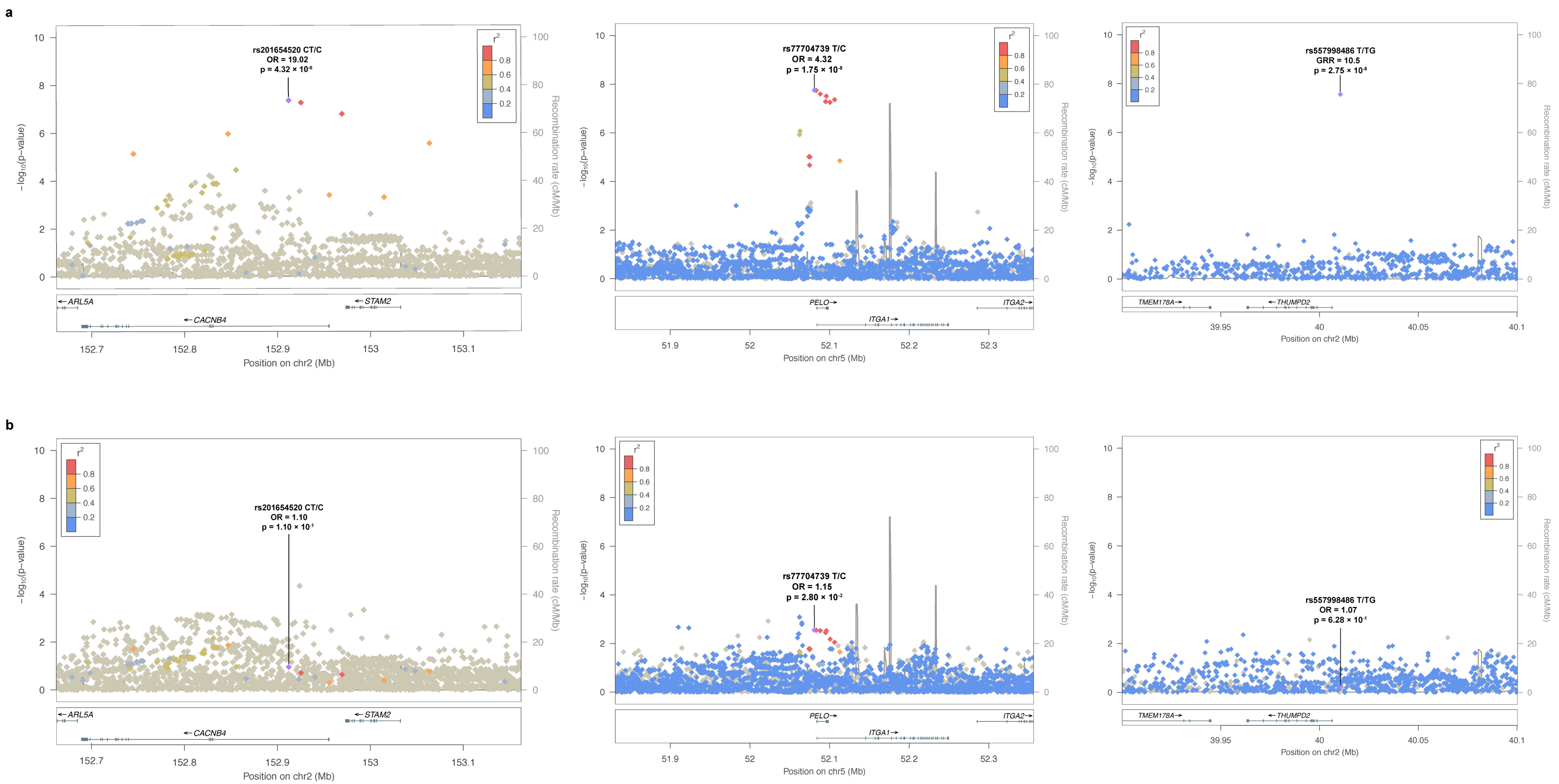
Supplementary Figure 20. Q-Q plots and Manhattan plots for peripheral vascular disease.



Supplementary Figure 21. Q-Q plots and Manhattan plots for acute reaction to stress.



Supplementary Figure 22. Q-Q plots and Manhattan plots for varicose veins.



Supplementary Figure 23. Comparison of the recessive and the additive model results for *CACNB4, PELO* and *THUMPD2* regions.

a LocusZooms for the recessive model for *CACNB4* (cardiovascular disease), *PELO* (type 2 diabetes) and *THUMPD2* (macular degeneration). **b** LocusZooms for the additive model for *CACNB4* (cardiovascular disease), *PELO* (type 2 diabetes) and *THUMPD2* (macular degeneration). These regions could not have been identified with the additive model.

Supplementary Table 1. Genome-wide significant top variants for each region.

										Lowest P-value Model		Additive Model		
Disease	rsID	CHR	Position	Allele A	Allele B	MAF	Info Score	GW Significant Models	Lowest P-value Model	P-Value	OR	P-Value	OR	Dominance deviation
Allergic Rhinitis	rs2160203	2	102,960,824	A	G	0.245	1.00	add,gen	add	2.68E-09	0.91	2.68E-09	0.91	1.83E-01
Allergic Rhinitis	rs2399472	3	112,911,615	C	T	0.073	1.00	add,rec	add	1.55E-08	1.17	1.55E-08	1.17	6.66E-01
Allergic Rhinitis	rs13166760	5	110,424,583	C	T	0.454	0.99	add,gen	add	5.62E-10	0.92	5.62E-10	0.92	9.90E-01
Allergic Rhinitis	rs10112506	8	13,164,746	A	G	0.390	1.00	dom	dom	1.54E-08	0.89	8.61E-06	0.94	2.86E-04
Asthma	rs252716	5	110,425,063	G	C	0.430	0.99	add,gen	add	3.53E-09	1.10	3.53E-09	1.10	9.21E-01
Asthma	rs154073	5	137,858,067	C	T	0.429	1.00	rec,gen	rec	4.23E-09	1.18	6.06E-08	1.09	9.28E-03
Asthma	rs9272521	6	32,606,479	G	A	0.437	0.84	add,dom,rec,gen,het	add	6.43E-15	1.15	6.43E-15	1.15	8.37E-02
Asthma	rs142807069	9	6,213,820	A	G	0.189	0.95	add,dom	add	1.22E-08	1.13	1.22E-08	1.13	6.20E-01
Asthma	rs67053006	9	98,344,866	C	G	0.139	0.88	add	add	4.14E-08	0.87	4.14E-08	0.87	8.10E-01
Cancer	rs6675912	1	17,769,290	T	G	0.351	0.99	add,dom,gen	add	2.47E-11	1.10	2.47E-11	1.10	4.12E-01
Cancer	rs3769823	2	202,122,995	A	G	0.300	1.00	add,rec,gen	add	8.47E-11	0.91	8.47E-11	0.91	3.16E-01
Cancer	rs16891982	5	33,951,693	C	G	0.038	0.98	add,dom,rec,gen,het	add	2.32E-08	1.24	2.32E-08	1.24	9.69E-01
Cancer	rs12203592	6	396,321	C	T	0.177	1.00	add,dom,rec,gen,het	add	9.74E-50	1.30	9.74E-50	1.30	5.86E-01
Cancer	rs41263822	6	32,632,447	C	A	0.421	0.87	add,dom,rec,gen,het	add	1.86E-10	1.10	1.86E-10	1.10	2.24E-01
Cancer	rs11445081	9	16,913,836	T	A	0.430	0.97	add,gen	add	9.71E-09	1.08	9.71E-09	1.08	2.93E-01
Cancer	rs1126809	11	89,017,961	G	A	0.279	1.00	add,rec,gen	gen	7.48E-13	1.15/0.92	8.81E-11	1.10	2.36E-04
Cancer	rs138646839	13	112,115,591	C	T	0.005	0.83	gen	gen	3.54E-08	80.11/0.02	1.45E-07	1.68	*NA*
Cancer	rs34659644	16	89,796,017	G	A	0.079	0.91	add,dom,rec,gen,het	dom	9.20E-17	1.26	1.08E-15	1.23	3.95E-02
Cancer	rs2014497	18	28,442,343	A	G	0.008	0.99	add,dom	add	2.44E-08	1.50	2.44E-08	1.50	6.00E-01
Cancer	rs6059655	20	32,665,748	A	G	0.079	0.97	add,dom,rec,gen,het	add	1.15E-12	0.83	1.15E-12	0.83	6.07E-01
Cancer	rs75653149	20	34,726,973	C	A	0.071	0.92	add,dom,gen,het	dom	7.32E-09	1.18	2.57E-08	1.16	9.68E-02
Cardiovascular Disorders	rs10858023	1	114,448,752	C	T	0.350	0.99	add,dom,rec,gen	dom	2.11E-09	1.14	3.26E-08	1.09	1.94E-02
Cardiovascular Disorders	rs201654520	2	152,912,244	CT	C	0.017	0.97	rec	rec	4.32E-08	19.02	1.10E-01	1.10	4.36E-06
Cardiovascular Disorders	rs2466455	4	111,685,615	C	T	0.213	1.00	add,gen	add	9.44E-09	0.90	9.44E-09	0.90	1.23E-01
Depression	rs76025409	5	103,783,801	G	C	0.365	0.97	add	add	2.74E-08	1.11	2.74E-08	1.11	7.31E-01
Depression	rs56978738	12	128,551,715	GT	G	0.281	0.90	gen,het	het	3.15E-09	1.18	3.00E-03	0.94	1.10E-06
Dyslipidemia	rs11591147	1	55,505,647	G	T	0.014	0.83	add,dom,rec,gen,het	dom	2.49E-28	0.52	4.21E-28	0.52	1.13E-01
Dyslipidemia	rs3832016	1	109,818,158	C	CT	0.208	0.87	add,dom,rec,gen,het	add	2.14E-30	1.21	2.14E-30	1.21	1.29E-01
Dyslipidemia	rs1367117	2	21,263,900	G	A	0.314	1.00	add,dom,rec,gen,het	add	8.61E-40	1.20	8.61E-40	1.20	6.73E-01
Dyslipidemia	rs1260326	2	27,730,940	T	C	0.424	1.00	add,dom,rec,gen	add	1.67E-11	0.92	1.67E-11	0.92	9.95E-01
Dyslipidemia	rs6544713	2	44,073,881	T	C	0.320	1.00	add,dom,rec,gen,het	add	1.79E-12	0.91	1.79E-12	0.91	9.58E-01
Dyslipidemia	rs113346874	3	135,846,911	T	G	0.233	0.99	add,dom,rec,gen	add	2.63E-09	1.09	2.63E-09	1.09	6.17E-01
Dyslipidemia	rs12916	5	74,656,539	T	C	0.409	1.00	add,dom,rec,gen,het	add	4.02E-18	1.12	4.02E-18	1.12	8.13E-01
Dyslipidemia	rs6882076	5	156,390,297	T	C	0.371	0.99	add,rec,gen	add	5.68E-12	1.10	5.68E-12	1.10	8.80E-01
Dyslipidemia	rs55651120	6	32,619,835	C	T	0.105	0.91	add,dom,rec,gen,het	dom	5.30E-12	1.18	4.51E-11	1.15	3.10E-02
Dyslipidemia	rs140570886	6	161,013,013	T	C	0.014	0.94	add,dom,rec,gen,het	het	8.13E-20	1.69	4.63E-19	1.66	3.75E-02
Dyslipidemia	rs28601761	8	126,500,031	C	G	0.417	0.98	add,dom,rec,gen	add	7.08E-29	0.86	7.08E-29	0.86	6.34E-02
Dyslipidemia	rs532436	9	136,149,830	G	A	0.194	1.00	add,dom,rec,gen,het	add	8.10E-21	1.16	8.10E-21	1.16	1.38E-01
Dyslipidemia	rs66505542	11	116,623,213	TA	T	0.153	1.00	add,dom,rec,gen,het	add	3.34E-30	0.82	3.34E-30	0.82	2.18E-01
Dyslipidemia	rs72085277	11	126,241,852	TTCTG	T	0.141	0.91	dom	dom	2.96E-08	1.13	5.71E-08	1.11	4.33E-01
Dyslipidemia	rs2649999	12	121,380,544	T	C	0.346	0.97	add,gen	add	8.38E-09	0.92	8.38E-09	0.92	6.90E-01
Dyslipidemia	rs3764261	16	56,993,324	C	A	0.317	1.00	add,dom,rec,gen	add	2.85E-12	0.91	2.85E-12	0.91	1.80E-01
Dyslipidemia	rs34042070	16	72,101,525	C	G	0.200	0.99	add,dom,gen,het	add	4.53E-17	1.15	4.53E-17	1.15	9.52E-01
Dyslipidemia	rs75003668	17	64,195,431	A	G	0.034	0.96	add,dom,gen,het	het	2.13E-10	1.27	1.14E-09	1.25	5.56E-02
Dyslipidemia	rs17248720	19	11,198,187	C	T	0.118	0.98	add,dom,rec,gen,het	add	1.02E-61	0.72	1.02E-61	0.72	1.43E-01
Dyslipidemia	rs118147862	19	45,319,631	G	A	0.038	0.89	add,dom,rec,gen,het	het	1.30E-55	0.56	2.48E-52	0.58	1.74E-05
Dyslipidemia	rs681343	19	49,206,462	C	T	0.482	1.00	add,dom,rec,gen	rec	2.58E-10	1.15	4.65E-08	1.07	1.27E-03
Dyslipidemia	rs1012167	20	39,159,119	T	C	0.387	1.00	add,het	add	2.92E-08	0.93	2.92E-08	0.93	9.77E-01
Dyslipidemia	rs67648651	23	109,693,274	T	C	0.420	0.99	add	add	1.21E-10	0.83	1.21E-10	0.83	-
Hemorrhoids	rs186102686	13	76,281,808	C	T	0.004	0.90	add,dom,het	het	2.03E-08	1.99	2.18E-08	1.98	*NA*
Hernia Abdominopelvic	rs2494196	1	219,762,581	C	A	0.274	1.00	add	add	2.03E-08	1.13	2.03E-08	1.13	6.87E-01
Hernia Abdominopelvic	rs3791679	2	56,096,892	A	G	0.237	1.00	add,dom,rec,gen	add	7.21E-14	0.84	7.21E-14	0.84	0.6564

Hernia Abdominopelvic	rs113180595	4	27,019,359	T	C	0.004	0.87	add,dom,het	het	1.27E-08	2.18	1.59E-08	2.17	*NA*
Hernia Abdominopelvic	rs66798575	11	32,451,920	T	G	0.364	0.97	add,dom,gen	add	5.02E-10	0.88	5.02E-10	0.88	2.69E-01
Hypertension	rs1275923	2	26,932,796	C	T	0.391	0.96	add,dom,rec,gen	add	6.24E-14	0.90	6.24E-14	0.90	9.20E-02
Hypertension	rs1446802	2	176,532,019	A	G	0.500	1.00	rec	rec	4.42E-08	1.13	1.66E-06	1.07	6.85E-03
Hypertension	rs72375069	3	27,427,821	CAATT	C	0.354	1.00	dom	dom	2.37E-08	0.90	1.04E-07	0.93	5.64E-02
Hypertension	rs16998073	4	81,184,341	A	T	0.290	0.99	add,dom,gen	add	3.01E-09	1.09	3.01E-09	1.09	8.40E-01
Hypertension	rs56012466	7	151,406,788	G	A	0.279	0.93	rec	rec	4.22E-08	1.23	2.23E-05	1.07	3.74E-04
Hypertension	rs72850439	11	10,274,381	T	A	0.286	0.82	dom	dom	2.93E-08	1.12	1.12E-07	1.09	5.07E-02
Hypertension	rs7174250	15	81,018,587	C	T	0.463	1.00	add	add	4.01E-08	1.08	4.01E-08	1.08	8.06E-01
Hypertension	rs28792763	15	90,081,905	G	A	0.462	0.96	dom	dom	4.42E-08	0.88	4.14E-06	0.94	4.80E-03
Hypertension	rs57515981	15	91,412,850	A	AAAGGCAG	0.482	0.83	rec	rec	3.20E-08	1.15	7.56E-06	1.07	9.19E-04
Hypertension	rs112963849	17	1,959,826	C	A	0.082	0.92	add,dom	add	1.71E-08	1.15	1.71E-08	1.15	8.01E-01
Iron Deficiency	rs79798837	7	67,292,424	C	T	0.118	0.94	dom	dom	3.80E-08	0.74	1.69E-07	0.77	8.92E-02
Macular Degeneration	rs488380	1	196,664,505	C	T	0.377	1.00	add,dom,rec,gen,het	gen	4.05E-86	0.59/0.87	1.88E-84	0.60	1.77E-04
Macular Degeneration	rs557998486	2	40,010,523	T	TG	0.009	0.90	rec	rec	2.75E-08	> 100	6.28E-01	1.07	*NA*
Macular Degeneration	rs556679	6	31,894,355	C	T	0.111	1.00	add,dom,rec,gen,het	dom	3.27E-22	0.63	4.83E-21	0.66	1.45E-02
Macular Degeneration	rs3750847	10	124,215,421	C	T	0.215	1.00	add,dom,rec,gen,het	gen	1.41E-77	1.91/0.78	7.85E-73	1.72	1.15E-07
Macular Degeneration	rs550946885	19	6,722,832	GTTTTT	G	0.217	0.85	add,dom,gen	add	7.10E-10	1.23	7.10E-10	1.23	7.25E-01
Osteoporosis	rs4869744	6	151,908,012	T	C	0.292	1.00	add,dom,gen	add	4.06E-09	1.15	4.06E-09	1.15	2.12E-01
Osteoporosis	rs10242100	7	120,983,343	A	G	0.277	1.00	add,dom	add	2.62E-08	0.87	2.62E-08	0.87	3.07E-01
Osteoporosis	rs56154705	11	68,211,378	C	T	0.143	0.99	add,dom,gen	add	9.09E-10	1.21	9.09E-10	1.21	3.60E-01
Osteoporosis	rs7308105	12	54,424,123	T	C	0.365	0.99	add	add	4.64E-08	1.13	4.64E-08	1.13	7.41E-01
Osteoporosis	rs139959245	22	27,772,054	C	T	0.007	0.90	add	add	4.79E-08	1.91	4.79E-08	1.91	*NA*
Peripheral Vascular Disease	rs6025	1	169,519,049	T	C	0.028	1.00	add,dom,rec,gen,het	add	4.40E-12	0.64	4.40E-12	0.64	4.84E-01
Peripheral Vascular Disease	rs587729126	9	136,138,765	GCGCCCACCACTA	G	0.192	0.99	add,dom,gen	add	4.91E-10	1.20	4.91E-10	1.20	9.44E-01
Peripheral Vascular Disease	rs80274406	11	33,391,655	A	G	0.091	0.92	gen	gen	4.26E-08	0.51/2.26	1.76E-01	1.06	6.32E-06
Peripheral Vascular Disease	rs146399108	11	47,031,734	G	A	0.013	0.96	add	add	4.42E-08	1.67	4.42E-08	1.67	2.87E-01
Peripheral Vascular Disease	rs2932761	19	48,403,215	A	G	0.289	1.00	gen	gen	3.55E-08	0.87/1.27	3.04E-01	0.97	1.35E-08
Psychiatric Disorders	rs12712961	2	46,278,720	T	A	0.452	1.00	add	add	1.66E-08	1.10	1.66E-08	1.10	2.57E-01
Psychiatric Disorders	rs4736253	8	140,354,986	C	T	0.251	1.00	het	het	4.65E-08	0.87	4.60E-04	1.07	2.42E-05
Stress	rs577242570	2	184,407,101	T	G	0.004	0.85	add	add	4.56E-08	2.33	4.56E-08	2.33	*NA*
Type 2 Diabetes	rs1801282	3	12,393,125	C	G	0.120	1.00	dom,het	dom	2.27E-08	0.84	1.05E-07	0.86	7.07E-02
Type 2 Diabetes	rs547194177	3	185,507,515	ATTT	A	0.314	0.99	add,dom,gen	add	3.84E-10	1.13	3.84E-10	1.13	9.52E-01
Type 2 Diabetes	rs77704739	5	52,080,909	T	C	0.036	1.00	rec	rec	1.75E-08	4.32	2.80E-03	1.15	1.92E-07
Type 2 Diabetes	rs3891173	6	32,634,675	A	T	0.287	0.80	add,dom,rec,gen	rec	7.27E-11	1.42	5.25E-08	1.14	5.69E-04
Type 2 Diabetes	rs849133	7	28,192,280	C	T	0.498	1.00	add,rec,gen	add	3.11E-11	0.89	3.11E-11	0.89	1.24E-01
Type 2 Diabetes	rs13266634	8	118,184,783	C	T	0.304	1.00	add,gen	add	8.96E-10	0.88	8.96E-10	0.88	5.86E-01
Type 2 Diabetes	rs10811662	9	22,134,253	G	A	0.176	1.00	add	add	3.92E-08	0.87	3.92E-08	0.87	4.01E-01
Type 2 Diabetes	rs34872471	10	114,754,071	T	C	0.295	0.98	add,dom,rec,gen,het	add	1.03E-52	1.36	1.03E-52	1.36	9.92E-01
Varicose Veins	rs62250779	3	32,652,184	G	A	0.073	0.92	gen	gen	2.13E-08	0.36/3.61	5.60E-03	1.17	9.58E-04
Varicose Veins	rs2383896	8	74,284,818	A	G	0.479	0.99	add	add	5.00E-08	1.17	5.00E-08	1.17	9.88E-01
Varicose Veins	rs117798068	13	88,346,617	T	C	0.011	0.80	add,dom,het	het	8.41E-09	2.07	1.59E-08	2.03	*NA*

CHR = Chromosome, Position = Position Hg19, Alleles A= Non-effect Allele, Allele B= Effect Allele, MAF=Minor Allele Frequency, GW= Genome Wide, OR= Odds Ratio

add= Additive, dom=Dominant, rec= Recessive, het= Heterodominant, gen= Genotypic

Supplementary Table 2. UK Biobank biomarkers associated with new regions

									Additive		Lowest P-value Model	
Phenotype (Cases/Controls)	Nearest Gene	rsID	CHR	position	Alleles	MAF	Lowest P-value Model	Field	beta (CI 95%)	P-value	beta (CI 95%)	P-value
Asthma (9,209/47,428)	*ETF1*	rs154073	5	137,858,067	C/T	0.429	Recessive	IGF-1 (357,380)	0.01 (0.01-0.02)	5.20×10^{-8}	0.01 (0.01-0.02)	1.20×10^{-3}
								Testosterone (325,662)	-0.005 (-0.01--0.003)	1.01×10^{-5}	-0.01 (-0.01--0.01)	4.42×10^{-7}
Cancer (17,131/39,506)	*TEX29*	rs138646839	13	112,115,591	C/T	0.005	Genotypic	Alanine aminotransferase (359,162)	-0.01 (-0.04-0.03)	7.59×10^{-1}	9.62 (8.71-10.54)	1.58×10^{-88}
Cardiovascular (15,009/41,628)	*DCLRE1B*	rs10858023	1	114,448,752	C/T	0.35	Dominant	IGF-1 (357,380)	-0.01 (-0.01--0.003)	2.34×10^{-3}	-0.01 (-0.02--0.01)	3.95×10^{-4}
								Triglycerides (359,009)	-0.01 (-0.01--0.003)	1.72×10^{-3}	-0.01 (-0.02c)	4.95×10^{-4}
Type 2 Diabetes (6,967/49,670)	*PELO*	rs77704739	5	52,080,909	T/C	0.036	Recessive	Apolipoprotein A (327,072)	0.02 (0.01-0.03)	1.36×10^{-3}	-0.18 (-0.25--0.1)	5.54×10^{-6}
								Albumin (329,051)	0.04 (0.02-0.05)	4.81×10^{-9}	0.38 (0.30-0.46)	5.69×10^{-19}
								Apolipoprotein B (357,546)	-0.06 (-0.07--0.05)	2.15×10^{-26}	-0.39 (-0.47--0.31)	1.19×10^{-20}
								C-reactive protein (358,524)	-0.02 (-0.04--0.01)	1.65×10^{-5}	-0.18 (-0.26--0.1)	1.50×10^{-5}
								HDL cholesterol (328,906)	0.02 (0.01-0.03)	3.22×10^{-3}	-0.25 (-0.32--0.17)	4.23×10^{-11}
								LDL direct (358,623)	-0.05 (-0.06--0.04)	1.56×10^{-15}	-0.39 (-0.47--0.31)	1.46×10^{-20}
								Triglycerides (359,009)	0.01 (0.00-0.02)	1.56×10^{-2}	0.36 (0.28-0.44)	3.09×10^{-20}
								Lipoprotein A (286,064)	-0.01 (-0.02-0.004)	2.10×10^{-1}	-0.17 (-0.26--0.07)	4.64×10^{-4}
								Cholesterol (359,307)	-0.03 (-0.05--0.02)	3.88×10^{-9}	-0.37 (-0.45--0.29)	3.85×10^{-19}
Hernia Abdominopelvic (6,291/50,346)	*LOC102723886*	rs2494196	1	219,762,581	C/A	0.274	Additive	Total protein (328,694)	-0.02 (-0.02--0.01)	3.13×10^{-12}	-0.02 (-0.02--0.01)	3.13×10^{-12}
								Apolipoprotein A (327,072)	0.01 (0.01-0.02)	6.98×10^{-10}	0.01 (0.01-0.02)	6.98×10^{-10}
								HDL cholesterol (328,906)	0.02 (0.01-0.02)	3.13×10^{-15}	0.02 (0.01-0.02)	3.13×10^{-15}
								Calcium (328,933)	-0.01 (-0.01--0.003)	1.29×10^{-3}	-0.01 (-0.01--0.003)	1.29×10^{-3}
								Albumin (329,051)	-0.02 (-0.03--0.02)	3.39×10^{-16}	-0.02 (-0.03--0.02)	3.39×10^{-16}
								Glucose (328,681)	-0.01 (-0.02--0.01)	3.32×10^{-5}	-0.01 (-0.02--0.01)	3.32×10^{-5}
								Triglycerides (359,009)	-0.02 (-0.03--0.02)	1.12×10^{-19}	-0.02 (-0.03--0.02)	1.12×10^{-19}
								Urate (358,855)	-0.01 (-0.01--0.004)	2.22×10^{-4}	-0.01 (-0.01--0.004)	2.22×10^{-4}
								SHBG (325,845)	0.02 (0.01-0.02)	2.86×10^{-11}	0.02 (0.01-0.02)	2.86×10^{-11}
								Alanine aminotransferase (359,162)	-0.01 (-0.02--0.01)	1.12×10^{-5}	-0.01 (-0.02--0.01)	1.12×10^{-5}
								Phosphate (328,407)	0.01 (0.00-0.01)	3.78×10^{-4}	0.01 (0.00-0.01)	3.78×10^{-4}
								IGF-1 (357,380)	-0.01 (-0.01--0.005)	1.32×10^{-4}	-0.01 (-0.01--0.005)	1.32×10^{-4}
	STIM2	rs113180595	4	27,019,359	T/C	0.004	Heterodominant	Calcium (328,933)	-0.10 (-0.15--0.05)	2.31×10^{-4}	-0.10 (-0.15--0.05)	2.42×10^{-4}
Hypertension (28,391/28,246)	*LNPK*	rs1446802	2	176,532,019	A/G	0.5	Recessive	SHBG (325,845)	-0.01 (-0.01--0.01)	1.33×10^{-5}	-0.01 (-0.02--0.01)	2.38×10^{-5}
	HIC1	rs112963849	17	1,959,826	C/A	0.082	Additive	Calcium (328,933)	-0.01 (-0.02--0.01)	1.15×10^{-3}	-0.01 (-0.02--0.01)	1.15×10^{-3}
								Total protein (328,694)	-0.02 (-0.03--0.01)	2.19×10^{-6}	-0.02 (-0.03--0.01)	2.19×10^{-6}
								Triglycerides (359,009)	-0.01 (-0.02--0.01)	1.81×10^{-4}	-0.01 (-0.02--0.01)	1.81×10^{-4}
Macular Degeneration (3,685/52,952)	*THUMPD2*	rs557998486	2	40,010,523	T/TG	0.009	Recessive	C-reactive protein (358,524)	-0.01 (-0.03-0.02)	6.63×10^{-1}	1.11 (0.70-1.53)	1.15×10^{-4}
Psychiatric (8,624/48,013)	*PRKCE*	rs12712961	2	46,278,720	T/A	0.452	Additive	C-reactive protein (358,524)	0.01 (0.00-0.01)	1.62×10^{-3}	0.01 (0.00-0.01)	1.62×10^{-3}
Peripheral Vascular Disease (4,301/52,336)	*SNAR-A12*	rs2932761	19	48,403,215	A/G	0.289	Genotypic	IGF-1 (357,380)	0.01 (0.01-0.01)	4.82×10^{-5}	0.01 (0.00-0.02)	2.58×10^{-4}
								Vitamin D (343,604)	0.03 (0.03-0.04)	2.09×10^{-40}	0.03 (0.03-0.04)	2.63 × 10-^{39}

CHR = Chromosome, Position = Position Hg19, Alleles = Non-effect Allele / Effect Allele, MAF=Minor Allele Frequency, CI= Confidence Interval

Supplementary Table 3. Colocalization between GWAS results and eQTL loci

							GTEx			
rsID	CHR	Position	MAF	Phenotype (cases/controls)	Alleles	Nearest gene	eQTL Gene	p-value	Tissue	Colocalization probability*
rs2399472	3	112,911,615	0.073	*ALLERGIC RHINITIS* (13,936/42,701)	C/T	*LINC02044*	*CCDC80*	4.10E-05	Heart - Left Ventricle	0.07
rs10112506	8	13,164,746	0.39	*ALLERGIC RHINITIS* (13,936/42,701)	A/G	*DLC1*	*DLC1*	5.00E-08	Testis	0.89
rs154073	5	137,858,067	0.429	*ASTHMA* (9,209/47,428)	C/T	*ETF1*	*ETF1*	2.30E-07	Esophagus - Mucosa	0.97
rs10858023	1	114,448,752	0.35	*CARDIOVASCULAR DISORDERS* (15,009/41,628)	C/T	*DCLRE1B*	*AP4B1*	1.80E-11	Muscle - Skeletal	0.98
							PTPN22	1.60E-09	Colon - Transverse	0.71
								1.60E-09	Pancreas	0.93
rs77704739	5	52,080,909	0.036	*TYPE 2 DIABETES* (6,967/49,670)	T/C	*ITGA1 / PELO*	*PELO*	9.10E-20	Whole Blood	0.98
								9.10E-19	Cells - Transformed fibroblasts	0.98
								2.80E-18	Skin - Not Sun Exposed (Suprapubic)	0.98
								3.30E-17	Skin - Sun Exposed (Lower leg)	0.98
								9.50E-11	Esophagus - Mucosa	0.98
								2.60E-10	Nerve - Tibial	0.94
								1.20E-07	Adipose - Subcutaneous	0.91
								1.00E-06	Pancreas	0.50
								1.20E-06	Stomach	0.54
								2.80E-06	Cells - EBV-transformed lymphocytes	*NA*
								4.20E-06	Muscle - Skeletal	0.83
								5.00E-06	Artery - Aorta	0.27
rs2494196	1	219,762,581	0.274	*HERNIA ABDOMINOPELVIC* (6,291/50,346)	C/A	*LOC102723886*	RP11-392O17.1	2.60E-06	Skin - Not Sun Exposed (Suprapubic)	3.05E-03
								5.00E-05	Skin - Sun Exposed (Lower leg)	0.06
rs28792763	15	90,081,905	0.462	*HYPERTENSION* (28,391/28,246)	G/A	*LINC00928*	*TICRR*	1.20E-08	Cells - Transformed fibroblasts	0.04
								4.10E-08	Thyroid	0.04
								5.50E-08	Artery - Aorta	0.81
								6.10E-06	Nerve - Tibial	1.58E-03
								7.80E-06	Artery - Tibial	0.11
rs112963849	17	1,959,826	0.082	*HYPERTENSION* (28,391/28,246)	C/A	*HIC1*	*DPH1*	6.30E-08	Cells - Transformed fibroblasts	0.0000563
								1.40E-07	Adipose - Subcutaneous	0.0000553
								3.10E-07	Nerve - Tibial	0.0000505
								1.80E-06	Esophagus - Mucosa	0.0000504
								6.00E-06	Adipose - Visceral (Omentum)	0.32
							SRR	1.90E-05	Artery - Aorta	1.15E-03
rs12712961	2	46,278,720	0.452	*PSYCHIATRIC DISORDERS* (8,624/48,013)	T/A	*PRKCE*	*EPAS1*	1.50E-05	Cells - Transformed fibroblasts	0.57
rs80274406	11	33,391,655	0.091	*PERIPHERAL VASCULAR DISEASE* (4,301/52,336)	A/G	*HIPK3*	*KIAA1549L*	5.60E-06	Cells - Transformed fibroblasts	0.05
rs2932761	19	48,403,215	0.289	*PERIPHERAL VASCULAR DISEASE* (4,301/52,336)	A/G	*SNAR-A12*	*CTD-3098H1.2*	6.60E-08	Adrenal Gland	NA
								1.40E-06	Liver	NA
							SULT2A1	2.90E-06	Adrenal Gland	NA
							CTD-3098H1.2	8.00E-06	Small Intestine - Terminal Ileum	NA

CHR= chromosome; Position= genomic position hg19, MAF= minor allele frequency, Alleles= Non-effect allele / Effect allele

*Colocalization probability corresponds to the Bayesian posterior probability of colocalization between the GWAS and the eQTL loci

Supplementary Table 4. Cross-phenotype associations results

Locus	CHR	Variant ID	Disease A vs Disease B	Lowest P-value Model	Disease A		Disease B	
					OR (CI 95%)	p-value	OR (CI 95%)	p-value
ABO	9	chr9:136138765:D	Peripheral Vascular Disease-Dyslipidemia	Additive	1.20 (1.13-1.27)	4.91×10^{-10}	1.16 (1.13-1.2)	1.09×10^{-20}
		9:136151806_T_C	Dyslipidemia-Peripheral Vascular Disease	Dominant	1.16 (1.12-1.21)	2.79×10^{-16}	1.18 (1.11-1.27)	4.46×10^{-7}
		9:136184798_C_T	Dyslipidemia-Osteoarthritis	Heterodominant	1.16 (1.11-1.21)	1.66×10^{-10}	1.11 (1.06-1.16)	2.34×10^{-5}
CETP	16	rs3764261	Dyslipidemia-Macular Degeneration	Additive	0.91 (0.88-0.93)	2.85×10^{-12}	1.12 (1.07-1.19)	1.94×10^{-5}
ETF1	5	chr5:137858067	Asthma-Irritable Bowel	Recessive	1.18 (1.12-1.25)	4.23×10^{-9}	1.22 (1.12-1.34)	1.30×10^{-5}
HLA	6	6:31930441_G_T	Macular Degeneration-Dyslipidemia	Additive	1.18 (1.12-1.25)	4.69×10^{-9}	1.07 (1.04-1.1)	1.39×10^{-6}
		rs33941037	Cancer-Dyslipidemia	Additive	1.13 (1.08-1.17)	1.39×10^{-8}	1.11 (1.07-1.15)	1.81×10^{-7}
		rs62406303	Dyslipidemia-Asthma	Additive	1.12 (1.08-1.15)	1.46×10^{-10}	0.91 (0.87-0.95)	8.04×10^{-6}
		rs9274639:32636146:T:C	Asthma-Macular Degeneration	Additive	1.11 (1.07-1.15)	1.20×10^{-8}	1.14 (1.08-1.2)	6.98×10^{-6}
		rs55651120	Dyslipidemia-Cancer	Dominant	1.18 (1.13-1.24)	5.30×10^{-12}	1.15 (1.09-1.21)	5.60×10^{-8}
		rs9272266:32603416:G:A	Asthma-Dyslipidemia	Dominant	0.86 (0.82-0.9)	2.55×10^{-10}	1.10 (1.06-1.14)	8.66×10^{-8}
		rs75351515:32635420:G:A	Asthma-Type 2 Diabetes	Genotypic	1.13 (1.09-1.17)	6.64×10^{-13}	1.08 (1.04-1.12)	7.00×10^{-6}
		rs3210176:32627850:T:C	Type 2 Diabetes-Asthma	Genotypic	1.15 (1.10-1.29)	3.89×10^{-9}	1.10 (1.06-1.15)	2.20×10^{-6}
JAZF1	7	7:28154822_A_G	Type 2 Diabetes-Asthma	Genotypic	0.89 (0.86-0.93)	4.36×10^{-8}	0.92 (0.89-0.95)	6.87×10^{-6}
NECTIN2 / PVRL2	19	19:45388130_G_A	Dyslipidemia-Type 2 Diabetes	Additive	1.23 (1.19-1.28)	8.99×10^{-30}	0.89 (0.84-0.94)	1.03×10^{-5}
		19:45428234_G_A	Dyslipidemia-Cardiovascular	Additive	1.26 (1.21-1.31)	4.55×10^{-3}	1.11 (1.06-1.16)	1.44×10^{-5}
TRIB1	8	8:126507389_C_A	Dyslipidemia-Hypertension	Additive	0.88 (0.86-0.91)	2.53×10^{-19}	0.94 (0.91-0.97)	2.40×10^{-5}
WDR36	5	rs252716:110425063:G:C	Asthma-Allergic Rhinitis	Additive	1.10 (1.07-1.14)	3.53×10^{-9}	1.07 (1.04-1.1)	2.79×10^{-6}
		rs367937013:110424583:C:T	Allergic Rhinitis-Asthma	Additive	0.92 (0.89-0.94)	5.62×10^{-10}	0.92 (0.89-0.95)	2.19×10^{-7}

CHR = Chromosome

References

1000 Genomes Project Consortium, Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., and McVean, G.A. (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061-1073.

1000 Genomes Project Consortium, Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56-65.

1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A.*, et al.* (2015). A global reference for human genetic variation. Nature *526*, 68-74.

Alder, J.K., and Kass, D.J. (2017). Another building in the IPF Manhattan plot skyline. Lancet Respir Med *5*, 837-839.

Alioto, T.S., Buchhalter, I., Derdak, S., Hutter, B., Eldridge, M.D., Hovig, E., Heisler, L.E., Beck, T.A., Simpson, J.T., Tonon, L*., et al.* (2015). A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. Nat Commun *6*, 10001.

Altmuller, J., Palmer, L.J., Fischer, G., Scherb, H., and Wjst, M. (2001). Genomewide scans of complex human diseases: true linkage is hard to find. Am J Hum Genet *69*, 936-950.

Altshuler, D., Daly, M.J., and Lander, E.S. (2008). Genetic mapping in human disease. Science *322*, 881-888.

Amos, C.I., Dennis, J., Wang, Z., Byun, J., Schumacher, F.R., Gayther, S.A., Casey, G., Hunter, D.J., Sellers, T.A., Gruber, S.B*., et al.* (2017). The OncoArray Consortium: A Network for Understanding the Genetic Architecture of Common Cancers. Cancer Epidemiol Biomarkers Prev *26*, 126-135.

Anderson, C.A., Pettersson, F.H., Clarke, G.M., Cardon, L.R., Morris, A.P., and Zondervan, K.T. (2010). Data quality control in genetic case-control association studies. Nat Protoc *5*, 1564-1573.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T*., et al.* (2014). An atlas of active enhancers across human cell types and tissues. Nature *507*, 455-461.

Assimes, T.L., and Roberts, R. (2016). Genetics: Implications for Prevention and Management of Coronary Artery Disease. J Am Coll Cardiol *68*, 2797-2818.

Avery, O.T., Macleod, C.M., and 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. J Exp Med *79*, 137-158.

Badouin, H., Gouzy, J., Grassa, C.J., Murat, F., Staton, S.E., Cottret, L., Lelandais-Briere, C., Owens, G.L., Carrere, S., Mayjonade, B*., et al.* (2017). The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. Nature *546*, 148-152.

Bahcall, O.G. (2018). UK Biobank - a new era in genomic medicine. Nat Rev Genet *19*, 737.

Baichoo, S., Souilmi, Y., Panji, S., Botha, G., Meintjes, A., Hazelhurst, S., Bendou, H., Beste, E., Mpangase, P.T., Souiai, O*., et al.* (2018). Developing reproducible bioinformatics analysis workflows for heterogeneous computing environments to support African genomics. BMC Bioinformatics *19*, 457.

Bak, S., Gaist, D., Sindrup, S.H., Skytthe, A., and Christensen, K. (2002). Genetic liability in stroke: a long-term follow-up study of Danish twins. Stroke *33*, 769-774.

Balding, D.J. (2006). A tutorial on statistical methods for population association studies. Nat Rev Genet *7*, 781-791.

Bamshad, M., Lin, R.C., Law, D.J., Watkins, W.C., Krakowiak, P.A., Moore, M.E., Franceschini, P., Lala, R., Holmes, L.B., Gebuhr, T.C.*, et al.* (1997). Mutations in human TBX3 alter limb, apocrine and genital development in ulnar-mammary syndrome. Nat Genet *16*, 311-315.

Bateson, P. (2002). William Bateson: a biologist ahead of his time. J Genet *81*, 49-58.

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.

Biondi, B., and Klein, I. (2004). Hypothyroidism as a risk factor for cardiovascular disease. Endocrine *24*, 1-13.

Bodmer, W., and Bonilla, C. (2008). Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet *40*, 695-701.

Bonas-Guarch, S., Guindo-Martinez, M., Miguel-Escalada, I., Grarup, N., Sebastian, D., Rodriguez-Fos, E., Sanchez, F., Planas-Felix, M., Cortes-Sanchez, P., Gonzalez, 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.

Bonnelykke, K., Sleiman, P., Nielsen, K., Kreiner-Moller, E., Mercader, J.M., Belgrave, D., den Dekker, H.T., Husby, A., Sevelsted, A., Faura-Tellez, G.*, et al.* (2014). A genome-wide association study identifies CDHR3 as a susceptibility locus for early childhood asthma with severe exacerbations. Nat Genet *46*, 51-55.

Botstein, D., and Risch, N. (2003). Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat Genet *33 Suppl*, 228-237.

Botstein, D., White, R.L., Skolnick, M., and Davis, R.W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet *32*, 314-331.

Browning, B.L., Zhou, Y., and Browning, S.R. (2018). A One-Penny Imputed Genome from Next-Generation Reference Panels. Am J Hum Genet *103*, 338-348.

Bulik-Sullivan, B., Finucane, H.K., Anttila, V., Gusev, A., Day, F.R., Loh, P.R., ReproGen, C., Psychiatric Genomics, C., Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control, C., Duncan, L.*, et al.* (2015a). An atlas of genetic correlations across human diseases and traits. Nat Genet *47*, 1236-1241.

Bulik-Sullivan, B.K., Loh, P.R., Finucane, H.K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics, C., Patterson, N., Daly, M.J., Price, A.L., and Neale, B.M. (2015b). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet *47*, 291-295.

Busch, M., Kruger, A., Gross, S., Ittermann, T., Friedrich, N., Nauck, M., Dorr, M., and Felix, S.B. (2019). Relation of IGF-1 and IGFBP-3 with prevalent and incident atrial fibrillation in a population-based study. Heart Rhythm *16*, 1314-1319.

Bush, W.S., and Moore, J.H. (2012). Chapter 11: Genome-wide association studies. PLoS Comput Biol *8*, e1002822.

Bustamante, C.D., Burchard, E.G., and De la Vega, F.M. (2011). Genomics for the world. Nature *475*, 163-165.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J.*, et al.* (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203-209.

Campbell, C.D., Ogburn, E.L., Lunetta, K.L., Lyon, H.N., Freedman, M.L., Groop, L.C., Altshuler, D., Ardlie, K.G., and Hirschhorn, J.N. (2005). Demonstrating stratification in a European American population. Nat Genet *37*, 868-872.

Carrel, L., and Willard, H.F. (2005). X-inactivation profile reveals extensive variability in X-linked gene expression in females. Nature *434*, 400-404.

Cephus, J.Y., Stier, M.T., Fuseini, H., Yung, J.A., Toki, S., Bloodworth, M.H., Zhou, W., Goleniewska, K., Zhang, J., Garon, S.L.*, et al.* (2017). Testosterone Attenuates Group 2 Innate Lymphoid Cell-Mediated Airway Inflammation. Cell Rep *21*, 2487-2499.

Chanock, S.J., Manolio, T., Boehnke, M., Boerwinkle, E., Hunter, D.J., Thomas, G., Hirschhorn, J.N., Abecasis, G., Altshuler, D., Bailey-Wilson, J.E.*, et al.* (2007). Replicating genotype-phenotype associations. Nature *447*, 655-660.

Charchar, F.J., Bloomer, L.D., Barnes, T.A., Cowley, M.J., Nelson, C.P., Wang, Y., Denniff, M., Debiec, R., Christofidou, P., Nankervis, S.*, et al.* (2012). Inheritance of coronary artery disease in men: an analysis of the role of the Y chromosome. Lancet *379*, 915-922.

Chargaff, E. (1950). Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. Experientia *6*, 201-209.

Charlesworth, B., and Charlesworth, D. (2009). Darwin and genetics. Genetics *183*, 757-766.

Chen, Z., Chen, J., Collins, R., Guo, Y., Peto, R., Wu, F., Li, L., and China Kadoorie Biobank collaborative, g. (2011). China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. Int J Epidemiol *40*, 1652-1666.

Christophersen, I.E., Rienstra, M., Roselli, C., Yin, X., Geelhoed, B., Barnard, J., Lin, H., Arking, D.E., Smith, A.V., Albert, C.M.*, et al.* (2017). Large-scale analyses of common and rare variants identify 12 new loci associated with atrial fibrillation. Nat Genet *49*, 946-952.

Cirulli, E.T., and Goldstein, D.B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet *11*, 415-425.

Cummings, S.R., and Melton, L.J. (2002). Epidemiology and outcomes of osteoporotic fractures. Lancet *359*, 1761-1767.

Cutler, D., Lleras-Muney, A. (2007). Policy brief: education and health. Ann Arbor, MI, University of Michigan.

Czene, K., Lichtenstein, P., and Hemminki, K. (2002). Environmental and heritable causes of cancer among 9.6 million individuals in the Swedish Family-Cancer Database. Int J Cancer *99*, 260-266.

D. Cutting, M.C. (2006). Apache hadoop: an open-source software for reliable, scalable, distributed computing. .

Dahm, R. (2005). Friedrich Miescher and the discovery of DNA. Dev Biol *278*, 274-288.

Dai, J., Huang, M., Amos, C.I., Hung, R.J., Tardon, A., Andrew, A., Chen, C., Christiani, D.C., Albanes, D., Rennert, G.*, et al.* (2019). Genome-wide association study of INDELs identified four novel susceptibility loci associated with lung cancer risk. Int J Cancer.

Das, S., Abecasis, G.R., and Browning, B.L. (2018). Genotype Imputation from Large Reference Panels. Annu Rev Genomics Hum Genet *19*, 73-96.

Das, S., Forer, L., Schonherr, 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.

de Bakker, P.I., Ferreira, M.A., Jia, X., Neale, B.M., Raychaudhuri, S., and Voight, B.F. (2008). Practical aspects of imputation-driven meta-analysis of genome-wide association studies. Hum Mol Genet *17*, R122-128.

Delaneau, O., Howie, B., Cox, A.J., Zagury, J.F., and Marchini, J. (2013a). Haplotype estimation using sequencing reads. Am J Hum Genet *93*, 687-696.

Delaneau, O., Marchini, J., Genomes Project, C., and Genomes Project, C. (2014). Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. Nat Commun *5*, 3934.

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

Delaneau, O., Zagury, J.-F., Robinson, M., Marchini, J., and Dermitzakis, E. (2018). Integrative haplotype estimation with sub-linear complexity. bioRxiv, 493403.

Delaneau, O., Zagury, J.F., and Marchini, J. (2013b). Improved whole-chromosome phasing for disease and population genetic studies. Nat Methods *10*, 5-6.

Donahue, R.P., Bias, W.B., Renwick, J.H., and McKusick, V.A. (1968). Probable assignment of the Duffy blood group locus to chromosome 1 in man. Proc Natl Acad Sci U S A *61*, 949-955.

Duncan, L.E., Ratanatharathorn, A., Aiello, A.E., Almli, L.M., Amstadter, A.B., Ashley-Koch, A.E., Baker, D.G., Beckham, J.C., Bierut, L.J., Bisson, J.*, et al.* (2018). Largest GWAS of PTSD (N=20 070) yields genetic overlap with schizophrenia and sex differences in heritability. Mol Psychiatry *23*, 666-673.

Duntas, L.H., and Brenta, G. (2018). A Renewed Focus on the Association Between Thyroid Hormones and Lipid Metabolism. Front Endocrinol (Lausanne) *9*, 511.

Editorial, N. (2017). Gender imbalance in science journals is still pervasive. Nature *541*, 435-436.

Ellinor, P.T., Lunetta, K.L., Glazer, N.L., Pfeufer, A., Alonso, A., Chung, M.K., Sinner, M.F., de Bakker, P.I., Mueller, M., Lubitz, S.A.*, et al.* (2010). Common variants in KCNN3 are associated with lone atrial fibrillation. Nat Genet *42*, 240-244.

Elliott, L.T., Sharp, K., Alfaro-Almagro, F., Shi, S., Miller, K.L., Douaud, G., Marchini, J., and Smith, S.M. (2018). Genome-wide association studies of brain imaging phenotypes in UK Biobank. Nature *562*, 210-216.

Encode Project Consortium (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. Science *306*, 636-640.

Encode Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57-74.

Encode Project Consortium, Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigo, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T.*, et al.* (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature *447*, 799-816.

Estrada, K., Abuseiris, A., Grosveld, F.G., Uitterlinden, A.G., Knoch, T.A., and Rivadeneira, F. (2009). GRIMP: a web- and grid-based tool for high-speed analysis of large-scale genome-wide association using imputed data. Bioinformatics *25*, 2750-2752.

Estrada, K., Aukrust, I., Bjorkhaug, L., Burtt, N.P., Mercader, J.M., Garcia-Ortiz, H., Huerta-Chagoya, A., Moreno-Macias, H., Walford, G., Flannick, J.*, et al.* (2014). Association of a low-frequency variant in HNF1A with type 2 diabetes in a Latino population. JAMA *311*, 2305-2314.

Evangelou, E., and Ioannidis, J.P. (2013). Meta-analysis methods for genome-wide association studies and beyond. Nat Rev Genet *14*, 379-389.

Fadista, J., Manning, A.K., Florez, J.C., and Groop, L. (2016). The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. Eur J Hum Genet *24*, 1202-1205.

Falconer, D.S., and Mackay, T.F.C. (1996). Introduction to quantitative genetics, 4th ed edn (Harlow: Longman).

Ferreira, M.A., Matheson, M.C., Tang, C.S., Granell, R., Ang, W., Hui, J., Kiefer, A.K., Duffy, D.L., Baltic, S., Danoy, P.*, et al.* (2014). Genome-wide association analysis identifies 11 risk variants associated with the asthma with hay fever phenotype. J Allergy Clin Immunol *133*, 1564-1571.

Ferreira, M.A., Vonk, J.M., Baurecht, H., Marenholz, I., Tian, C., Hoffman, J.D., Helmer, Q., Tillander, A., Ullemar, V., van Dongen, J.*, et al.* (2017). Shared genetic origin of asthma, hay fever and eczema elucidates allergic disease biology. Nat Genet *49*, 1752-1757.

Ferreira, M.A.R., Mathur, R., Vonk, J.M., Szwajda, A., Brumpton, B., Granell, R., Brew, B.K., Ullemar, V., Lu, Y., Jiang, Y.*, et al.* (2019). Genetic Architectures of Childhood- and Adult-Onset Asthma Are Partly Distinct. Am J Hum Genet *104*, 665-684.

Feuk, L., Carson, A.R., and Scherer, S.W. (2006). Structural variation in the human genome. Nat Rev Genet *7*, 85-97.

Fischer, M., Broeckel, U., Holmer, S., Baessler, A., Hengstenberg, C., Mayer, B., Erdmann, J., Klein, G., Riegger, G., Jacob, H.J.*, et al.* (2005). Distinct heritable patterns of angiographic coronary artery disease in families with myocardial infarction. Circulation *111*, 855-862.

Fisher, R.A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. Trans R Soc Edinb *52*, 399-433.

Fishilevich, S., Nudel, R., Rappaport, N., Hadar, R., Plaschkes, I., Iny Stein, T., Rosen, N., Kohn, A., Twik, M., Safran, M.*, et al.* (2017). GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database (Oxford) *2017*.

Fry, A., Littlejohns, T.J., Sudlow, C., Doherty, N., Adamska, L., Sprosen, T., Collins, R., and Allen, N.E. (2017). Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. Am J Epidemiol *186*, 1026-1034.

Gao, F., Chang, D., Biddanda, A., Ma, L., Guo, Y., Zhou, Z., and Keinan, A. (2015). XWAS: A Software Toolset for Genetic Data Analysis and Association Studies of the X Chromosome. J Hered *106*, 666-671.

Gatz, M., Reynolds, C.A., Fratiglioni, L., Johansson, B., Mortimer, J.A., Berg, S., Fiske, A., and Pedersen, N.L. (2006). Role of genes and environments for explaining Alzheimer disease. Arch Gen Psychiatry *63*, 168-174.

Gauthier, J., Vincent, A.T., Charette, S.J., and Derome, N. (2018). A brief history of bioinformatics. Brief Bioinform.

Gaziano, J.M., Concato, J., Brophy, M., Fiore, L., Pyarajan, S., Breeling, J., Whitbourne, S., Deen, J., Shannon, C., Humphries, D*., et al.* (2016). Million Veteran Program: A mega-biobank to study genetic influences on health and disease. J Clin Epidemiol *70*, 214-223.

Ge, T., Chen, C.Y., Neale, B.M., Sabuncu, M.R., and Smoller, J.W. (2017). Phenome-wide heritability analysis of the UK Biobank. PLoS Genet *13*, e1006711.

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.

Gibson, G. (2010). Hints of hidden heritability in GWAS. Nat Genet *42*, 558-560.

Gibson, G. (2012). Rare and common variants: twenty arguments. Nat Rev Genet *13*, 135-145.

Goodwin, S., McPherson, J.D., and McCombie, W.R. (2016). Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet *17*, 333-351.

Goyette, P., Boucher, G., Mallon, D., Ellinghaus, E., Jostins, L., Huang, H., Ripke, S., Gusareva, E.S., Annese, V., Hauser, S.L*., et al.* (2015). High-density mapping of the MHC identifies a shared role for HLA-DRB1*01:03 in inflammatory bowel diseases and heterozygous advantage in ulcerative colitis. Nat Genet *47*, 172-179.

Grarup, N., Moltke, I., Andersen, M.K., Bjerregaard, P., Larsen, C.V.L., Dahl-Petersen, I.K., Jorsboe, E., Tiwari, H.K., Hopkins, S.E., Wiener, H.W*., et al.* (2018). Identification of novel high-impact recessively inherited type 2 diabetes risk variants in the Greenlandic population. Diabetologia *61*, 2005-2015.

Gravel, S., Henn, B.M., Gutenkunst, R.N., Indap, A.R., Marth, G.T., Clark, A.G., Yu, F., Gibbs, R.A., Genomes, P., and Bustamante, C.D. (2011). Demographic history and rare allele sharing among human populations. Proc Natl Acad Sci U S A *108*, 11983-11988.

Griffiths, A.J.F. (2000). An introduction to genetic analysis, 7th edn (New York: Freeman).

GTEx Consortium (2013). The Genotype-Tissue Expression (GTEx) project. Nat Genet *45*, 580-585.

GTEx Consortium, Laboratory, D.A., Coordinating Center -Analysis Working, G., Statistical Methods groups-Analysis Working, G., Enhancing, G.g., Fund, N.I.H.C., Nih/Nci, Nih/Nhgri, Nih/Nimh, Nih/Nida*, et al.* (2017). Genetic effects on gene expression across human tissues. Nature *550*, 204-213.

Gudmundsson, J., Sulem, P., Gudbjartsson, D.F., Masson, G., Agnarsson, B.A., Benediktsdottir, K.R., Sigurdsson, A., Magnusson, O.T., Gudjonsson, S.A., Magnusdottir, D.N*., et al.* (2012). A study based on whole-genome sequencing yields a rare variant at 8q24 associated with prostate cancer. Nat Genet *44*, 1326-1329.

Guglielmi, G. (2019a). Eastern European universities score highly in university gender ranking. Nature.

Guglielmi, G. (2019b). Facing up to injustice in genome science. Nature *568*, 290-293.

Gusella, J.F., Wexler, N.S., Conneally, P.M., Naylor, S.L., Anderson, M.A., Tanzi, R.E., Watkins, P.C., Ottina, K., Wallace, M.R., Sakaguchi, A.Y*., et al.* (1983).

A polymorphic DNA marker genetically linked to Huntington's disease. Nature *306*, 234-238.

GWAS Catalog, GWAS Catalog, https://www.ebi.ac.uk/gwas/, July 13th, 2019

Hail Team, https://github.com/hail-is/hail/releases/tag/0.2.13., 0.2.13-81ab564db2b4, 2016

Hamburg, M.A., and Collins, F.S. (2010). The path to personalized medicine. N Engl J Med *363*, 301-304.

Han, Y., Rand, K.A., Hazelett, D.J., Ingles, S.A., Kittles, R.A., Strom, S.S., Rybicki, B.A., Nemesure, B., Isaacs, W.B., Stanford, J.L.*, et al.* (2016). Prostate Cancer Susceptibility in Men of African Ancestry at 8q24. J Natl Cancer Inst *108*.

Hargittai, I. (2009). The tetranucleotide hypothesis: a centennial. Struct Chem *20*, 753-756.

Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S.*, et al.* (2012). GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res *22*, 1760-1774.

Hershey, A.D., and Chase, M. (1952). Independent functions of viral protein and nucleic acid in growth of bacteriophage. J Gen Physiol *36*, 39-56.

Hewitt, J., Walters, M., Padmanabhan, S., and Dawson, J. (2016). Cohort profile of the UK Biobank: diagnosis and characteristics of cerebrovascular disease. BMJ Open *6*, e009161.

Hinds, D.A., Stuve, L.L., Nilsen, G.B., Halperin, E., Eskin, E., Ballinger, D.G., Frazer, K.A., and Cox, D.R. (2005). Whole-genome patterns of common DNA variation in three human populations. Science *307*, 1072-1079.

Hinrichs, A.L., Larkin, E.K., and Suarez, B.K. (2009). Population stratification and patterns of linkage disequilibrium. Genet Epidemiol *33 Suppl 1*, S88-92.

Hirschhorn, J.N., and Daly, M.J. (2005). Genome-wide association studies for common diseases and complex traits. Nat Rev Genet *6*, 95-108.

Hodgkin, J. (1998). Seven types of pleiotropy. Int J Dev Biol *42*, 501-505.

Hoffmann, T.J., Ehret, G.B., Nandakumar, P., Ranatunga, D., Schaefer, C., Kwok, P.Y., Iribarren, C., Chakravarti, A., and Risch, N. (2017). Genome-wide association analyses using electronic health records identify new loci influencing blood pressure variation. Nat Genet *49*, 54-64.

Hoffmann, T.J., Kvale, M.N., Hesselson, S.E., Zhan, Y., Aquino, C., Cao, Y., Cawley, S., Chung, E., Connell, S., Eshragh, J.*, et al.* (2011a). Next generation genome-wide association tool: design and coverage of a high-throughput European-optimized SNP array. Genomics *98*, 79-89.

Hoffmann, T.J., Zhan, Y., Kvale, M.N., Hesselson, S.E., Gollub, J., Iribarren, C., Lu, Y., Mei, G., Purdy, M.M., Quesenberry, C.*, et al.* (2011b). Design and coverage of high throughput genotyping arrays optimized for individuals of East Asian, African American, and Latino race/ethnicity using imputation and a novel hybrid SNP selection algorithm. Genomics *98*, 422-430.

Hood, L., and Rowen, L. (2013). The Human Genome Project: big science transforms biology and medicine. Genome Med *5*, 79.

Hormozdiari, F., Kostem, E., Kang, E.Y., Pasaniuc, B., and Eskin, E. (2014). Identifying causal variants at loci with multiple signals of association. Genetics *198*, 497-508.

Hormozdiari, F., van de Bunt, M., Segre, A.V., Li, X., Joo, J.W.J., Bilow, M., Sul, J.H., Sankararaman, S., Pasaniuc, B., and Eskin, E. (2016). Colocalization of

GWAS and eQTL Signals Detects Target Genes. Am J Hum Genet *99*, 1245-1260.

Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., and Abecasis, G.R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet *44*, 955-959.

Howie, B., Marchini, J., and Stephens, M. (2011). Genotype imputation with thousands of genomes. G3 (Bethesda) *1*, 457-470.

Howie, B.N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet *5*, e1000529.

Huang, J., Howie, B., McCarthy, S., Memari, Y., Walter, K., Min, J.L., Danecek, P., Malerba, G., Trabetti, E., Zheng, H.F.*, et al.* (2015). Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. Nat Commun *6*, 8111.

Huang, L., Li, Y., Singleton, A.B., Hardy, J.A., Abecasis, G., Rosenberg, N.A., and Scheet, P. (2009). Genotype-imputation accuracy across worldwide human populations. Am J Hum Genet *84*, 235-250.

Ingebrigtsen, T., Thomsen, S.F., Vestbo, J., van der Sluis, S., Kyvik, K.O., Silverman, E.K., Svartengren, M., and Backer, V. (2010). Genetic influences on Chronic Obstructive Pulmonary Disease - a twin study. Respir Med *104*, 1890-1895.

International HapMap, C., Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P.*, et al.* (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851-861.

International HapMap Consortium (2003). The International HapMap Project. Nature *426*, 789-796.

International HapMap Consortium (2005). A haplotype map of the human genome. Nature *437*, 1299-1320.

Jha, V., Garcia-Garcia, G., Iseki, K., Li, Z., Naicker, S., Plattner, B., Saran, R., Wang, A.Y., and Yang, C.W. (2013). Chronic kidney disease: global dimension and perspectives. Lancet *382*, 260-272.

Kantarci, O.H., Barcellos, L.F., Atkinson, E.J., Ramsay, P.P., Lincoln, R., Achenbach, S.J., De Andrade, M., Hauser, S.L., and Weinshenker, B.G. (2006). Men transmit MS more often to their children vs women: the Carter effect. Neurology *67*, 305-310.

Karlsson Linner, 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.* (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.

Katzmarzyk, P.T., Perusse, L., Rice, T., Gagnon, J., Skinner, J.S., Wilmore, J.H., Leon, A.S., Rao, D.C., and Bouchard, C. (2000). Familial resemblance for coronary heart disease risk: the HERITAGE Family Study. Ethn Dis *10*, 138-147.

Kenny, E.E., Timpson, N.J., Sikora, M., Yee, M.C., Moreno-Estrada, A., Eng, C., Huntsman, S., Burchard, E.G., Stoneking, M., Bustamante, C.D.*, et al.* (2012). Melanesian blond hair is caused by an amino acid change in TYRP1. Science *336*, 554.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res *12*, 996-1006.

Kerem, B., Rommens, J.M., Buchanan, J.A., Markiewicz, D., Cox, T.K., Chakravarti, A., Buchwald, M., and Tsui, L.C. (1989). Identification of the cystic fibrosis gene: genetic analysis. Science *245*, 1073-1080.

Kerminen, S., Martin, A.R., Koskela, J., Ruotsalainen, S.E., Havulinna, A.S., Surakka, I., Palotie, A., Perola, M., Salomaa, V., Daly, M.J.*, et al.* (2019). Geographic Variation and Bias in the Polygenic Scores of Complex Diseases and Traits in Finland. Am J Hum Genet *104*, 1169-1181.

Khramtsova, E.A., Davis, L.K., and Stranger, B.E. (2019). The role of sex in the genomics of human complex traits. Nat Rev Genet *20*, 173-190.

King, R.A., Rotter, J.I., and Motulsky, A.G. (1992). The genetic basis of common diseases (New York ; Oxford: Oxford University Press).

Klaver, C.C., Wolfs, R.C., Assink, J.J., van Duijn, C.M., Hofman, A., and de Jong, P.T. (1998). Genetic risk of age-related maculopathy. Population-based familial aggregation study. Arch Ophthalmol *116*, 1646-1651.

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.* (2005). Complement factor H polymorphism in age-related macular degeneration. Science *308*, 385-389.

Knowler, W.C., Barrett-Connor, E., Fowler, S.E., Hamman, R.F., Lachin, J.M., Walker, E.A., Nathan, D.M., and Diabetes Prevention Program Research, G. (2002). Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. N Engl J Med *346*, 393-403.

Knowlton, R.G., Cohen-Haguenauer, O., Van Cong, N., Frezal, J., Brown, V.A., Barker, D., Braman, J.C., Schumm, J.W., Tsui, L.C., Buchwald, M.*, et al.* (1985). A polymorphic DNA marker linked to cystic fibrosis is located on chromosome 7. Nature *318*, 380-382.

Konig, I.R., Loley, C., Erdmann, J., and Ziegler, A. (2014). How to include chromosome X in your genome-wide association study. Genet Epidemiol *38*, 97-103.

Krementsov, D.N., Case, L.K., Dienz, O., Raza, A., Fang, Q., Ather, J.L., Poynter, M.E., Boyson, J.E., Bunn, J.Y., and Teuscher, C. (2017). Genetic variation in chromosome Y regulates susceptibility to influenza A virus infection. Proc Natl Acad Sci U S A *114*, 3491-3496.

Kruglyak, L., and Nickerson, D.A. (2001). Variation is the spice of life. Nat Genet *27*, 234-236.

Lam, M., Awasthi, S., Watson, H.J., Goldstein, J., Panagiotaropoulou, G., Trubetskoy, V., Karlsson, R., Frei, O., Fan, C.C., De Witte, W.*, et al.* (2019). RICOPILI: Rapid Imputation for COnsortias PIpeLIne. Bioinformatics.

Lander, E.S., and Botstein, D. (1986). Mapping complex genetic traits in humans: new methods using a complete RFLP linkage map. Cold Spring Harb Symp Quant Biol *51 Pt 1*, 49-62.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W.*, et al.* (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.

Lango Allen, H., Estrada, K., Lettre, G., Berndt, S.I., Weedon, M.N., Rivadeneira, F., Willer, C.J., Jackson, A.U., Vedantam, S., Raychaudhuri, S.*, et al.* (2010).

Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature *467*, 832-838.

Lappalainen, I., Almeida-King, J., Kumanduri, V., Senf, A., Spalding, J.D., Ur-Rehman, S., Saunders, G., Kandasamy, J., Caccamo, M., Leinonen, R*., et al.* (2015). The European Genome-phenome Archive of human data consented for biomedical research. Nat Genet *47*, 692-695.

Lariviere, V., Ni, C., Gingras, Y., Cronin, B., and Sugimoto, C.R. (2013). Bibliometrics: global gender disparities in science. Nature *504*, 211-213.

Leitsalu, L., Haller, T., Esko, T., Tammesoo, M.L., Alavere, H., Snieder, H., Perola, M., Ng, P.C., Magi, R., Milani, L*., et al.* (2015). Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. Int J Epidemiol *44*, 1137-1147.

Lenz, T.L., Deutsch, A.J., Han, B., Hu, X., Okada, Y., Eyre, S., Knapp, M., Zhernakova, A., Huizinga, T.W., Abecasis, G*., et al.* (2015). Widespread non-additive and interaction effects within HLA loci modulate the risk of autoimmune diseases. Nat Genet *47*, 1085-1090.

Lettre, G., Lange, C., and Hirschhorn, J.N. (2007). Genetic model testing and statistical power in population-based association studies of quantitative traits. Genet Epidemiol *31*, 358-362.

Levene, P.A. (1919). The structure of yeast nucleic acid. IV. Ammonia hydrolysis. Journal of Biological Chemistry *40*, 415-424.

Li, M., Li, C., and Guan, W. (2008). Evaluation of coverage variation of SNP chips for genome-wide association studies. Eur J Hum Genet *16*, 635-643.

Li, Y., Willer, C., Sanna, S., and Abecasis, G. (2009). Genotype imputation. Annu Rev Genomics Hum Genet *10*, 387-406.

Lichtenstein, P., Holm, N.V., Verkasalo, P.K., Iliadou, A., Kaprio, J., Koskenvuo, M., Pukkala, E., Skytthe, A., and Hemminki, K. (2000). Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. N Engl J Med *343*, 78-85.

Lippert, C., Listgarten, J., Liu, Y., Kadie, C.M., Davidson, R.I., and Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. Nat Methods *8*, 833-835.

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.

Lobo, I. (2008). Multifactorial inheritance and genetic disease. Nature Education *1(1):5*, 5.

Loh, P.R., Danecek, P., Palamara, P.F., Fuchsberger, C., Y, A.R., H, K.F., 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., Kichaev, G., Gazal, S., Schoech, A.P., and Price, A.L. (2018). Mixed-model association for biobank-scale datasets. Nat Genet *50*, 906-908.

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.

Loh, P.R., Tucker, G., Bulik-Sullivan, B.K., Vilhjalmsson, B.J., Finucane, H.K., Salem, R.M., Chasman, D.I., Ridker, P.M., Neale, B.M., Berger, B*., et al.* (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat Genet *47*, 284-290.

Lohmueller, K.E., Pearce, C.L., Pike, M., Lander, E.S., and Hirschhorn, J.N. (2003). Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat Genet *33*, 177-182.

Longo, D.L., and Drazen, J.M. (2016). Data Sharing. N Engl J Med *374*, 276-277.

Lordan, F., Tejedor, E., Ejarque, J., Rafanell, R., Álvarez, J., Marozzo, F., Lezzi, D., Sirvent, R., Talia, D., and Badia, R.M. (2014). ServiceSs: An Interoperable Programming Framework for the Cloud. Journal of Grid Computing *12*, 67-91.

Low, S.K., Takahashi, A., Ebana, Y., Ozaki, K., Christophersen, I.E., Ellinor, P.T., Consortium, A.F., Ogishima, S., Yamamoto, M., Satoh, M.*, et al.* (2017). Identification of six new genetic loci associated with atrial fibrillation in the Japanese population. Nat Genet *49*, 953-958.

Lu, C., Wen, Y., Hu, W., Lu, F., Qin, Y., Wang, Y., Li, S., Yang, S., Lin, Y., Wang, C.*, et al.* (2016). Y chromosome haplogroups based genome-wide association study pinpoints revelation for interactions on non-obstructive azoospermia. Sci Rep *6*, 33363.

Maan, A.A., Eales, J., Akbarov, A., Rowland, J., Xu, X., Jobling, M.A., Charchar, F.J., and Tomaszewski, M. (2017). The Y chromosome: a blueprint for men's health? Eur J Hum Genet *25*, 1181-1188.

MacDonald, M.E., Ambrose, C.M., Duyao, M.P., Myers, R.H., Lin, C., Srinidhi, L., Barnes, G., Taylor, S.A., James, M., Groot, N.*, et al.* (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell *72*, 971-983.

Machiela, M.J., and 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*, 3555-3557.

Malik, R., Chauhan, G., Traylor, M., Sargurupremraj, M., Okada, Y., Mishra, A., Rutten-Jacobs, L., Giese, A.K., van der Laan, S.W., Gretarsdottir, S.*, et al.* (2018). Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. Nat Genet *50*, 524-537.

Manning, A., Highland, H.M., Gasser, J., Sim, X., Tukiainen, T., Fontanillas, P., Grarup, N., Rivas, M.A., Mahajan, A., Locke, A.E.*, et al.* (2017). A Low-Frequency Inactivating AKT2 Variant Enriched in the Finnish Population Is Associated With Fasting Insulin Levels and Type 2 Diabetes Risk. Diabetes *66*, 2019-2032.

Manolio, T.A., Brooks, L.D., and Collins, F.S. (2008). A HapMap harvest of insights into the genetics of common disease. J Clin Invest *118*, 1590-1605.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A.*, et al.* (2009). Finding the missing heritability of complex diseases. Nature *461*, 747-753.

Marchini, J., and Howie, B. (2010). Genotype imputation for genome-wide association studies. Nat Rev Genet *11*, 499-511.

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.

Martin-Timon, I., Sevillano-Collantes, C., Segura-Galindo, A., and Del Canizo-Gomez, F.J. (2014). Type 2 diabetes and cardiovascular disease: Have all risk factors the same strength? World J Diabetes *5*, 444-470.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J.*, et al.* (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science *337*, 1190-1195.

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M., Fuchsberger, C., Danecek, P., Sharp, K*., et al.* (2016). A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet *48*, 1279-1283.

McKusick, V.A. (1976). Letter: Pleiotropism. Am J Hum Genet *28*, 301-302.

McKusick, V.A. (1991). Current trends in mapping human genes. FASEB J *5*, 12-20.

Miescher, F. (1871). Ueber die chemische Zusammensetzung der Eiterzellen. Med-Chem Unters *4*, 441-460.

Miller, M., Stone, N.J., Ballantyne, C., Bittner, V., Criqui, M.H., Ginsberg, H.N., Goldberg, A.C., Howard, W.J., Jacobson, M.S., Kris-Etherton, P.M*., et al.* (2011). Triglycerides and cardiovascular disease: a scientific statement from the American Heart Association. Circulation *123*, 2292-2333.

Mills, M.C., and Rahal, C. (2019). A scientometric review of genome-wide association studies. Commun Biol *2*, 9.

Mitt, M., Kals, M., Parn, K., Gabriel, S.B., Lander, E.S., Palotie, A., Ripatti, S., Morris, A.P., Metspalu, A., Esko, T*., et al.* (2017). Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. Eur J Hum Genet *25*, 869-876.

Molins, B., Romero-Vazquez, S., Fuentes-Prior, P., Adan, A., and Dick, A.D. (2018). C-Reactive Protein as a Therapeutic Target in Age-Related Macular Degeneration. Front Immunol *9*, 808.

Moltke, I., Grarup, N., Jorgensen, M.E., Bjerregaard, P., Treebak, J.T., Fumagalli, M., Korneliussen, T.S., Andersen, M.A., Nielsen, T.S., Krarup, N.T*., et al.* (2014). A common Greenlandic TBC1D4 variant confers muscle insulin resistance and type 2 diabetes. Nature *512*, 190-193.

Morgan, T.H. (1910). Sex Limited Inheritance in Drosophila. Science *32*, 120-122.

Mucci, L.A., Hjelmborg, J.B., Harris, J.R., Czene, K., Havelick, D.J., Scheike, T., Graff, R.E., Holst, K., Moller, S., Unger, R.H*., et al.* (2016). Familial Risk and Heritability of Cancer Among Twins in Nordic Countries. JAMA *315*, 68-76.

Muniz-Fernandez, F., Carreno-Torres, A., Morcillo-Suarez, C., and Navarro, A. (2011). Genome-wide association studies pipeline (GWASpi): a desktop application for genome-wide SNP analysis and management. Bioinformatics *27*, 1871-1872.

Nagai, A., Hirata, M., Kamatani, Y., Muto, K., Matsuda, K., Kiyohara, Y., Ninomiya, T., Tamakoshi, A., Yamagata, Z., Mushiroda, T*., et al.* (2017). Overview of the BioBank Japan Project: Study design and profile. J Epidemiol *27*, S2-S8.

Ngo, S.T., Steyn, F.J., and McCombe, P.A. (2014). Gender differences in autoimmune disease. Front Neuroendocrinol *35*, 347-369.

Nica, A.C., Montgomery, S.B., Dimas, A.S., Stranger, B.E., Beazley, C., Barroso, I., and Dermitzakis, E.T. (2010). Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. PLoS Genet *6*, e1000895.

Nicolae, D.L., Gamazon, E., Zhang, W., Duan, S., Dolan, M.E., and Cox, N.J. (2010). Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet *6*, e1000888.

Nielsen, J.B., Thorolfsdottir, R.B., Fritsche, L.G., Zhou, W., Skov, M.W., Graham, S.E., Herron, T.J., McCarthy, S., Schmidt, E.M., Sveinbjornsson, G*., et al.*

(2018). Biobank-driven genomic discovery yields new insight into atrial fibrillation biology. Nat Genet *50*, 1234-1239.

Ninds Stroke Genetics Network, and International Stroke Genetics Consortium (2016). Loci associated with ischaemic stroke and its subtypes (SiGN): a genome-wide association study. Lancet Neurol *15*, 174-184.

Nirenberg, M.W., and Matthaei, J.H. (1961). The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides. Proc Natl Acad Sci U S A *47*, 1588-1602.

Nordestgaard, B.G., and Varbo, A. (2014). Triglycerides and cardiovascular disease. Lancet *384*, 626-635.

O'Connell, J., Gurdasani, D., Delaneau, O., Pirastu, N., Ulivi, S., Cocca, M., Traglia, M., Huang, J., Huffman, J.E., Rudan, I.*, et al.* (2014). A general approach for haplotype phasing across the full spectrum of relatedness. PLoS Genet *10*, e1004234.

O'Connell, J., Sharp, K., Shrine, N., Wain, L., Hall, I., Tobin, M., Zagury, J.F., Delaneau, O., and Marchini, J. (2016). Haplotype estimation for biobank-scale data sets. Nat Genet *48*, 817-820.

Paltoo, D.N., Rodriguez, L.L., Feolo, M., Gillanders, E., Ramos, E.M., Rutter, J.L., Sherry, S., Wang, V.O., Bailey, A., Baker, R.*, et al.* (2014). Data use under the NIH GWAS data sharing policy and future directions. Nat Genet *46*, 934-938.

Parkes, M., Cortes, A., van Heel, D.A., and Brown, M.A. (2013). Genetic insights into common pathways and complex relationships among immune-mediated diseases. Nat Rev Genet *14*, 661-673.

Paul, D. (2000). A double-edged sword. Nature *405*, 515.

Pennisi, E. (2012). Genomics. ENCODE project writes eulogy for junk DNA. Science *337*, 1159, 1161.

Pers, T.H., Karjalainen, J.M., Chan, Y., Westra, H.J., Wood, A.R., Yang, J., Lui, J.C., Vedantam, S., Gustafsson, S., Esko, T.*, et al.* (2015). Biological interpretation of genome-wide association studies using predicted gene functions. Nat Commun *6*, 5890.

Pickrell, J.K., Berisa, T., Liu, J.Z., Segurel, 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.

Pividori, M., Schoettler, N., Nicolae, D.L., Ober, C., and Im, H.K. (2019). Shared and distinct genetic risk factors for childhood-onset and adult-onset asthma: genome-wide and transcriptome-wide studies. Lancet Respir Med *7*, 509-522.

Polderman, T.J., Benyamin, B., de Leeuw, C.A., Sullivan, P.F., van Bochoven, A., Visscher, P.M., and Posthuma, D. (2015). Meta-analysis of the heritability of human traits based on fifty years of twin studies. Nat Genet *47*, 702-709.

Popejoy, A.B., and Fullerton, S.M. (2016). Genomics is failing on diversity. Nature *538*, 161-164.

Poulsen, P., Kyvik, K.O., Vaag, A., and Beck-Nielsen, H. (1999). Heritability of type II (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance--a population-based twin study. Diabetologia *42*, 139-145.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet *38*, 904-909.

Price, A.L., Spencer, C.C., and Donnelly, P. (2015). Progress and promise in understanding the genetic basis of common diseases. Proc Biol Sci *282*, 20151684.

Pritchard, J.K. (2001). Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet *69*, 124-137.

Pritchard, J.K., and Cox, N.J. (2002). The allelic architecture of human disease genes: common disease-common variant...or not? Hum Mol Genet *11*, 2417-2423.

Pruim, R.J., Welch, R.P., Sanna, S., Teslovich, T.M., Chines, P.S., Gliedt, T.P., Boehnke, M., Abecasis, G.R., and Willer, C.J. (2010). LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics *26*, 2336-2337.

Raychaudhuri, S., Plenge, R.M., Rossin, E.J., Ng, A.C., International Schizophrenia, C., Purcell, S.M., Sklar, P., Scolnick, E.M., Xavier, R.J., Altshuler, D.*, et al.* (2009). Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. PLoS Genet *5*, e1000534.

Reich, D.E., and Lander, E.S. (2001). On the allelic spectrum of human disease. Trends Genet *17*, 502-510.

Rhee, S.Y., Wood, V., Dolinski, K., and Draghici, S. (2008). Use and misuse of the gene ontology annotations. Nat Rev Genet *9*, 509-515.

Richmond, M.L. (2007). Opportunities for women in early genetics. Nat Rev Genet *8*, 897-902.

Riordan, J.R., Rommens, J.M., Kerem, B., Alon, N., Rozmahel, R., Grzelczak, Z., Zielenski, J., Lok, S., Plavsic, N., Chou, J.L.*, et al.* (1989). Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. Science *245*, 1066-1073.

Risch, N., and Merikangas, K. (1996). The future of genetic studies of complex human diseases. Science *273*, 1516-1517.

Risch, N.J. (2000). Searching for genetic determinants in the new millennium. Nature *405*, 847-856.

Roadmap Epigenomics, C., Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J.*, et al.* (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317-330.

Robinson, E.B., Lichtenstein, P., Anckarsater, H., Happe, F., and Ronald, A. (2013). Examining and interpreting the female protective effect against autistic behavior. Proc Natl Acad Sci U S A *110*, 5258-5262.

Rommens, J.M., Iannuzzi, M.C., Kerem, B., Drumm, M.L., Melmer, G., Dean, M., Rozmahel, R., Cole, J.L., Kennedy, D., Hidaka, N.*, et al.* (1989). Identification of the cystic fibrosis gene: chromosome walking and jumping. Science *245*, 1059-1065.

Roselli, C., Chaffin, M.D., Weng, L.C., Aeschbacher, S., Ahlberg, G., Albert, C.M., Almgren, P., Alonso, A., Anderson, C.D., Aragam, K.G.*, et al.* (2018). Multi-ethnic genome-wide association study for atrial fibrillation. Nat Genet *50*, 1225-1233.

Rubinacci, S., Delaneau, O., and Marchini, J. (2019). Genotype imputation using the Positional Burrows Wheeler Transform. bioRxiv, 797944.

Saber, H., Himali, J.J., Beiser, A.S., Shoamanesh, A., Pikula, A., Roubenoff, R., Romero, J.R., Kase, C.S., Vasan, R.S., and Seshadri, S. (2017). Serum Insulin-Like Growth Factor 1 and the Risk of Ischemic Stroke: The Framingham Study. Stroke *48*, 1760-1765.

Sachidanandam, R., Weissman, D., Schmidt, S.C., Kakol, J.M., Stein, L.D., Marth, G., Sherry, S., Mullikin, J.C., Mortimore, B.J., Willey, D.L.*, et al.* (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature *409*, 928-933.

Salanti, G., Southam, L., Altshuler, D., Ardlie, K., Barroso, I., Boehnke, M., Cornelis, M.C., Frayling, T.M., Grallert, H., Grarup, N.*, et al.* (2009). Underlying genetic models of inheritance in established type 2 diabetes associations. Am J Epidemiol *170*, 537-545.

Schaid, D.J., Chen, W., and Larson, N.B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. Nat Rev Genet *19*, 491-504.

Schaub, M.A., Boyle, A.P., Kundaje, A., Batzoglou, S., and Snyder, M. (2012). Linking disease associations with regulatory information in the human genome. Genome Res *22*, 1748-1759.

Schildkraut, J.M., Risch, N., and Thompson, W.D. (1989). Evaluating genetic association among ovarian, breast, and endometrial cancer: evidence for a breast/ovarian cancer relationship. Am J Hum Genet *45*, 521-529.

Schork, N.J., Murray, S.S., Frazer, K.A., and Topol, E.J. (2009). Common vs. rare allele hypotheses for complex diseases. Curr Opin Genet Dev *19*, 212-219.

Scott, L.J., Mohlke, K.L., Bonnycastle, L.L., Willer, C.J., Li, Y., Duren, W.L., Erdos, M.R., Stringham, H.M., Chines, P.S., Jackson, A.U.*, et al.* (2007). A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science *316*, 1341-1345.

Segre, A.V., Consortium, D., investigators, M., Groop, L., Mootha, V.K., Daly, M.J., and Altshuler, D. (2010). Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. PLoS Genet *6*.

Sezgin, E., Lind, J.M., Shrestha, S., Hendrickson, S., Goedert, J.J., Donfield, S., Kirk, G.D., Phair, J.P., Troyer, J.L., O'Brien, S.J.*, et al.* (2009). Association of Y chromosome haplogroup I with HIV progression, and HAART outcome. Hum Genet *125*, 281-294.

Sidore, C., Busonero, F., Maschio, A., Porcu, E., Naitza, S., Zoledziewska, M., Mulas, A., Pistis, G., Steri, M., Danjou, F.*, et al.* (2015). Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. Nat Genet *47*, 1272-1281.

Sivakumaran, S., Agakov, F., Theodoratou, E., Prendergast, J.G., Zgaga, L., Manolio, T., Rudan, I., McKeigue, P., Wilson, J.F., and Campbell, H. (2011). Abundant pleiotropy in human complex diseases and traits. Am J Hum Genet *89*, 607-618.

Skaletsky, H., Kuroda-Kawaguchi, T., Minx, P.J., Cordum, H.S., Hillier, L., Brown, L.G., Repping, S., Pyntikova, T., Ali, J., Bieri, T.*, et al.* (2003). The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature *423*, 825-837.

Speliotes, E.K., Willer, C.J., Berndt, S.I., Monda, K.L., Thorleifsson, G., Jackson, A.U., Lango Allen, H., Lindgren, C.M., Luan, J., Magi, R.*, et al.* (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nat Genet *42*, 937-948.

Spjuth, O., Bongcam-Rudloff, E., Hernandez, G.C., Forer, L., Giovacchini, M., Guimera, R.V., Kallio, A., Korpelainen, E., Kandula, M.M., Krachunov, M.*, et al.* (2015). Experiences with workflows for automating data-intensive bioinformatics. Biol Direct *10*, 43.

Stahl, E.A., Wegmann, D., Trynka, G., Gutierrez-Achury, J., Do, R., Voight, B.F., Kraft, P., Chen, R., Kallberg, H.J., Kurreeman, F.A.*, et al.* (2012). Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. Nat Genet *44*, 483-489.

Steinthorsdottir, V., Thorleifsson, G., Reynisdottir, I., Benediktsson, R., Jonsdottir, T., Walters, G.B., Styrkarsdottir, U., Gretarsdottir, S., Emilsson, V., Ghosh, S.*, et al.* (2007). A variant in CDKAL1 influences insulin response and risk of type 2 diabetes. Nat Genet *39*, 770-775.

Stephens, Z.D., Lee, S.Y., Faghri, F., Campbell, R.H., Zhai, C., Efron, M.J., Iyer, R., Schatz, M.C., Sinha, S., and Robinson, G.E. (2015). Big Data: Astronomical or Genomical? PLoS Biol *13*, e1002195.

Stranger, B.E., Stahl, E.A., and Raj, T. (2011). Progress and promise of genome-wide association studies for human complex trait genetics. Genetics *187*, 367-383.

Stringer, S., Polderman, T.J.C., and Posthuma, D. (2017). Majority of human traits do not show evidence for sex-specific genetic and environmental effects. Sci Rep *7*, 8688.

Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M.H.*, et al.* (2015). An integrated map of structural variation in 2,504 human genomes. Nature *526*, 75-81.

Sullivan, P.F., Kendler, K.S., and Neale, M.C. (2003). Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. Arch Gen Psychiatry *60*, 1187-1192.

Szumilas, M. (2010). Explaining odds ratios. J Can Acad Child Adolesc Psychiatry *19*, 227-229.

Tabassum, R., Ramo, J.T., Ripatti, P., Koskela, J.T., Kurki, M., Karjalainen, J., Palta, P., Hassan, S., Nunez-Fontarnau, J., Kiiskinen, T.T.J.*, et al.* (2019). Genetic architecture of human plasma lipidome and its link to cardiovascular disease. Nat Commun *10*, 4329.

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.* (2019). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. bioRxiv, 563866.

Tam, V., Patel, N., Turcotte, M., Bosse, Y., Pare, G., and Meyre, D. (2019). Benefits and limitations of genome-wide association studies. Nat Rev Genet.

The Genome of the Netherlands Consortium (2014). Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat Genet *46*, 818-825.

Traglia, M., Bseiso, D., Gusev, A., Adviento, B., Park, D.S., Mefford, J.A., Zaitlen, N., and Weiss, L.A. (2017). Genetic Mechanisms Leading to Sex Differences Across Common Diseases and Anthropometric Traits. Genetics *205*, 979-992.

Traylor, M., Farrall, M., Holliday, E.G., Sudlow, C., Hopewell, J.C., Cheng, Y.C., Fornage, M., Ikram, M.A., Malik, R., Bevan, S.*, et al.* (2012). Genetic risk factors for ischaemic stroke and its subtypes (the METASTROKE collaboration): a meta-analysis of genome-wide association studies. Lancet Neurol *11*, 951-962.

Tryka, K.A., Hao, L., Sturcke, A., Jin, Y., Wang, Z.Y., Ziyabari, L., Lee, M., Popova, N., Sharopova, N., Kimura, M.*, et al.* (2014). NCBI's Database of Genotypes and Phenotypes: dbGaP. Nucleic Acids Res *42*, D975-979.

Tsui, L.C., and Dorfman, R. (2013). The cystic fibrosis gene: a molecular genetic perspective. Cold Spring Harb Perspect Med *3*, a009472.

Tukiainen, T., Pirinen, M., Sarin, A.P., Ladenvall, C., Kettunen, J., Lehtimaki, T., Lokki, M.L., Perola, M., Sinisalo, J., Vlachopoulou, E.*, et al.* (2014). Chromosome X-wide association study identifies Loci for fasting insulin and height and evidence for incomplete dosage compensation. PLoS Genet *10*, e1004127.

Tukiainen, T., Villani, A.C., Yen, A., Rivas, M.A., Marshall, J.L., Satija, R., Aguirre, M., Gauthier, L., Fleharty, M., Kirby, A.*, et al.* (2017). Landscape of X chromosome inactivation across human tissues. Nature *550*, 244-248.

Turkheimer, E., Haley, A., Waldron, M., D'Onofrio, B., and Gottesman, II (2003). Socioeconomic status modifies heritability of IQ in young children. Psychol Sci *14*, 623-628.

Turner, S.D. (2018). qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. Journal of Open Source Software *3*, 731.

UK10K Consortium, Walter, K., Min, J.L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., Perry, J.R., Xu, C., Futema, M.*, et al.* (2015). The UK10K project identifies rare variants in health and disease. Nature *526*, 82-90.

UK Biobank, UK Biobank leads the way in genetics research to tackle chronic diseases, https://www.ukbiobank.ac.uk/2019/09/uk-biobank-leads-the-way-in-genetics-research-to-tackle-chronic-diseases/, Oct 13, 2019

van Leeuwen, E.M., Kanterakis, A., Deelen, P., Kattenberg, M.V., Genome of the Netherlands, C., Slagboom, P.E., de Bakker, P.I., Wijmenga, C., Swertz, M.A., Boomsma, D.I.*, et al.* (2015). Population-specific genotype imputations using minimac or IMPUTE2. Nat Protoc *10*, 1285-1296.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A.*, et al.* (2001). The sequence of the human genome. Science *291*, 1304-1351.

Visscher, P.M., Brown, M.A., McCarthy, M.I., and Yang, J. (2012). Five years of GWAS discovery. Am J Hum Genet *90*, 7-24.

Visscher, P.M., and Goddard, M.E. (2019). From R.A. Fisher's 1918 Paper to GWAS a Century Later. Genetics *211*, 1125-1130.

Visscher, P.M., Hill, W.G., and Wray, N.R. (2008). Heritability in the genomics era--concepts and misconceptions. Nat Rev Genet *9*, 255-266.

Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. Am J Hum Genet *101*, 5-22.

Voight, B.F., Kang, H.M., Ding, J., Palmer, C.D., Sidore, C., Chines, P.S., Burtt, N.P., Fuchsberger, C., Li, Y., Erdmann, J.*, et al.* (2012). The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. PLoS Genet *8*, e1002793.

Võsa, U., Claringbould, A., Westra, H.-J., Bonder, M.J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Kasela, S.*, et al.* (2018). Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. bioRxiv, 447367.

Wainschtein, P., Jain, D.P., Yengo, L., Zheng, Z., Cupples, L.A., Shadyab, A.H., McKnight, B., Shoemaker, B.M., Mitchell, B.D., Psaty, B.M.*, et al.* (2019). Recovery of trait heritability from whole genome sequence data. bioRxiv, 588020.

Watson, J.D., and Crick, F.H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature *171*, 737-738.

Wellcome Trust Case Control, C. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature *447*, 661-678.

Wellcome Trust Case Control Consortium, Maller, J.B., McVean, G., Byrnes, J., Vukcevic, D., Palin, K., Su, Z., Howson, J.M., Auton, A., Myers, S*., et al.* (2012). Bayesian refinement of association signals for 14 loci in 3 common diseases. Nat Genet *44*, 1294-1301.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L*., et al.* (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res *42*, D1001-1006.

Wetterstrand, K.A. (2019). DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: wwwgenomegov/sequencingcostsdata Accessed 30 June 2019.

WHO (2018). Noncommunicable diseases country profiles 2018.

Willer, C.J., Li, Y., and Abecasis, G.R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics *26*, 2190-2191.

Willyard, C. (2018). New human gene tally reignites debate. Nature *558*, 354-355.

Wise, A.L., Gyi, L., and Manolio, T.A. (2013). eXclusion: toward integrating the X chromosome in genome-wide association analyses. Am J Hum Genet *92*, 643-647.

Wittke-Thompson, J.K., Pluzhnikov, A., and Cox, N.J. (2005). Rational inferences about departures from Hardy-Weinberg equilibrium. Am J Hum Genet *76*, 967-986.

Wojcik, G.L., Graff, M., Nishimura, K.K., Tao, R., Haessler, J., Gignoux, C.R., Highland, H.M., Patel, Y.M., Sorokin, E.P., Avery, C.L*., et al.* (2019). Genetic analyses of diverse populations improves discovery for complex traits. Nature *570*, 514-518.

Wood, A.R., Tyrrell, J., Beaumont, R., Jones, S.E., Tuke, M.A., Ruth, K.S., consortium, G., Yaghootkar, H., Freathy, R.M., Murray, A*., et al.* (2016). Variants in the FTO and CDKAL1 loci have recessive effects on risk of obesity and type 2 diabetes, respectively. Diabetologia *59*, 1214-1221.

Wright, S. (1920). The Relative Importance of Heredity and Environment in Determining the Piebald Pattern of Guinea-Pigs. Proc Natl Acad Sci U S A *6*, 320-332.

Xu, H., Dorn, G.W., 2nd, Shetty, A., Parihar, A., Dave, T., Robinson, S.W., Gottlieb, S.S., Donahue, M.P., Tomaselli, G.F., Kraus, W.E*., et al.* (2018). A Genome-Wide Association Study of Idiopathic Dilated Cardiomyopathy in African Americans. J Pers Med *8*.

Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W*., et al.* (2010). Common SNPs explain a large proportion of the heritability for human height. Nat Genet *42*, 565-569.

Yang, J., Weedon, M.N., Purcell, S., Lettre, G., Estrada, K., Willer, C.J., Smith, A.V., Ingelsson, E., O'Connell, J.R., Mangino, M*., et al.* (2011). Genomic inflation factors under polygenic inheritance. Eur J Hum Genet *19*, 807-812.

Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauley, M., Franklin, M.J., Shenker, S., and Stoica, I. (2012). Resilient distributed datasets: a fault-tolerant abstraction for in-memory cluster computing. In Proceedings of the

9th USENIX conference on Networked Systems Design and Implementation (San Jose, CA: USENIX Association), pp. 2-2.

Zaitlen, N., Kraft, P., Patterson, N., Pasaniuc, B., Bhatia, G., Pollack, S., and Price, A.L. (2013). Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. PLoS Genet *9*, e1003520.

Zaitlen, N., Pasaniuc, B., Sankararaman, S., Bhatia, G., Zhang, J., Gusev, A., Young, T., Tandon, A., Pollack, S., Vilhjalmsson, B.J.*, et al.* (2014). Leveraging population admixture to characterize the heritability of complex traits. Nat Genet *46*, 1356-1362.

Zeng, P., Zhao, Y., Qian, C., Zhang, L., Zhang, R., Gou, J., Liu, J., Liu, L., and Chen, F. (2015). Statistical analysis for genome-wide association study. J Biomed Res *29*, 285-297.

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.* (2018). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat Genet *50*, 1335-1341.

Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. Nat Genet *44*, 821-824.

Zhu, Z., Bakshi, A., Vinkhuyzen, A.A., Hemani, G., Lee, S.H., Nolte, I.M., van Vliet-Ostaptchouk, J.V., Snieder, H., LifeLines Cohort, S., Esko, T.*, et al.* (2015). Dominance genetic variation contributes little to the missing heritability for human complex traits. Am J Hum Genet *96*, 377-385.

Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M.R., Powell, J.E., Montgomery, G.W., Goddard, M.E., Wray, N.R., Visscher, P.M.*, et al.* (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat Genet *48*, 481-487.

Appendix

Appendix 1. Re-analysis of public genetic data reveals a rare X-chromosomal variant associated with type 2 diabetes.

Bonàs-Guarch S, Guindo-Martínez M, Miguel-Escalada I, Grarup N, Sebastian D, Rodriguez-Fos E, Sánchez F, Planas-Fèlix M, Cortes-Sánchez P, González S, Timshel P, Pers TH, Morgan CC, Moran I, Atla G, González JR, Puiggros M, Martí J, Andersson EA, Díaz C, Badia RM, Udler M, Leong A Kaur V, Flannick J, Jørgensen T, Linneberg A, Jørgensen M, Witte DR, Christensen C, Brandslund I, Appel EV, Scott RA, Luan J, Langenberg C, Wareham NJ, Pedersen O, Zorzano A, Florez JC, Hansen T, Ferrer J, Mercader JM, Torrents D. *Nat Commun*. 2018 Jan 22;9(1):32.

Contribution:

- Implementation of the combination of results from multiple reference panels.
- Analysis of the QCed GERA cohort using GUIDANCE, with 1000G phase 1 and UK10K as reference panels for genotype imputation. Only the additive model was implemented during the association test.
- Revision of the manuscript.

nature
COMMUNICATIONS

ARTICLE

DOI: 10.1038/s41467-017-02380-9 OPEN

Corrected: Publisher correction

Re-analysis of public genetic data reveals a rare X-chromosomal variant associated with type 2 diabetes

Sílvia Bonàs-Guarch et al.[#]

The reanalysis of existing GWAS data represents a powerful and cost-effective opportunity to gain insights into the genetics of complex diseases. By reanalyzing publicly available type 2 diabetes (T2D) genome-wide association studies (GWAS) data for 70,127 subjects, we identify seven novel associated regions, five driven by common variants (*LYPLAL1*, *NEUROG3*, *CAMKK2*, *ABO*, and *GIP* genes), one by a low-frequency (*EHMT2*), and one driven by a rare variant in chromosome Xq23, rs146662075, associated with a twofold increased risk for T2D in males. rs146662075 is located within an active enhancer associated with the expression of Angiotensin II Receptor type 2 gene (*AGTR2*), a modulator of insulin sensitivity, and exhibits allelic specific activity in muscle cells. Beyond providing insights into the genetics and pathophysiology of T2D, these results also underscore the value of reanalyzing publicly available data using novel genetic resources and analytical approaches.

Correspondence and requests for materials should be addressed to J.M.M. (email: mercader@broadinstitute.org) or to D.T. (email: david.torrents@bsc.es)
#A full list of authors and their affliations appears at the end of the paper

During the last decade, hundreds of genome-wide association studies (GWAS) have been performed with the aim of providing a better understanding of the biology of complex diseases, improving their risk prediction, and ultimately discovering novel therapeutic targets[1]. However, the majority of the published GWAS have only reported primary findings, which generally explain a small fraction of the estimated heritability. To examine the missing heritability, most strategies involve generating new genetic and clinical data. Very rarely are new studies based on the revision and reanalysis of existing genetic data by applying more powerful analytic techniques and resources after the primary GWAS findings are published. These cost-effective reanalysis strategies are now possible, given emerging (1) data-sharing initiatives with large amounts of primary genetic data for multiple human genetic diseases, as well as (2) new and improved GWAS methodologies and resources. Notably, genotype imputation with novel sequence-based reference panels can now substantially increase the genetic resolution of GWASs from previously genotyped data sets[2], reaching good-quality imputation of low frequency (minor allele frequency [MAF]: $0.01 \le \text{MAF} < 0.05$) and rare variants ($\text{MAF} < 0.01$), increasing the power to identify novel associations, and fine map the known ones. Moreover, the availability of publicly available primary genetic data allows the homogeneous integration of multiple data sets from different origins providing more accurate meta-analysis results, particularly at the low ranges of allele frequency. Finally, the vast majority of reported GWAS analyses omits the X chromosome, despite representing 5% of the genome and coding for more than 1,500 genes[3]. The reanalysis of publicly available data also enables interrogation of this chromosome.

We hypothesized that a unified reanalysis of multiple publicly available data sets, applying homogeneous standardized quality control (QC), genotype imputation, and association methods, as well as novel and denser sequence-based reference panels for imputation would provide new insights into the genetics and the pathophysiology of complex diseases. To test this hypothesis, we focused this study on type 2 diabetes (T2D), one of the most prevalent complex diseases for which many GWAS have been performed during the past decade[4]. These studies have allowed the identification of more than 100 independent loci, most of them driven by common variants, with a few exceptions[5]. Despite these efforts, there is still a large fraction of genetic heritability hidden in the data, and the role of low-frequency variants, although recently proposed to be minor[6], has still not been fully explored. The availability of large T2D genetic data sets in combination with larger and more comprehensive genetic variation reference panels[2], provides the opportunity to impute a significantly increased fraction of low-frequency and rare variants, and to study their contribution to the risk of developing this disease. This strategy also allows us to fine map known associated loci, increasing the chances of finding causal variants and understanding their functional impact. We therefore gathered publicly available T2D GWAS cohorts with European ancestry, comprising a total of 13,857 T2D cases and 62,126 controls, to which we first applied harmonization and quality control protocols covering the whole genome (including the X chromosome). We then performed imputation using 1000 Genomes Project (1000G)[7] and UK10K[2] reference panels, followed by association testing. By using this strategy, we identified novel associated regions driven by common, low-frequency and rare variants, fine mapped and functionally annotated the existing and novel ones, allowing us to describe a regulatory mechanism disrupted by a novel rare and large-effect variant identified at the X chromosome.

Results

Overall analysis strategy. As shown in Fig. 1, we first gathered all T2D case-control GWAS individual-level data that were available through the EGA and dbGaP databases (i.e., Gene Environment-Association Studies [GENEVA], Wellcome Trust Case Control Consortium [WTCCC], Finland–United States Investigation of NIDDM Genetics [FUSION], Resource for Genetic Epidemiology Research on Aging [GERA], and Northwestern NuGENE project [NuGENE]). We harmonized these cohorts, applied standardized quality control procedures, and filtered out low-quality variants and samples (Methods and Supplementary Notes). After this process, a total of 70,127 subjects (70KforT2D, 12,931 cases, and 57,196 controls, Supplementary Data 1) were retained for downstream analysis. Each of these cohorts was then imputed to the 1000G and UK10K reference panels using an integrative method, which selected the results from the reference panel that provided the highest accuracy for each variant, according to IMPUTE2 info score (Methods). Finally, the results from each of these cohorts were meta-analyzed (Fig. 1), obtaining a total of 15,115,281 variants with good imputation quality (IMPUTE2 info score ≥ 0.7, $\text{MAF} \ge 0.001$, and I^2 heterogeneity score < 0.75), across 12,931 T2D cases and 57,196 controls. Of these, 6,845,408 variants were common ($\text{MAF} \ge 0.05$), 3,100,848 were low-frequency ($0.01 \le \text{MAF} < 0.05$), and 5,169,025 were rare ($0.001 \le \text{MAF} < 0.01$). Merging the imputation results derived from the two reference panels substantially improved the number of good-quality imputed variants, particularly within the low-frequency and rare spectrum, compared to the imputation results obtained with each of the panels separately. For example, a set of 5,169,025 rare variants with good quality was obtained after integrating 1000G and UK10K results, while only 2,878,263 rare variants were imputed with 1000G and 4,066,210 with UK10K (Supplementary Fig. 1A). This strategy also allowed us to impute 1,357,753 indels with good quality (Supplementary Fig. 1B).

To take full advantage of publicly available genetic data, we used three main meta-analytic approaches to adapt to the three most common strategies for genetic data sharing: individual-level genotypes, summary statistics, and single-case queries through the Type 2 Diabetes Knowledge Portal (T2D Portal) (http://www.type2diabetesgenetics.org/). We first meta-analyzed all summary statistics results from the DIAGRAM trans-ancestry meta-analysis[8] (26,488 cases and 83,964 controls), selecting 1,918,233 common variants (MAF ≥ 0.05), mostly imputed from HapMap, with the corresponding fraction of non-overlapping samples in our 70KforT2D set, i.e. the GERA and the NuGENE cohorts, comprising a total of 7,522 cases and 50,446 controls (Fig. 1, Supplementary Data 1). Second, the remaining variants (13,197,048), which included mainly non-HapMap variants ($\text{MAF} < 0.05$) or variants not tested above, were meta-analyzed using all five cohorts from the 70KforT2D resource (Supplementary Data 1). Finally, low-frequency variants located in coding regions and with $p \le 1 \times 10^{-4}$ were meta-analyzed using the non-overlapping fraction of samples with the data from the T2D Portal through the interrogation of exome array and whole-exome sequence data from ~80,000 and ~17,000 individuals, respectively[6].

Pathway and functional enrichment analysis. To explore whether our results recapitulate the pathophysiology of T2D, we performed gene-set enrichment analysis with all the variants with $p \le 1 \times 10^{-5}$ using DEPICT[9] (Methods). This analysis showed enrichment of genes expressed in pancreas (ranked first in tissue enrichment analysis, $p = 7.8 \times 10^{-4}$, FDR < 0.05, Supplementary Data 2) and cellular response to insulin stimulus (ranked second in gene-set enrichment analysis, $p = 3.9 \times 10^{-8}$, FDR = 0.05,

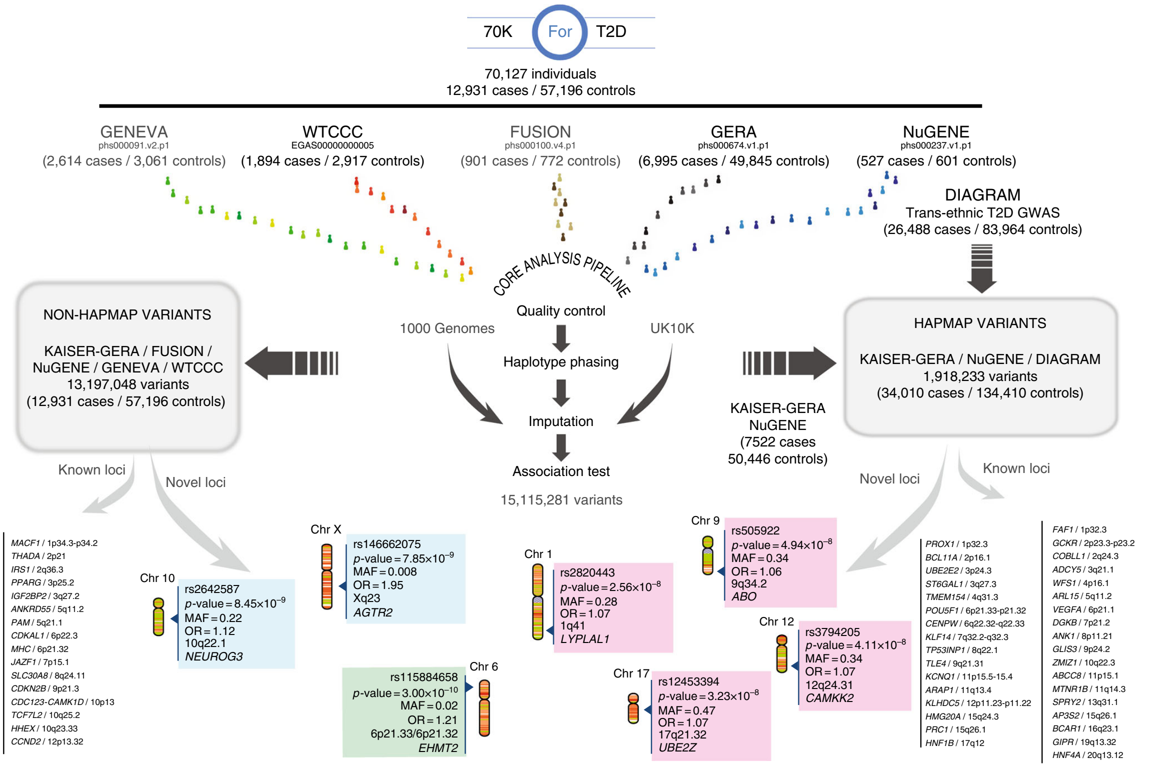


Fig. 1 Discovery and replication strategy. Publicly available GWAS datasets representing a total of 12,931 cases and 57,196 controls (70KforT2D) were first quality controlled, phased, and imputed, using 1000G and UK10K separately. For those variants that were present in the DIAGRAM trans-ethnic meta-analysis, we used the summary statistics to meta-analyze our results with the cohorts that had no overlap with any of the cohorts included in the DIAGRAM trans-ethnic meta-analysis. With this first meta-analysis, we discovered four novel loci (within magenta panels). For the rest of the variants, we meta-analyzed all the 70KforT2D data sets, which resulted in two novel loci (in blue panels). All the variants that were coding and showed a p-value of $\leq 1 \times 10^{-4}$ were tested for replication by interrogating the summary statistics in the Type 2 Diabetes Knowledge Portal (T2D Portal) (http://www.type2diabetesgenetics.org/). This uncovered a novel low-frequency variant in the *EHMT2* gene (highlighted with a green panel)

Supplementary Data 3, Supplementary Fig. 2, Supplementary Fig. 3), in concordance with the current knowledge of the molecular basis of T2D.

In addition, variant set enrichment analysis of the T2D-associated credible sets across regulatory elements defined in isolated human pancreatic islets showed a significant enrichment for active regulatory enhancers (Supplementary Fig. 4), suggesting that causal SNPs within associated regions have a regulatory function, as previously reported[10].

Fine-mapping and functional characterization of T2D loci. The three association strategies allowed us to identify 57 genome-wide significant associated loci ($p \leq 5 \times 10^{-8}$), of which seven were not previously reported as associated with T2D (Table 1). The remaining 50 loci have been previously reported and included, for example, two low-frequency variants recently discovered in Europeans, one located within one of the *CCND2* introns (rs76895963), and a missense variant within the *PAM*[5] gene. Furthermore, we confirmed that the magnitude and direction of the effect of all the associated variants ($p \leq 0.001$) were highly consistent with those reported previously ($\rho = 0.92$, $p = 1 \times 10^{-248}$, Supplementary Fig. 5). In addition, the direction of effect was consistent with all 139 previously reported variants, except three that were discovered in east and south Asian populations (Supplementary Data 4).

The high coverage of genetic variation ascertained in this study allowed us to fine-map known and novel loci, providing more candidate causal variants for downstream functional interpretations. We constructed 99% credible variant sets[11] for each of these loci, i.e. the subset of variants that have, in aggregate, 99% probability of containing the true causal variant for all 57 loci (Supplementary Data 5). As an important improvement over previous T2D genetic studies, we identified small structural variants within the credible sets, consisting mostly of insertions and deletions between 1 and 1,975 nucleotides. In fact, out of the 8,348 variants included within the credible sets for these loci, 927 (11.1%) were indels, of which 105 were genome-wide significant (Supplementary Data 6). Interestingly, by integrating imputed results from 1000G and UK10K reference panels, we gained up to 41% of indels, which were only identified by either one of the two reference panels, confirming the advantage of integrating the results from both reference panels. Interestingly, 15 of the 71 previously reported loci that we replicated ($p \leq 5.3 \times 10^{-4}$ after correcting for multiple testing) have an indel as the top variant, highlighting the potential role of this type of variation in the susceptibility for T2D. For example, within the *IGF2BP2* intron, a well-established and functionally validated locus for T2D[12], we found that 12 of the 57 variants within its 99% credible set correspond to indels with genome-wide significance ($5.6 \times 10^{-16} < p < 2.4 \times 10^{-15}$), which collectively represented 18.4% posterior probability of being causal.

Table 1 Novel T2D-associated loci

			OR (95% CI) *P*-value			
Novel Locus	**Chr**	**rsID--Risk Allele**	**Stage1 Discovery Meta-analysis**	**Stage2 Replication Meta-analysis**	**Stage1 + Stage2 Combined Meta-analysis**	**MAF**
LYPLAL1/ZC3H11B (1q41)	1	rs2820443-T	1.08 (1.04–1.13) $2.94 \times 10^{-4\ a}$	1.06 (1.03–1.09) $2.10 \times 10^{-5\ b}$	1.07 (1.04–1.09) $2.56 \times 10^{-8\ c}$	0.28
EHMT2 (6p21.33–p21.32)	6	rs115884658-A	1.34 (1.18–1.53) $1.00 \times 10^{-5\ a}$	1.17 (1.09–1.26) $2.90 \times 10^{-6\ c,\ d}$	1.21 (1.14–1.29) $3.00 \times 10^{-10\ c}$	0.02
ABO (9q34.2)	9	rs505922-C	1.07 (1.03–1.11) $6.93 \times 10^{-4\ a}$	1.06 (1.03–1.09) $1.90 \times 10^{-5\ b}$	1.06 (1.04–1.09) $4.94 \times 10^{-8\ c}$	0.34
NEUROG3 (10q22.1)	10	rs2642587-G	1.12 (1.08–1.16) $8.45 \times 10^{-9\ e}$	-	-	0.22
CAMKK2 (12q24.31)	12	rs3794205-G	1.09 (1.05–1.14) $4.18 \times 10^{-5\ a}$	1.06 (1.03–1.09) $1.60 \times 10^{-4\ b}$	1.07 (1.04–1.10) $4.11 \times 10^{-8\ c}$	0.32
CALCOCO2/ATP5G1/ UBE2Z/SNF8/GIP (17q21.32)	17	rs12453394-A	1.08 (1.04–1.12) $7.86 \times 10^{-5\ a}$	1.07 (1.03–1.11) $9.60 \times 10^{-5\ b}$	1.07 (1.05–1.10) $3.23 \times 10^{-8\ c}$	0.47
AGTR2 (Xq23)	X	rs146662075-T	3.09 (2.06–4.60) $3.24 \times 10^{-8\ f}$	1.57 (1.19–2.07) $1.42 \times 10^{-3\ g}$	1.95 (1.56–2.45) 7.85×10^{-9}	0.008

Chr chromosome, *OR* odds ratio, *MAF* minor allele frequency

[a] Imputed based public GWAS discovery meta-analysis (NuGENE + GERA cohort, 7,522 cases and 50,446 controls)

[b] Transancestry DIAGRAM Consortium (26,488 cases and 83,964 controls)[c] Meta *P*-value estimated using a weighted *Z*-score method due to unavailable SE information from Stage 2 replication cohorts[d] T2D Diabetes Genetic Portal (Exome-Chip + Exome Sequencing, 35,789 cases and 56,738 controls)[e] Full imputed based public GWAS meta-analysis (NuGENE + GERA cohort + GENEVA + FUSION + WTCCC, 12,931 cases and 57,196 controls)

[f] 70KforT2D Men Cohort (GERA cohort + GENEVA + FUSION, 5,277 cases and 15,702 controls older than 55 years)

[g] Replication Men Cohort SIGMA UK10K imputation + InterAct + Danish Cohort (case control and follow-up) + Partners Biobank + UK Biobank (18,370 cases and 88,283 controls older than 55 years and OGTT > 7.8 mmol l^{-1}, when available)

To prioritize causal variants within all the identified associated loci, we annotated their corresponding credible sets using the Variant Effector Predictor (VEP) for coding variants[13] (Supplementary Data 7), and the Combined Annotation-Dependent Depletion (CADD)[14] and LINSIGHT[15] tools for non-coding variation (Supplementary Data 8 and 9). In addition, we tested the effect of all variants on expression across multiple tissues by interrogating GTEx[16] and RNA-sequencing gene expression data from pancreatic islets[17].

Novel T2D-associated loci driven by common variants. Beyond the detailed characterization of the known T2D-associated regions, we also identified seven novel loci, among which, five were driven by common variants with modest effect sizes (1.06 < OR < 1.12; Table 1, Fig. 2, Supplementary Fig. 6 and 7).

Within the first novel T2D-associated locus in chromosome 1q41 (*LYPLAL1-ZC3H11B*, rs2820443, OR = 1.07 [1.04–1.09], $p = 2.6 \times 10^{-8}$), several variants have been previously associated with waist-to-hip ratio, height, visceral adipose fat in females, adiponectin levels, fasting insulin, and non-alcoholic fatty liver disease[18–23]. Among the genes in this locus, *LYPLAL1*, which encodes for lysophospholypase-like 1, appears to be the most likely effector gene, as it has been found to be downregulated in mouse models of diet-induced obesity and upregulated during adipogenesis[24].

Second, a novel locus at chromosome 9q34.2 region (*ABO*, rs505922, OR = 1.06 [1.04–1.09], $p = 4.9 \times 10^{-8}$) includes several variants that have been previously associated with other metabolic traits. For example, the variant rs651007, in linkage disequilibrium (LD) with rs505922 ($r^2 = 0.507$), has been shown to be associated with fasting glucose[25], and rs514659 (r^2 with top = 1) is associated with an increased risk for cardiometabolic disorders[26]. One of the variants within the credible set was the single base-pair frame-shift deletion defining the blood group O[27]. In concordance with previous results that linked O blood type with a lower risk of developing T2D[28], the frame-shift deletion determining the blood group type O was associated with a protective effect for T2D in our study (rs8176719, $p = 3.4 \times 10^{-4}$, OR = 0.95 [0.91–0.98]). In addition, several variants within this credible set are associated with the expression of the *ABO* gene in multiple tissues including skeletal muscle, adipose tissue, and pancreatic islets (Supplementary Data 9, Supplementary Data 10).

Third, a novel locus at chromosome 10q22.1 locus (*NEUROG3/ COL13A1/RPL5P26*, rs2642587, OR = 1.12 [1.08–1.16], $p = 8.4 \times 10^{-9}$) includes *NEUROG3* (Neurogenin3), which is an essential regulator of pancreatic endocrine cell differentiation[29]. Mutations in this gene have been reported to cause permanent neonatal diabetes, but a role of this gene in T2D has not been yet reported[30].

The lead common variant of the fourth novel locus at chromosome 12q24.31 (rs3794205, OR = 1.07 [1.04–1.10], $p = 4.1 \times 10^{-8}$) lies within an intron of the *CAMKK2* gene, previously implicated in cytokine-induced beta-cell death[31]. However, other variants within the corresponding credible set could also be causal, such as a missense variant within the *P2RX7*, a gene previously associated with glucose homeostasis in humans and mice[32], or another variant (rs11065504, r^2 with lead variant = 0.81) found to be associated with the regulation of the *P2RX4* gene in tibial artery and in whole blood, according to GTEx (Supplementary Data 9).

The fifth novel locus driven by common variants is located within 17q21.32 (rs12453394, OR = 1.07 [1.05–1.10], $p = 3.23 \times 10^{-8}$). It includes three missense variants located within the *CALCOCO2*, *SNF8*, and *GIP* genes. *GIP* encodes for glucose-dependent insulinotropic peptide, a hormonal mediator of enteral regulation of insulin secretion[33]. Variants in the GIP receptor (GIPR) have been previously associated with insulin response to oral glucose challenge and beta-cell function[34], proposing *GIP* as a plausible candidate effector gene of this locus[35].

A new T2D signal driven by a low-frequency variant. Furthermore, we selected all low-frequency ($0.01 \leq \text{MAF} < 0.05$) variants with $p \leq 1 \times 10^{-4}$ in the 70KforT2D meta-analysis that

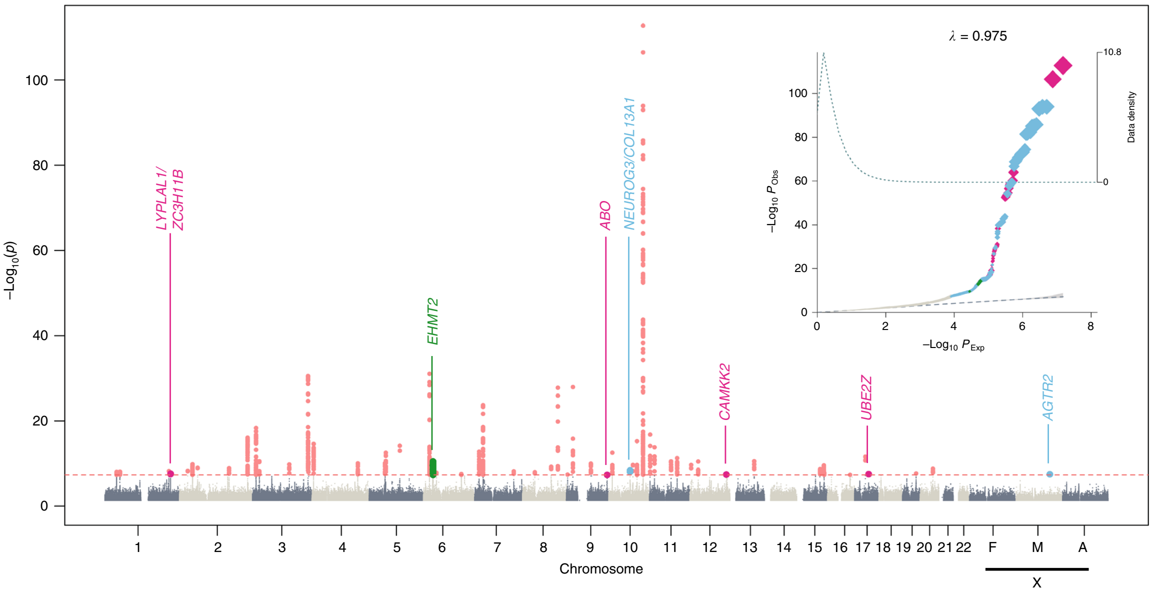


Fig. 2 Manhattan and quantile–quantile plot (QQ-plot) of the discovery and replication genome-wide meta-analysis. The upper corner represents the QQ-plot. Expected $-\log_{10}$ p-values under the null hypothesis are represented in the x axis, while observed $-\log_{10}$ p-values are represented in the y axis. Observed p-values were obtained according to the suitable replication dataset used (as shown in Fig. 1) and were depicted using different colors. HapMap variants were meta-analyzed using the trans-ethnic summary statistics from the DIAGRAM study and our meta-analysis based on the Genetic Epidemiology Research on Aging (GERA) cohort and the northwestern NuGENE project, and that resulted in novel associations depicted in magenta. The rest of non-HapMap variants meta-analyzed using the full 70KforT2D cohort are represented in gray, and the fraction of novel GWAS-significant variants is highlighted in light blue. Coding low-frequency variants meta-analyzed using the 70KforT2D and the T2D Portal data that resulted in novel GWAS-significant associations are depicted in green. The shaded area of the QQ-plot indicates the 95% confidence interval under the null and a density function of the distribution of the p-values was plotted using a dashed line. The λ is a measure of the genomic inflation and corresponds to the observed median $\chi 2$ test statistic divided by the median expected χ^2 test statistic under the null hypothesis. The Manhattan plot, representing the $-\log_{10}$ p-values, was colored as explained in the QQ-plot. All known GWAS-significant associated variants within known T2D genes are also depicted in red. X chromosome results for females (F), males (M), and all individuals (A) are also included

were annotated as altering protein sequences, according to VEP. This resulted in 15 coding variants that were meta-analyzed with exome array and whole-exome sequencing data from a total of ~97,000 individuals[6] after excluding the overlapping cohorts between the different data sets. This analysis highlighted a novel genome-wide association driven by a low-frequency missense variant (Ser58Phe) within the *EHMT2* gene at chromosome 6p21.33 (rs115884658, OR = 1.21 [1.14–1.29], $p = 3.00 \times 10^{-10}$; Fig. 2, Supplementary Figures 6 and 7). *EHMT2* is involved in the mediation of FOXO1 translocation induced by insulin[36]. Since this variant is less than 1 Mb away from *HLA-DQA1*, a locus reported to be associated with T2D[37,] we performed a series of reciprocal conditional analyses and excluded the possibility that our analysis was capturing previously reported T2D[8, 37] or T1D[38–40] signals (Supplementary Data 11). Beyond this missense *EHMT2* variant, other low-frequency variants within the corresponding credible set may also be causal. For example, rs115333512 (r^2 with lead variant = 0.28) is associated with the expression of *CLIC1* in several tissues according to GTEx (multitissue meta-analysis $p = 8.9 \times 10^{-16}$, Supplementary Data 9). In addition, this same variant is associated with the expression of the first and second exon of the *CLIC1* mRNA in pancreatic islet donors ($p(\text{exon } 1) = 1.4 \times 10^{-19}$, $p(\text{exon } 2) = 1.9 \times 10^{-13}$, Supplementary Data 10). Interestingly, *CLIC1* has been reported as a direct target of metformin by mediating the antiproliferative effect of this drug in human glioblastoma[41]. All these findings support *CLIC1,* as an additional possible effector transcript, likely driven by rs115333512.

A novel rare X chromosome variant associated with T2D. Similar to other complex diseases, the majority of published large-scale T2D GWAS studies have omitted the analysis of the X chromosome, with the notable exception of the identification of a T2D-associated region near the *DUSP9* gene in 2010[42]. To fill this gap, we tested the X chromosome genetic variation for association with T2D. To account for heterogeneity of the effects and for the differences in imputation performance between males and females, the association was stratified by sex and tested separately, and then meta-analyzed. This analysis was able to replicate the *DUSP9* locus, not only through the known rs5945326 variant (OR = 1.15, $p = 0.049$), but also through a three-nucleotide deletion located within a region with several promoter marks in liver (rs61503151 [GCCA/G], OR = 1.25, $p = 3.5 \times 10^{-4}$), and in high LD with the first reported variant ($r^2 = 0.62$). Conditional analyses showed that the originally reported variant was no longer significant (OR = 1.01, $p = 0.94$) when conditioning on the newly identified variant, rs61503151. On the other hand, when conditioning on the previously reported variant, rs5945326, the effect of the newly identified indel remained significant and with a larger effect size (OR = 1.33, $p = 0.003$), placing this deletion, as a more likely candidate causal variant for this locus (Supplementary Data 14).

In addition, we identified a novel genome-wide significant signal in males at the Xq23 locus driven by a rare variant (rs146662075, MAF = 0.008, OR = 2.94 [2.00–4.31], $p = 3.5 \times 10^{-8}$; Fig. 3a). Two other variants in LD with the top variant, rs139246371 (chrX:115329804, OR = 1.65, $p = 3.5 \times 10^{-5}$, $r^2 =$

0.37 with the top variant) and rs6603744 (chrX:115823966, OR = 1.28, $p = 1.7 \times 10^{-4}$, $r^2 = 0.1$ with the top variant), comprised the 99% credible set and supported the association. We tested in detail the accuracy of the imputation for the rs146662075 variant by comparing the imputed results from the same individuals genotyped by two different platforms (Methods) and found that the imputation was highly accurate in males only when using UK10K, but not in females, nor when using 1000G ($R^2_{[\text{UK10K,males}]} = 0.94$, $R^2_{[\text{UK10K,females}]} = 0.66$, $R^2_{[\text{1000G,males}]} = 0.62$, and $R^2_{[\text{1000G,females}]} = 0.43$; Supplementary Fig. 8). Whether this association is specific to men, or whether it also affects female carriers, remains to be clarified with datasets that allow accurate imputation on females, or with direct genotyping or sequencing.

To further validate and replicate this association, we next analyzed four independent data sets (SIGMA[6], INTERACT[43], Partners Biobank[44], and UK Biobank[45]), by performing imputation with the UK10K reference panel. In addition, a fifth cohort was genotyped de novo for the rs146662075 variant in several Danish sample sets. The initial meta-analysis, including the five replication data sets did not reach genome-wide significance (OR = 1.57, $p = 1.2 \times 10^{-5}$; Supplementary Fig. 9A), and revealed a strong degree of heterogeneity (heterogeneity $p_{\text{het}} = 0.004$), which appeared to be driven by the replication cohorts.

As a complementary replication analysis, within one of the case-control studies, there was a nested prospective cohort study, the Inter99, which consisted of 1,652 nondiabetic male subjects genotyped for rs146662075, of which 158 developed T2D after 11 years of follow-up. Analysis of incident diabetes in this cohort confirmed the association with the same allele, as previously seen in the case-control studies, with carriers of the rare T allele having increased risk of developing incident diabetes, compared to the C carriers (Cox-proportional hazards ratio (HR) = 3.17 [1.3–7.7], $p = 0.011$, Fig. 3b). Nearly 30% of carriers of the T risk allele developed incident T2D during 11 years of follow-up, compared to only 10% of noncarriers.

To understand the strong degree of heterogeneity observed after adding the replication datasets, we compared the clinical and demographic characteristics of the discovery and replication cohorts, and found that the majority of the replication datasets contained control subjects that were significantly younger than 55 years, the average age at the onset of T2D reported in this study and in Caucasian populations[46]. This was particularly clear for the Danish cohort (age controls [95%CI] = 46.9 [46.6–47.2] vs. age cases [95%CI] = 60.7 [60.4–61.0]) and for INTERACT (age controls [95%CI] = 51.7 [51.4–52.1] vs. age cases [95%CI] = 54.8 [54.6–55.1]; Supplementary Fig. 10). Given the supporting results with the Inter99 prospective cohort, we performed an additional analysis using a stricter definition of controls, to minimize the presence of prediabetics or individuals that may further develop diabetes after reaching the average age at the onset. For this, we applied two additional exclusion criteria: (i) subjects younger than 55 years and (ii), when possible, excluding individuals with measured 2-h plasma glucose values during oral glucose tolerance test (OGTT) above 7.8 mmol l^{-1}, a threshold employed to identify impaired glucose tolerance (prediabetes)[47], or controls with family history of T2D, both being strong risk factors for developing T2D. While the application of the first filter alone did not yield genome-wide significant results (Supplementary Fig. 9B), upon excluding individuals with prediabetes or a family history of T2D, the replication results were significant and consistent with the initial discovery results (OR = 1.57 [1.19–2.07], $p = 0.0014$). The combined analysis of the discovery and replication cohorts resulted in genome-wide significance, confirming the association of rs146662075 with T2D (OR = 1.95 [1.56–2.45], $p = 7.8 \times 10^{-9}$, Fig. 3c).

Allele-specific enhancer activity of the rs146662075 variant. We next explored the possible molecular mechanism behind this association, by using different genomic resources and experimental approaches. The credible set of this region contained three variants, with the leading SNP alone (rs146662075), showing 78% posterior probability of being causal (Supplementary Fig. 7, Supplementary Data 5), as well as the highest CADD (scaled *C*-score = 15.68; Supplementary Data 8), and LINSIGHT score (Supplementary Data 9). rs146662075 lies within a chromosomal region enriched in regulatory (DNase I) and active enhancer (H3K27ac) marks, between the *AGTR2* (at 103 kb) and the *SLC6A14* (at 150 kb) genes. The closest gene *AGTR2*, which encodes for the angiotensin II receptor type 2, has been previously associated with insulin secretion and resistance[48–50]. From the analysis of available epigenomic data sets[51], we found no evidences of H3K27ac or other enhancer regulatory marks in human pancreatic islets; whereas a significant association was observed between the presence of H3K27ac enhancer marks and the expression of *AGTR2* across multiple tissues (Fisher test $p = 4.45 \times 10^{-3}$), showing the highest signal of both H3K27ac and *AGTR2* RNA-seq expression, but not with other genes from the same topologically associated domain (TAD), in fetal muscle (Fig. 4a; Supplementary Figure 11).

We next studied whether the region encompassing the rs146662075 variant could act as a transcriptional enhancer and whether its activity was allele-specific. For this, we linked the DNA region with either the T (risk) or the C (non-risk) allele, to a minimal promoter and performed luciferase assays in a mouse myoblast cell line. The luciferase analysis showed an average 4.4-fold increased activity for the disease-associated T allele, compared to the expression measured with the common C allele, suggesting an activating function of the T allele, or a repressive function of the C allele (Fig. 4b). Consistent with these findings, electrophoretic mobility shift assays using nuclear protein extracts from mouse myoblast cell lines, differentiated myotubes, and human fetal muscle cell line, revealed sequence-specific binding activity of the C allele, but not the rare T allele (Fig. 4c). Overall, these data indicate that the risk T allele prevents the binding of a nuclear protein that is associated with decreased activity of an *AGTR2*-linked enhancer.

Discussion

Through harmonizing and reanalyzing publicly available T2D GWAS data, and performing genotype imputation with two whole-genome sequence-based reference panels, we are able to perform deeper exploration of the genetic architecture of T2D. This strategy allowed us to impute and test for association with T2D more than 15 million of high-quality imputed variants, including low-frequency, rare, and small insertions and deletions, across chromosomes 1–22 and X.

The reanalysis of these data confirmed a large fraction of already-known T2D loci, and identified novel potential causal variants by fine mapping and functionally annotating each locus.

This reanalysis also allowed us to identify seven novel associations, five driven by common variants in or near *LYPLAL1*, *NEUROG3*, *CAMKK2*, *ABO*, and *GIP;* a low-frequency variant in *EHMT2*, and a rare variant in the X chromosome. This rare variant identified in Xq23 chromosome was located near the *AGTR2* gene, and showed nearly twofold increased risk for T2D in males, which represents, to our knowledge, the largest effect size identified so far in Europeans, and a magnitude similar to other variants with large effects identified in other populations[52, 53].

Our study complemented other efforts that also aim at unraveling the genetics behind T2D through the generation of new

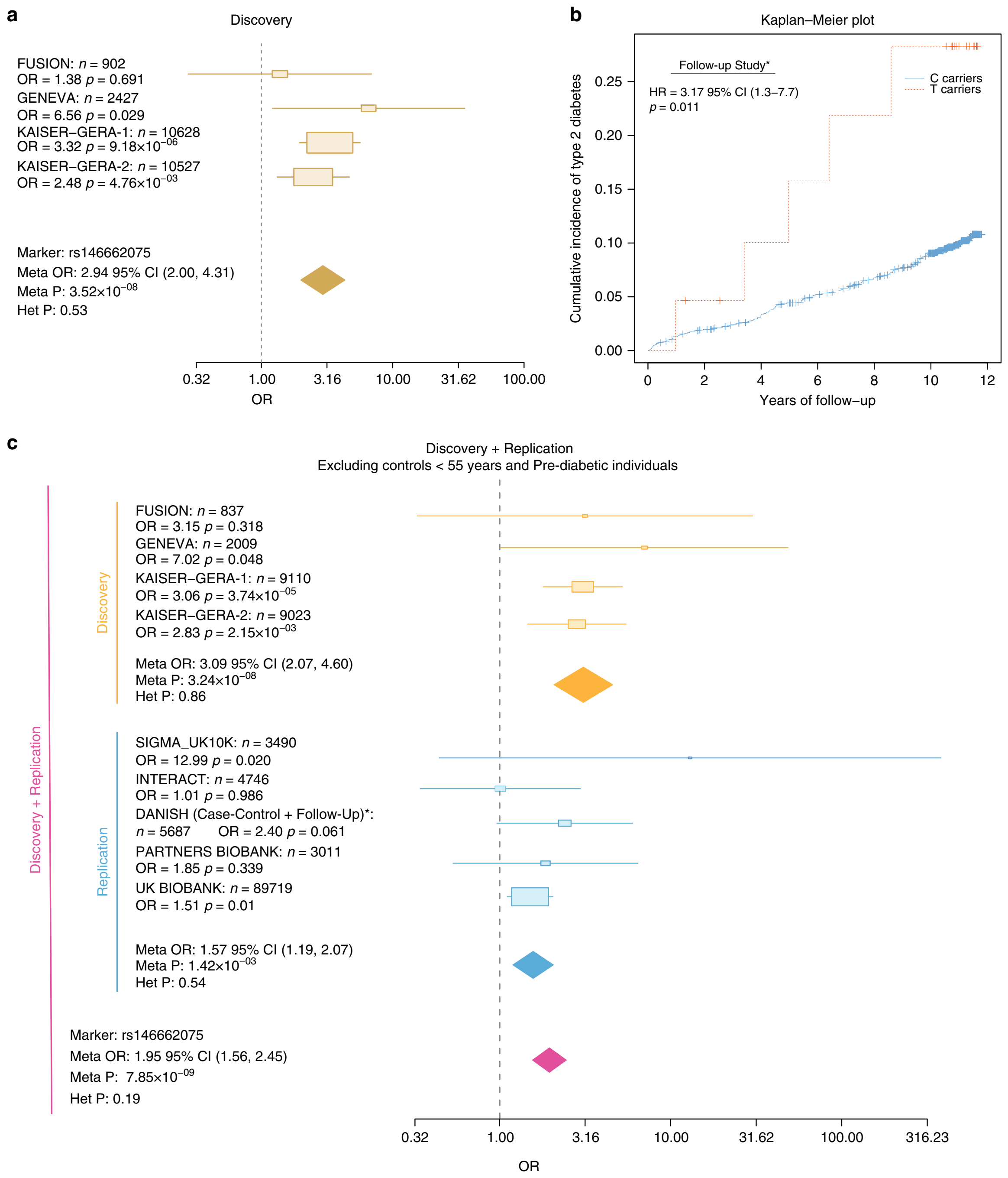
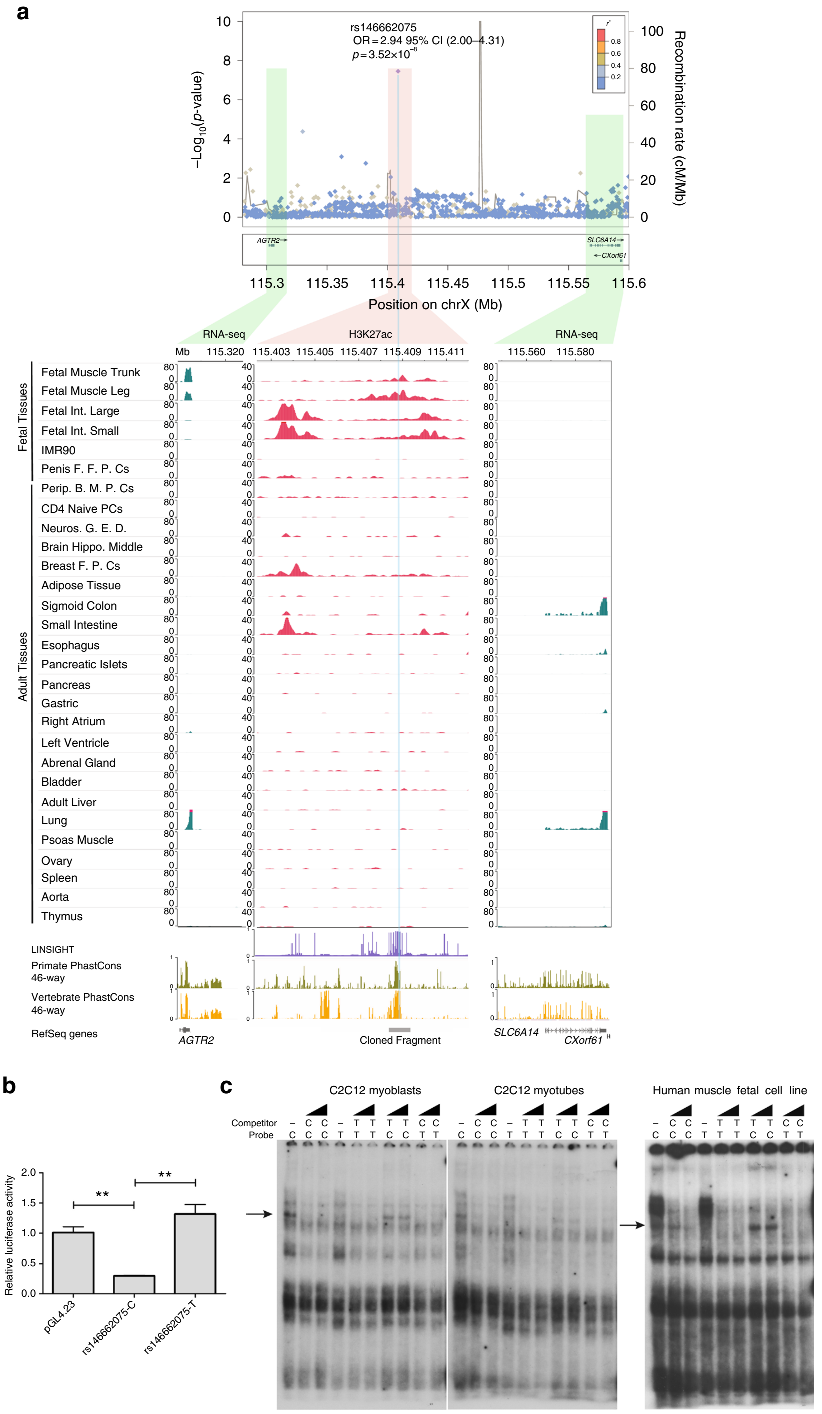


Fig. 3 Discovery and replication of rs14666075 association signal. **a** Forest plot of the discovery of rs146662075 variant. Cohort-specific odds ratios are denoted by boxes proportional to the size of the cohort and 95% CI error bars. The combined OR estimated for all the data sets is represented by a diamond, where the diamond width corresponds to 95% CI bounds. The *p*-value for the meta-analysis (Meta P) and for the heterogeneity (Het P) of odds ratio is shown. **b** Kaplan–Meier plot showing the cumulative incidence of T2D for a 11 years follow-up. The red line represents the T carriers and in light blue, C carriers are represented ($n = 1{,}652$, cases = 158). **c** Forest plot after excluding controls younger than 55 years, OGTT >7.8 mmol l^{-1}, and controls with family history of T2D in both the discovery and replication cohorts when available

genetic data[6, 54]. For example, we provided for the first time a comprehensive coverage of structural variants, which point to previously unobserved candidate causal variants in known and novel loci, as well as a comprehensive coverage of the X chromosome through sequence-based imputation.

This study also highlights the importance of a strict classification of both cases and controls, in order to identify rare variants associated with disease. Our initial discovery of the Xq23 locus was only replicated when the control group was restricted to T2D-free individuals who were older than 55 years (average age



at the onset of T2D), had normal glucose tolerance, and no family history of T2D. This is in line with previous results obtained for a T2D population-specific variant found in Inuit within the *TBC1D4* gene, which was only significant when using OGTT as criteria for classifying cases and controls, but not when using HbA1c[52]. Our observation that 30% of the rs146662075 risk allele carriers developed T2D over 11 years of follow-up, compared to 10% of noncarriers, further supports the association of this variant and suggests that an early identification of these subjects through genotyping may be useful to tailor pharmacological or lifestyle intervention to prevent or delay the onset of T2D.

Using binding and gene-reporter analyses, we demonstrated a functional role of this variant and proposed a possible mechanism behind the pathophysiology of T2D in T risk allele carriers, in which this rare variant could favor a gain of function of *AGTR2*, previously associated with insulin resistance[48]. *AGTR2* appears, therefore, as a potential therapeutic target for this disease, which would be in line with previous studies showing that the blockade of the renin–angiotensin system in mice[55] and in humans[56] prevents the onset of T2D, and restores normoglycemia[57, 58].

Overall, beyond our significant contribution toward expanding the number of genetic associations with T2D, our study also highlights the potential of the reanalysis of public data, as a complement to large studies that use newly generated data. This study informs the open debate in favor of data sharing and democratization initiatives[4, 59], for investigating the genetics and pathophysiology of complex diseases, which may lead to new preventive and therapeutic applications.

Methods

Quality filtering for imputed variants. In order to assess genotype imputation quality and to determine an accurate post-imputation quality filter, we made use of the Wellcome Trust Case Control Consortium (WTCCC)[40] data available through the European Genotype Archive (EGA, https://www.ebi.ac.uk/ega/studies/EGAS00000000028). The genotyping data and the subjects included in the following tests were filtered according to the guidelines provided by the WTCCC, whose criteria of exclusion are in line with standard quality filters for GWAS[60]. We used the 1958 British Birth cohort (~3,000 samples, 58C) that was genotyped by Affymetrix v6.0 and Illumina 1.2M chips. After applying the quality-filtering criteria, 2,706 and 2,699 subjects from the Affymetrix and Illumina data, respectively, were available for the 58C samples, leaving an intersection of 2,509 individuals genotyped by both platforms. After variant quality filtering and excluding all the variants with minor allele frequency (MAF) below 0.01, 717,556, and 892,516 variants remained for 58C Affymetrix and Illumina platforms, respectively.

We used a two-step genotype imputation approach based on prephasing the study genotypes into full haplotypes with SHAPEIT2[61] to ameliorate the computational burden required for genotype imputation through IMPUTE2[62]. We used the GTOOL software (http://www.well.ox.ac.uk/~cfreeman/software/gwas/gtool.html, version 0.7.5) to homogenize strand annotation by merging the imputed results obtained from each set of genotyped data. To ensure that there were no strand orientation issues, we excluded all C/G and A/T SNPs. To perform genotype imputation, we used two sequence-based reference panels: the 1000G Phase1 (June 2014) release[7] and the UK10K[2].

We evaluated genotype imputation for each reference panel considering 2,509 58C individuals that were genotyped by both independent genotyping platforms. Four scenarios were considered: (a) fraction of variants originally genotyped (GT) by both Illumina (IL) and Affymetrix (Affy) platforms (both GT), (b) variants genotyped by Affy, but not present in IL array (Affy GT), (c) variants genotyped by IL, but not present in the Affy array (IL GT), and (d) variants not typed in IL nor in the Affy arrays, and therefore, imputed from IL and Affy data sets (d). This last scenario comprised the largest fraction of variants.

As the individuals typed (and imputed) using Affy and IL SNPs as backbones were the same, we expected no statistical differences when comparing the allele and genotype frequencies with any of the variants. The quality of the imputed variants was evaluated using the allelic dosage R^2 correlation coefficient, between the genotype dosages estimated when imputing using Affy or IL as the backbone. The Affy GT and IL GT SNPs were used to evaluate the correspondence between the allelic dosage R^2 scores and the IMPUTE2 info scores for the imputed genotypes. The linear model, between the allelic dosage R^2 and the IMPUTE2-info, was used to set an info score threshold of 0.7, which corresponds to an allelic dosage R^2 of 0.5. The correlation between R^2 and info score was uniform across all reference panels and platforms.

The 70KforT2D resource. We collected genetic individual-level data for T2D case/control studies from five independent datasets, Gene Environment-Association Studies initiative [GENEVA], Wellcome Trust Case Control Consortium [WTCCC], Finland–United States Investigation of NIDDM Genetics [FUSION], Resource for Genetic Epidemiology Research on Aging [GERA], and the Northwestern NUgene project [NuGENE] publicly available in the dbGaP (http://www.ncbi.nlm.nih.gov/gap) and EGA (https://www.ebi.ac.uk/ega/home) public repositories, comprising a total of 13,201 cases and 59,656 controls (for the description of each cohort, see Supplementary Note 1 and Supplementary Data 1).

Each dataset was independently harmonized and quality controlled with a three-step protocol, including two stages of SNP removal and an intermediate stage of sample exclusion. The exclusion criteria for variants were (i) missing call rate ≥ 0.05, (ii) significant deviation from Hardy–Weinberg equilibrium (HWE) $p \le 1 \times 10^{-6}$ for controls and $p \le 1 \times 10^{-20}$ for the entire cohort, (iii) significant differences in the proportion of missingness between cases and controls $p \le 1 \times 10^{-6}$, and (iv) MAF < 0.01 (for the GERA cohort, we considered a MAF of 0.001). The exclusion criteria for samples were i) gender discordance between the reported and genetically predicted sex, ii) subject relatedness (pairs with $\pi \ge 0.125$ from which we removed the individual with the highest proportion of missingness), iii) missing call rates per sample ≥ 0.02, and iv) population structure showing more than four standard deviations within the distribution of the study population according to the first four principal components.

We performed genotype imputation independently for each cohort by prephasing the genotypes to whole haplotypes with SHAPEIT2 and then, we performed genotype imputation with IMPUTE2. We tested for association with additive logistic regression using SNPTEST, seven derived principal components sex, age, and body-mass index (BMI), except for WTCCC, for which age and BMI were not available (Supplementary Data 1). To maximize power and accuracy, we combined the association results from 1000G Phase1 integrated haplotypes (June, 2014)[7] and UK10K (http://www.uk10k.org/) reference panels by choosing for each variant, the reference panel that provided the best IMPUTE2 info score. For 1000G-based genotype imputation in chromosome X (chrX), we used the "v3. macGT1" release (August, 2012). For chrX, we restricted the analysis to non-pseudoautosomal (non-PAR) regions and stratified the association analysis by sex to account for hemizygosity for males, while for females, we followed an autosomal model. Also, we did not apply HWE filtering in the X chromosome variants. Finally, for the GERA cohort due to the large computational burden that comprises the whole genotype imputation process in such a large sample size, we randomly split this cohort into two homogeneous subsets of ~30,000 individuals each, in order to minimize the memory requirements.

We included variants with IMPUTE2 info score ≥ 0.7, MAF ≥ 0.001, and for autosomal variants, HWE controls $p > 1 \times 10^{-6}$. Further details about genotype imputation and covariate information used in association testing are summarized in Supplementary Data 1.

70KforT2D and inclusion of previous summary statistics data. We meta-analyzed the different sets from the 70KforT2D data set with METAL[63], using the inverse variance-weighted fixed effect model. We included variants with I^2 heterogeneity < 75. This filter was not applied to the final X chromosome data set, after meta-analyzing the results from males and females separately (which were already filtered by $I^2 < 75$).

For the meta-analysis with the DIAGRAM trans-ethnic study[8], we excluded from the whole 70KforT2D datasets those cohorts that overlapped with the DIAGRAM data. Therefore, we meta-analyzed the GERA and NuGENE cohorts (7,522 cases and 50,446 controls) from the 70KforT2D analysis with the trans-ethnic summary statistics results. As standard errors were not provided for the

Fig. 4 Functional characterization of rs146662075 association signal. **a** Signal plot for X chromosome region surrounding rs146662075. Each point represents a variant, with its *p*-value (on a −log10 scale, *y* axis) derived from the meta-analysis results from association testing in males. The *x* axis represents the genomic position (hg19). Below, representation of H3K27ac and RNA-seq in a subset of cell types is shown. The association between RNA-seq signals and H3K27ac marks suggests that *AGTR2* is the most likely regulated gene by the enhancer that harbors rs146662075. **b** The presence of the common allelic variant rs146662075-C reduces enhancer activity in luciferase assays performed in a mouse myoblast cell line. **c** Electrophoretic mobility shift assay in C2C12 myoblast cell lines, C2C12-differentiated myotubes, and human fetal myoblasts showed allele-specific binding of a ubiquitous nuclear complex. The arrows indicate the allele-specific binding event. Competition was carried out using 50- and 100-fold excess of the corresponding unlabeled probe

DIAGRAM trans-ethnic meta-analysis, we performed a sample size based meta-analysis, which converts the direction of the effect and the p-value into a Z-score. In addition, we also performed an inverse variance-weighted fixed effect meta-analysis to estimate the final effect sizes. This approach required the estimation of the beta and standard errors from the summary statistics (p-value and odds ratio).

For the meta-analysis of coding low-frequency variants with the Type 2 Diabetes Knowledge Portal (T2D Portal)[6], we included from the 70KforT2D data set the NuGENE and GERA cohorts (7,522 cases and 50,446 controls), to avoid overlapping samples. Like in the previous scenario, standard errors were not provided for the T2D Portal data and we used a sample size based meta-analysis with METAL. However, to estimate the effect sizes, we also calculated the standard errors from the p-values and odds ratios, and we performed an inverse variance-weighted fixed effect meta-analysis.

See further details about the cohorts in Supplementary Note 1.

Pathway and enrichment analysis. Summary statistics that resulted from the 70KforT2D meta-analysis were analyzed by Data-driven Expression-Prioritized Integration for Complex Traits (DEPICT)[9] to prioritize likely causal genes, to highlight enriched pathways, and to identify the most relevant tissues/cell types; DEPICT relies on publicly available gene sets (including molecular pathways) and leverages gene expression data from 77,840 gene expression arrays, to perform gene prioritization and gene-set enrichment based on predicted gene function and the so-called reconstituted gene sets. A reconstituted gene set contains a membership probability for each gene and conversely, each gene is functionally characterized by its membership probabilities across 14,461 reconstituted gene sets. As an input to DEPICT, we used all summary statistics from autosomal variants with $p < 1 \times 10^{-5}$ in the 70KforT2D meta-analysis. We used an updated version of DEPICT, which handled 1000G Phase1-integrated haplotypes (June 2014, www.broadinstitute.org/depict). DEPICT was run using 3,412 associated SNPs ($p < 1 \times 10^{-5}$), from which we identified independent SNPs using PLINK and the following parameters: --clump-p1 5e-8, --clump-p2 1e-5, --clump-r2 0.6, and --clump-kb 250. We used LD $r^2 > 0.5$ distance to define locus limits yielding 70 autosomal loci comprising 119 genes (note that this is not the same locus definition that we used elsewhere in the text). We ran DEPICT with default settings, i.e., using 500 permutations to adjust for bias and 50 replications to estimate false discovery rate (FDR). We used normalized expression data from 77,840 Affymetrix microarrays to reconstitute gene sets[9]. The resulting 14,461 reconstituted gene sets were tested for enrichment analysis. A total of 209 tissue or cell types expression data assembled from 37,427 Affymetrix U133 Plus 2.0 Array samples were used for enrichment analysis in tissue/cell-type expression. DEPICT identified 103 reconstituted gene sets significantly enriched (FDR < 5%) for genes found among the 70 loci associated to T2D. We did not consider reconstituted sets in which genes of the original gene set were not nominally enriched (Wilcoxon rank-sum test), as these are expected to be enriched in the reconstituted gene set by design. The lack of enrichment makes the interpretation of the reconstituted gene set challenging because the label of the reconstituted gene set will not be accurate. Hence, the following reconstituted gene sets were removed from the results (Wilcoxon rank sum and P-values in parentheses): MP:0004247 gene set ($p = 0.73$), GO:0070491 gene set ($p = 0.14$), MP:0004086 gene set ($p = 0.17$), MP:0005491 gene set ($p = 0.54$), GO:0005159 gene set ($p = 0.04$), MP:0005666 gene set ($p = 0.05$), ENSG00000128641 gene set ($p = 0.02$), MP:0006344 gene set ($p = 0.42$), MP:0004188 gene set ($p = 0.22$), MP:0002189 gene set ($p = 0.02$), MP:0000003 gene set ($p = 0.08$), ENSG00000116604 gene set ($p = 0.13$), GO:0005158 gene set ($p = 0.07$), and MP:0001715 gene set ($p = 0.01$). After applying the filters described above, there were 89 significantly enriched reconstituted gene sets. We used the affinity propagation tool to cluster related reconstituted gene sets (network diagram script available from https://github.com/perslab/DEPICT).

We also used the VSE R package to compute the enrichment or depletion of genetic variants comprised in the 57 credible sets listed in Supplementary Data 5 across regulatory genomic annotations, as described in[64]. Each GWAS lead variant from the final meta-analysis was considered as a tag SNP and variants from the corresponding 99% credible set (Supplementary Data 5) in LD with the tag SNP ($R^2 \geq 0.4$), as a cluster or associated variant set (AVS). In order to account for the size and structure of the AVS, a null distribution was built based on random permutations of the AVS. Each permuted variant set was matched to the original AVS, cluster by cluster using HapMap data by size and structure. This Matched Random Variant Set (MRVS) was calculated using 500 permutations. Significant enrichments or depletions were considered when the Bonferroni-adjusted p-value was < 0.01. Human islet regulatory elements (C1–C5) were obtained from[10].

Definition of 99% credible sets of GWAS-significant loci. For each genome-wide significant region locus, we identified the fraction of variants that have, in aggregate, 99% probability of containing the causal T2D-associated variant. By using our 70KforT2D meta-analysis based on imputed data (NuGENE, GERA, FUSION, GENEVA, and WTCCC data sets, comprising 12,231 cases and 57,196 controls), we defined the 99% credible set of variants for each locus with a Bayesian refinement approach[11] (we considered variants with an $R^2 > 0.1$ with their respective leading SNP).

Credible sets of variants are analogous to confidence intervals as we assume that the credible set for each associated region contains, with 99% probability, the true causal SNP if this has been genotyped or imputed. The credible set construction provides, for each variant placed within a certain associated locus, a posterior probability of being the causal one[11]. We estimated the approximate Bayes' factor (ABF) for each variant as

$$\mathrm{ABF} = \sqrt{1 - r}\, e^{(rz^2/2)},$$

where

$$r = \frac{0.04}{(\mathrm{SE}^2 + 0.04)},$$

$$z = \frac{\beta}{\mathrm{SE}}.$$

The β and the SE are the estimated effect size and the corresponding standard error resulting from testing for association under a logistic regression model. The posterior probability for each variant was obtained as

$$\text{Posterior Probability}_i = \frac{\mathrm{ABF_i}}{T},$$

where ABF_i corresponds to the approximate Bayes' factor for the marker i and T represents the sum of all the ABF values from the candidate variants enclosed in the interval being evaluated. This calculation assumes that the prior of the β corresponds to a Gaussian with mean 0 and variance 0.04, which is also the same prior commonly employed by SNPTEST, the program being used for calculating single-variant associations.

Finally, we ranked variants according to the ABF (in decreasing order) and from this ordered list, we calculated the cumulative posterior probability. We included variants in the 99% credible set of each region until the SNP that pushed the cumulative posterior probability of association over 0.99.

The 99% credible sets of variants for each of the 57 GWAS-significant regions are summarized in Supplementary Data 5.

Characterization of indels. We examined whether indels from the 99% credible sets were present or absent in the 1000G Phase1 or UK10K reference panels, and also checked whether they were present or not in the 1000G Phase3 reference panel. All the information has been summarized in Supplementary Data 6. We also visually inspected the aligned BAM files of the most relevant indels from both projects to discard that they could be alignment artifacts.

Functional annotation of the 99% credible set variants. To determine the effect of 99% credible set variants on genes, transcripts, and protein sequence, we used the variant effect predictor (VEP, GRCh37.p13 assembly)[13]. The VEP application determines the effect of variants (SNPs, insertions, deletions, CNVs, or structural variants) on genes, transcripts, proteins, and regulatory regions. We used as input the coordinates of variants within 99% credible sets and the corresponding alleles, to find out the affected genes and RefSeq transcripts and the consequence on the protein sequence by using the GRCh37.p13 assembly. We also manually checked all these annotations with the Exome Aggregation Consortium data set (ExAC, http://exac.broadinstitute.org) and the most updated VEP server based on the GRCh38.p7 assembly. All these annotations are provided in Supplementary Data 7.

We used combined annotation-dependent depletion (CADD) scoring function to prioritize functional, deleterious, and disease causal variants. We obtained the scaled C-score (PHRED-like scaled C-score ranking each variant with respect to all possible substitutions of the human genome) metric for each 99% credible set variant, as it highly ranks causal variants within individual genome sequences[14] (Supplementary Data 8). We also used the LINSIGHT score to prioritize functional variants, which measures the probability of negative selection on noncoding sites by combining a generalized linear model for functional genomic data with a probabilistic model of molecular evolution[15]. For each credible set variant, we retrieved the precomputed LINSIGHT score at that particular nucleotide site, as well as the mean LINSIGHT precomputed score for a region of 20 bp centered on each credible set variant, respectively (https://github.com/CshlSiepelLab/LINSIGHT). These metrics are summarized in Supplementary Data 9.

In order to prioritize functional regulatory variants, we used the V6 release from the GTEx data that provides gene-level expression quantifications and eQTL results based on the annotation with GENCODE v19. This release included 450 genotyped donors, 8,555 RNA-seq samples across 51 tissues, and two cell lines, which led to the identification of eQTLs across 44 tissues[16]. Moreover, RNA-seq data from human pancreatic islets from 89 deceased donors cataloged as eQTLs and exon use (sQTL) were also integrated with the GWAS data to prioritize candidate regulatory variants[17] but in pancreatic islets, which is a target tissue for T2D. Both analyses are summarized in Supplementary Data 10 and Supplementary Data 11, respectively.

Conditional analysis. To confirm the independence between novel loci and previously known T2D signals, we performed reciprocal conditional analyses (Supplementary Data 5, Supplementary Data 12, Supplementary Data 13, and Supplementary Data 14). We included the conditioning SNP as a covariate in the logistic regression model, assuming that every residual signal that arises corresponds to a secondary signal independent from this conditioning SNP. We applied this method to the *EHMT2* locus (less than 1Mb away from the *HLA* where T2D and T1D signals have been identified), to confirm that this association was independent of previously reported T2D signals and also to discard that this association is also driven by possible contamination of T1D diagnosed as T2D cases. We conditioned on the top variant identified in this study and the top variant from the 99% credible set analysis, but also on the top variants previously described for T2D and T1D[8, 38–40]. For this purpose, we used the full 70KforT2D resource (NuGENE, GERA, FUSION, GENEVA, and WTCCC cohorts imputed with 1000G and UK10K reference panels). Finally, all the results were meta-analyzed as explained in previous sections. These analyses are provided in Supplementary Data 13. This approach was also applied to confirm that the novel *CAMKK2* signal at rs3794205 is independent of known T2D signals at the *HNF1A* locus (rs1169288, rs1800574, and chr12:121440833:D)[54], which is summarized in Supplementary Data 12. Moreover, this approach confirmed known secondary signals in the 9p21 locus[65] which allowed us to build 99% credible sets based on the results from the conditional analyses (included in Supplementary Data 5), and allowed us to identify the most likely causal variant for the *DUSP9* locus (Supplementary Data 14).

Replication of the rare variant association at Xq23. To replicate the association of the rs146662075 variant, we performed genotype imputation with the UK10K reference panel in four independent data sets: the InterAct case-cohort study[43], the Slim Initiative in Genomic Medicine for the Americas (SIGMA) consortium GWAS data set[6], the Partners HealthCare Biobank (Partners Biobank) data set[44], and the UK Biobank cohort[45]. Phasing was performed with SHAPEIT2 and the IMPUTE2 software was used for genotype imputation.

The current UK Biobank data release did not contain imputed data for the X chromosome, for which phasing and imputation had to be analyzed in-house. The data release used comprises X chromosome QCed genotypes of 488,377 participants, which were assayed using two arrays sharing 95% of marker content (Applied Biosystems$^{\mathrm{TM}}$ UK BiLEVE Axiom$^{\mathrm{TM}}$ Array and the Applied Biosystems$^{\mathrm{TM}}$ UK Biobank Axiom$^{\mathrm{TM}}$ Array). We included samples and markers that were used as input for phasing by UK Biobank investigators. At the sample level, we also excluded women, individuals with missing call rate > 5% or showing gender discordance between the reported and the genetically predicted sex. At the variant level, we excluded markers with MAF < 0.1% and with missing call rate > 5%. The final set of 16,463 X chromosome markers and 222,725 male individuals was split into six subsets due to the huge computational burden that would require phasing into whole haplotypes the entire data set. We also excluded indels, variants with MAF < 1%, and variants showing deviation of Hardy–Weinberg equilibrium with $p < 1 \times 10^{-20}$ before the imputation step. In addition, from those pairs of relatives reported to be third degree or higher according to UK Biobank, we excluded from each pair the individual with the lowest call rate. We then tested the rs146662075 variant for association with type 2 diabetes using SNPTEST v2.5.1 and the threshold method. To avoid contamination from other types of diabetes mellitus, we excluded from the entire sample data set, individuals with ICD10 codes falling in any of these categories: E10 (insulin-dependent diabetes mellitus), E13 (other specified diabetes mellitus), and E14 (unspecified diabetes mellitus). Then, we designated as T2D cases those individuals with E11 (non-insulin-dependent diabetes mellitus) ICD10 codes, and the rest as controls. Moreover, we only kept as control subjects those individuals without reported family history of diabetes mellitus and older than 55 years, which is the average age at the onset of T2D.

We also genotyped de novo the rs146662075 variant with KASPar SNP genotyping system (LGC Genomics, Hoddeson, UK) in the Danish cohort, which comprises data from five sample sets (Supplementary Note 2 also for the genotyping and QC analysis for this variant).

We used Cox-proportional hazard regression models to assess the association of the variant with the risk of incident T2D in 1,652 nondiabetic male subjects genotyped in the Inter99 cohort (part of the Danish cohort) that were followed for 11 years on average. The follow-up analysis was restricted to male individuals younger than 45 years who were 56 years old after 11 years of follow-up. Individuals with self-reported diabetes at the baseline examination and individuals present in the Danish National diabetes registry before the baseline examination were also excluded. To include the follow-up study as a part of the replication cohorts, we used a meta-analysis method that accounts for overlapping samples (MAOS)[66], as we had to control for the sample overlap between the follow-up and the case-control study from the Danish samples.

See Supplementary Note 2 for a larger description of each of the five replication cohorts and how they have been processed.

We meta-analyzed the association results from these five replication data sets with the 70KforT2D data sets. In the final meta-analysis, we excluded whenever it was possible (a) controls younger than 55 years and (b) with OGTT > 7.8 mmol l^{-1} or with family history of T2D.

In silico functional characterization of rs146662075. This variant is located in an intergenic region, flanked by *AGTR2* and *SLC6A14* genes, and within several DNase I hypersensitive sites. We searched for regulatory marks (i.e., H3K4me1 and H3K27ac marks) through the HaploReg web server (http://archive.broadinstitute.org/mammals/haploreg/haploreg.php), in order to assess which type of regulatory element was associated with the rs146662075 variant.

To further evaluate the putative regulatory role of rs146662075, we used the WashU EpiGenome Browser (http://epigenomegateway.wustl.edu/browser/, last access on June 2016). We used the following public data hubs: (1) the reference human epigenomes from the Roadmap Epigenomics Consortium track hubs and (2) the Roadmap Epigenomics Integrative Analysis Hub. These data were released by the NIH Roadmap Epigenomics Mapping Consortium[51]. RNA-seq data were used to evaluate whether gene expression of any of the closest genes (*AGTR2* and *SLC6A14* genes, fixed scale at 80 RPKM) correlated with the presence of the H3K27ac enhancer marks (a more strict mark for active enhancers in contrast with H3K4me1[67], which were highlighted by the HaploReg search) at the rs146662075 location. For visualizing the H3K27ac marks around rs146662075, we focused on a region of 8 kb and we used a fixed scale at 40 $-\log_{10}$ Poisson *p*-value of the counts relative to the expected background count (λ_{local}).

The NIH Roadmap Epigenomics Consortium data from standardized epigenomes also allowed us to further interrogate which target gene within the same topologically associating domain (TAD) was more likely to be regulated by this rs146662075 enhancer. We used H3K27ac narrow peaks from 59 tissues called using MACSv2 with a *p*-value threshold of 0.01 from 98 consolidated epigenomes to seek for enhancer marks in a given tissue (the presence of H3K27ac peak). To assess gene expression for any of the putative target genes within TAD, we used the RPKM expression matrix for 57 consolidated epigenomes (http://egg2.wustl.edu/roadmap/data/byDataType/rna/) and gene expression quantifications for fetal muscle leg, fetal muscle trunk, and fetal stomach provided by ENCODE (https://www.encodeproject.org/). With this, we were able to test for each of the genes, the association between gene expression and enhancer activity in 31 tissues with a Fisher's exact test.

Allele-specific enhancer activity at rs146662075. The mouse C2C12 cell line (ATCC CRL-1772) was grown in DMEM medium supplemented with 10% FBS and was induced to differentiate in DMEM with 10% horse serum for 4 days.

The human fetal myoblast cell line was established by Prof. Giulio Cossu (Institute of Inflammation and Repair, University of Manchester)[68]. The authors played no role in the procurement of the tissue. Cells were cultured in DMEM medium supplemented with 10% fetal calf serum and was induced to differentiate in DMEM with 2% horse serum for 4 days.

To perform an electrophoretic mobility shift assay, nuclear extracts from mouse myoblast C2C12 cells and the human myoblast cell line (ATCC CRL-1772) were obtained as described before[69]. Double-stranded oligonucleotides containing either the common or rare variants of rs146662075 were labeled using dCTP [α-32P] (Perkin Elmer). Oligonucleotide sequences are as follows (SNP location is underlined): probe-C-F: 5′-gatcTTTGAACACcGAGGGGAAAAT-3′ and R:5′-gatcATTTTCCCCTC gGTGTTCAAA-3′ and probe-T-F: 5′- gatcTTTGAACACtGAGGGGAAAAT-3′ and R: 5′-gatcATTTTCCCCTCaGTGTTCAAA-3′. Assay specificity was assessed by preincubation of nuclear extracts with 50- and 100-fold excess of unlabeled wild-type or mutant probes, followed by electrophoresis on a 5% nondenaturing polyacrylamide gel. Findings were confirmed by repeating binding assays on separate days.

For evaluating if the activity of the rs146662075 enhancer was allele specific, we performed a luciferase assay. A region of 969 bp surrounding rs146662075 was amplified from human genomic DNA using F: 5′-GCTAGCATATGGAGGTGATTTGT-3′ and R: 5′-GGCACTTCCTTCTCTGGTAGA-3′ oligonucleotides and cloned into pENTR/D-TOPO (Invitrogen). Allelic variant rs146662075T was introduced by site-directed mutagenesis using the following primers: F: 5′-CCTTTTTTTACTTTGAACACTGAGGGGAAAATCATGCTTGGC-3′ and R: 5′-GCCAAGCATGATTTTCCCCTCAGTGTTCAAAGTAAAAAAAGG-3′. Enhancer sequences were shuttled into pGL4.23[luc2/minP] vector (Promega) adapted for Gateway cloning (pGL4.23-GW, **2**) using Gateway LR Clonase II Enzyme mix (Invitrogen). Correct cloning was confirmed both by Sanger sequencing and restriction digestion.

C2C12 (ATCC CRL-1772) and 293T (ATCC CRL-3216) cells were transfected in quadruplicates with 500 ng of pGL4.23-GW enhancer containing vectors and 0.2 ng of Renilla normalizer plasmid. Transfections were carried out in 24-well plates using Lipofectamine 2000 and Opti-MEM (Thermo Fisher Scientific) following the manufacturer's instructions. Luciferase activity was measured 48 h after transfection using Dual-Luciferase Reporter Assay System (Promega). Firefly luciferase activity was normalized to Renilla luciferase activity, and the results were expressed as a normalized ratio to the empty pGL4.23[luc2/minP] vector backbone. Experiments were repeated three times. Statistical significance was evaluated through a Student's *t*-test.

Data availability. The association results are available at the Type 2 Diabetes Knowledge portal (www.type2diabetesgenetics.org/) and the complete summary statistics are available for download at http://cg.bsc.es/70kfort2d/

Received: 13 April 2017 Accepted: 24 November 2017
Published online: 22 January 2018

References

1. Welter, D. et al. The NHGRI GWAS catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001–D1006 (2014).
2. Huang, J. et al. Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat. Commun.* **6**, 8111 (2015).
3. Tukiainen, T. et al. Chromosome X-wide association study identifies loci for fasting insulin and height and evidence for incomplete dosage compensation. *PLoS Genet.* **10**, e1004127 (2014).
4. Flannick, J. & Florez, J. C. Type 2 diabetes: genetic data sharing to advance complex disease research. *Nat. Rev. Genet.* **17**, 535–549 (2016).
5. Steinthorsdottir, V. et al. Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. *Nat. Genet.* **46**, 294–298 (2014).
6. Fuchsberger, C. et al. The genetic architecture of type 2 diabetes. *Nature* **536**, 41–47 (2016).
7. Abecasis, G. R. et al. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
8. DIAbetes Genetics Replication and Meta-analysis (DIAGRAM) Consortium et al. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat. Genet.* **46**, 234–244 (2014).
9. Pers, T. H. et al. Biological interpretation of genome-wide association studies using predicted gene functions. *Nat. Commun.* **6**, 5890 (2015).
10. Pasquali, L. et al. Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nat. Genet.* **46**, 136–143 (2014).
11. Wellcome Trust Case Control Consortium et al. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat. Genet.* **44**, 1294–1301 (2012).
12. Diabetes Genetics Initiative of Broad Institute of Harvard and MIT et al. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* **316**, 1331–1336 (2007).
13. McLaren, W. et al. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069–2070 (2010).
14. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
15. Huang, Y. F., Gulko, B. & Siepel, A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat. Genet.* **49**, 618–624 (2017).
16. Mele, M. et al. Human genomics. The human transcriptome across tissues and individuals. *Science* **348**, 660–665 (2015).
17. Fadista, J. et al. Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism. *Proc. Natl Acad. Sci. USA* **111**, 13924–13929 (2014).
18. Manning, A. K. et al. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nat. Genet.* **44**, 659–669 (2012).
19. Randall, J. C. et al. Sex-stratified genome-wide association studies including 270,000 individuals show sexual dimorphism in genetic loci for anthropometric traits. *PLoS Genet.* **9**, e1003500 (2013).
20. Berndt, S. I. et al. Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. *Nat. Genet.* **45**, 501–512 (2013).
21. Dastani, Z. et al. Novel loci for adiponectin levels and their influence on type 2 diabetes and metabolic traits: a multi-ethnic meta-analysis of 45,891 individuals. *PLoS Genet.* **8**, e1002607 (2012).
22. Fox, C. S. et al. Genome-wide association for abdominal subcutaneous and visceral adipose reveals a novel locus for visceral fat in women. *PLoS. Genet.* **8**, e1002695 (2012).
23. Speliotes, E. K. et al. Genome-wide association analysis identifies variants associated with nonalcoholic fatty liver disease that have distinct effects on metabolic traits. *PLoS Genet.* **7**, e1001324 (2011).
24. Lei, X., Callaway, M., Zhou, H., Yang, Y. & Chen, W. Obesity associated Lyplal1 gene is regulated in diet induced obesity but not required for adipocyte differentiation. *Mol. Cell. Endocrinol.* **411**, 207–213 (2015).
25. Wessel, J. et al. Low-frequency and rare exome chip variants associate with fasting glucose and type 2 diabetes susceptibility. *Nat. Commun.* **6**, 5897 (2015).
26. the CARDIoGRAMplusC4D Consortium. A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat Genet.* **47**, 1121–1130 (2015).
27. Yamamoto, F., Clausen, H., White, T., Marken, J. & Hakomori, S. Molecular genetic basis of the histo-blood group ABO system. *Nature* **345**, 229–233 (1990).
28. Fagherazzi, G., Gusto, G., Clavel-Chapelon, F., Balkau, B. & Bonnet, F. ABO and Rhesus blood groups and risk of type 2 diabetes: evidence from the large E3N cohort study. *Diabetologia* **58**, 519–522 (2015).
29. Gradwohl, G., Dierich, A., LeMeur, M. & Guillemot, F. Neurogenin3 is required for the development of the four endocrine cell lineages of the pancreas. *Proc. Natl Acad. Sci. USA* **97**, 1607–1611 (2000).
30. Rubio-Cabezas, O. et al. Permanent neonatal diabetes and enteric anendocrinosis associated with biallelic mutations in NEUROG3. *Diabetes* **60**, 1349–1353 (2011).
31. Beck, A. et al. An siRNA screen identifies transmembrane 7 superfamily member 3 (TM7SF3), a seven transmembrane orphan receptor, as an inhibitor of cytokine-induced death of pancreatic beta cells. *Diabetologia* **54**, 2845–2855 (2011).
32. Todd, J. N. et al. Variation in glucose homeostasis traits associated with P2RX7 polymorphisms in mice and humans. *J. Clin. Endocrinol. Metab.* **100**, E688–E696 (2015).
33. Hinke, S. A., Hellemans, K. & Schuit, F. C. Plasticity of the beta cell insulin secretory competence: preparing the pancreatic beta cell for the next meal. *J. Physiol.* **558**, 369–380 (2004).
34. Saxena, R. et al. Genetic variation in GIPR influences the glucose and insulin responses to an oral glucose challenge. *Nat. Genet.* **42**, 142–148 (2010).
35. Lyssenko, V. et al. Pleiotropic effects of GIP on islet function involve osteopontin. *Diabetes* **60**, 2424–2433 (2011).
36. Arai, T., Kano, F. & Murata, M. Translocation of forkhead box O1 to the nuclear periphery induces histone modifications that regulate transcriptional repression of PCK1 in HepG2 cells. *Genes. Cells* **20**, 340–357 (2015).
37. Cook, J. P. & Morris, A. P. Multi-ethnic genome-wide association study identifies novel locus for type 2 diabetes susceptibility. *Eur. J. Hum. Genet.* **24**, 1175–1180 (2016).
38. Barrett, J. C. et al. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat. Genet.* **41**, 703–707 (2009).
39. Hakonarson, H. et al. A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene. *Nature* **448**, 591–594 (2007).
40. Wellcome Trust Case Control Consortium et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
41. Gritti, M. et al. Metformin repositioning as antitumoral agent: selective antiproliferative effects in human glioblastoma stem cells, via inhibition of CLIC1-mediated ion current. *Oncotarget* **5**, 11252–11268 (2014).
42. Voight, B. F. et al. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat. Genet.* **42**, 579–589 (2010).
43. Langenberg, C. et al. Gene-lifestyle interaction and type 2 diabetes: the EPIC interact case-cohort study. *PLoS. Med.* **11**, e1001647 (2014).
44. Karlson, E. W., Boutin, N. T., Hoffnagle, A. G. & Allen, N. L. Building the partners healthcare biobank at partners personalized medicine: informed consent, return of research results, recruitment lessons and operational considerations. *J Pers Med* **6**, 2 (2016).
45. Bycroft, C. et al. Genome-wide genetic data on ~500,000 UK Biobank participants. Preprint at *bioRxiv* https://doi.org/10.1101/166298 (2017).
46. Becerra, M. B. & Becerra, B. J. Disparities in age at diabetes diagnosis among Asian Americans: Implications for early preventive measures. *Prev. Chronic Dis.* **12**, E146 (2015).
47. Bartoli, E., Fra, G. P. & Carnevale Schianca, G. P. The oral glucose tolerance test (OGTT) revisited. *Eur. J. Intern. Med.* **22**, 8–12 (2011).
48. Shao, C., Zucker, I. H. & Gao, L. Angiotensin type 2 receptor in pancreatic islets of adult rats: a novel insulinotropic mediator. *Am. J. Physiol. Endocrinol. Metab.* **305**, E1281–E1291 (2013).
49. Yvan-Charvet, L. et al. Deletion of the angiotensin type 2 receptor (AT2R) reduces adipose cell size and protects from diet-induced obesity and insulin resistance. *Diabetes* **54**, 991–999 (2005).
50. Liu, M., Jing, D., Wang, Y., Liu, Y. & Yin, S. Overexpression of angiotensin II type 2 receptor promotes apoptosis and impairs insulin secretion in rat insulinoma cells. *Mol. Cell. Biochem.* **400**, 233–244 (2015).
51. Roadmap Epigenomics, C. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
52. Moltke, I. et al. A common greenlandic TBC1D4 variant confers muscle insulin resistance and type 2 diabetes. *Nature* **512**, 190–193 (2014).
53. Sigma Type 2 Diabetes Consortium. et al. Association of a low-frequency variant in HNF1A with type 2 diabetes in a Latino population. *JAMA* **311**, 2305–2314 (2014).
54. Gaulton, K. J. et al. Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. *Nat. Genet.* **47**, 1415–1425 (2015).
55. Frantz, E. D., Crespo-Mascarenhas, C., Barreto-Vianna, A. R., Aguila, M. B. & Mandarim-de-Lacerda, C. A. Renin-angiotensin system blockers protect pancreatic islets against diet-induced obesity and insulin resistance in mice. *PLoS ONE* **8**, e67192 (2013).
56. Leung, P. S. Mechanisms of protective effects induced by blockade of the renin-angiotensin system: novel role of the pancreatic islet angiotensin-generating system in Type 2 diabetes. *Diabet. Med.* **24**, 110–116 (2007).
57. Geng, D. F., Jin, D. M., Wu, W., Liang, Y. D. & Wang, J. F. Angiotensin converting enzyme inhibitors for prevention of new-onset type 2 diabetes

mellitus: a meta-analysis of 72,128 patients. *Int. J. Cardiol.* **167**, 2605–2610 (2013).

58. Investigators, D. T. et al. Effect of ramipril on the incidence of diabetes. *N. Engl. J. Med.* **355**, 1551–1562 (2006).
59. The ups and downs of data sharing in science. *Nature* **534**, 435-436 (2016).
60. Anderson, C. A. et al. Data quality control in genetic case-control association studies. *Nat. Protoc.* **5**, 1564–1573 (2010).
61. Delaneau, O., Zagury, J. F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* **10**, 5–6 (2013).
62. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529 (2009).
63. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
64. Cowper-Sal lari, R. et al. Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. *Nat. Genet.* **44**, 1191–1198 (2012).
65. Shea, J. et al. Comparing strategies to fine-map the association of common SNPs at chromosome 9p21 with type 2 diabetes and myocardial infarction. *Nat. Genet.* **43**, 801–805 (2011).
66. Lin, D. Y. & Sullivan, P. F. Meta-analysis of genome-wide association studies with overlapping subjects. *Am. J. Hum. Genet.* **85**, 862–872 (2009).
67. Creyghton, M. P. et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl Acad. Sci. USA* **107**, 21931–21936 (2010).
68. Cossu, G., Cicinelli, P., Fieri, C., Coletta, M. & Molinaro, M. Emergence of TPA-resistant 'satellite' cells during muscle histogenesis of human limb. *Exp. Cell. Res.* **160**, 403–411 (1985).
69. Boj, S. F., Parrizas, M., Maestro, M. A. & Ferrer, J. A transcription factor regulatory circuit in differentiated pancreatic cells. *Proc. Natl Acad. Sci. USA* **98**, 14481–14486 (2001).

Acknowledgements

This work has been sponsored by the grant SEV-2011-00067 of Severo Ochoa Program, awarded by the Spanish Government. This work was supported by an EFSD/Lilly research fellowship. Josep M. Mercader was supported by Sara Borrell Fellowship from the Instituto Carlos III and Beatriu de Pinós fellowship from the Agency for Management of University and Research Grants (AGAUR). Sílvia Bonàs was FI-DGR Fellowship from FI-DGR 2013 from Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR, Generalitat de Catalunya). This study makes use of data generated by the WTCCC. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113. This study also makes use of data generated by the UK10K Consortium, derived from samples from UK10K COHORT IMPUTATION (EGAS00001000713). A full list of the investigators who contributed to the generation of the data is available in www.UK10K.org. Funding for UK10K was provided by the Wellcome Trust under award WT091310. We acknowledge PRACE for awarding us to access MareNostrum supercomputer, based in Spain at Barcelona. The technical support group, particularly Pablo Ródenas and Jorge Rodríguez, from the Barcelona Supercomputing Center is gratefully acknowledged. This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 667191. Mercè Planas-Fèlix is funded by the Obra Social Fundación la Caixa fellowship under the Severo Ochoa 2013 program. Work from Irene Miguel-Escalada, Ignasi Moran, Goutham Atla, and Jorge Ferrer was supported by the National Institute for Health Research (NIHR) Imperial Biomedical Research Centre, the Wellcome Trust (WT101033), Ministerio de Economía y Competitividad (BFU2014-54284-R) and Horizon 2020 (667191). Irene Miguel-Escalada has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska–Curie grant agreement No 658145. We acknowledge Prof. Giulio Cossu (Institute of Inflammation and Repair, University of Manchester) for providing the muscle myoblast cell line. We also acknowledge the InterAct and SIGMA Type 2 Diabetes Consortia for access to the data to replicate the rs146662075 variant. A full list of the investigators of the SIGMA Type 2 Diabetes and the InterAct consortia is provided in Supplementary Notes 3 and 4. The Novo Nordisk Foundation Center for Basic Metabolic Research is an independent research center at the University of Copenhagen partially funded by an unrestricted donation from the Novo Nordisk Foundation (www.metabol.ku.dk). This research has been conducted using the UK Biobank Resource (application number 16803). We also acknowledge Bianca C. Porneala, MS for his technical assistance in the collection and curation of the genotype and phenotype data from Partners Biobank. We also thank Marcin von Grotthuss for their support for uploading the summary statistics data to the Type 2 Diabetes Genetic Portal (AMP-T2D portal). Finally, we thank all the Computational Genomics group at the BSC for their helpful discussions and valuable comments on the manuscript.

Author contributions

S.B-G., J.M.M., and D.T. conceived, planned, and performed the main analyses. S.B-G., J.M.M., and D.T. wrote the manuscript. M.G-M., F.S., P.C-S., M.P., C.D., and R.M.B. developed a framework for large-scale imputation analyses. E.R-F., P.T., and T.H.P. performed pathway analysis. I.M-E. performed the enrichment analysis. M.P-F. and S.G. performed structural variant analyses. N.G., J.R-G., J.M., E.A.A., M.U., A.L., V.K., J.F., T.J., A.L., M.E.J., D.R.W., C.C., I.B., E.V.A., R.A.S., J.L., C.L., N.J.W., O.P., J.C.F., and T.H. contributed with additional data and analyses. G.A., I.M., and C.C.M. performed additional bioinformatics analyses. D.S. and A.Z. contributed muscle cell lines. I.M-E. and J.F. performed luciferase and electrophoretic mobility shift assays. J.M.M. and D.T. designed and supervised the study. All authors reviewed and approved the final manuscript.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-017-02380-9.

Competing interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Sílvia Bonàs-Guarch[1], Marta Guindo-Martínez[1], Irene Miguel-Escalada[2,3,4], Niels Grarup[5], David Sebastian[3,6,7], Elias Rodriguez-Fos[1], Friman Sánchez[1,8], Mercè Planas-Fèlix[1], Paula Cortes-Sánchez[1], Santi González[1], Pascal Timshel[5,9], Tune H. Pers[5,9,10,11], Claire C. Morgan[4], Ignasi Moran[4], Goutham Atla[2,3,4], Juan R. González[12,13,14], Montserrat Puiggros[1], Jonathan Martí[8], Ehm A. Andersson[5], Carlos Díaz[8], Rosa M. Badia[8,15], Miriam Udler[16,17], Aaron Leong[17,18], Varindepal Kaur[17], Jason Flannick[16,17,19], Torben Jørgensen[20,21,22], Allan Linneberg[20,23,24], Marit E. Jørgensen[25,26], Daniel R. Witte[27,28], Cramer Christensen[29], Ivan Brandslund[30,31], Emil V. Appel[5], Robert A. Scott[32], Jian'an Luan[32], Claudia Langenberg[32], Nicholas J. Wareham[32], Oluf Pedersen[5], Antonio Zorzano[3,6,7], Jose C Florez[16,17,33], Torben Hansen[5,34], Jorge Ferrer[2,3,4], Josep Maria Mercader[1,16,17] & David Torrents[1,35]

[1]Barcelona Supercomputing Center (BSC), Joint BSC-CRG-IRB Research Program in Computational Biology, 08034 Barcelona, Spain. [2]Genomic Programming of Beta-cells Laboratory, Institut d'Investigacions August Pi i Sunyer (IDIBAPS), 08036 Barcelona, Spain. [3]Instituto de Salud Carlos III, Centro de Investigación Biomédica en Red de Diabetes y Enfermedades Metabólicas Asociadas (CIBERDEM), 28029 Madrid, Spain. [4]Section of Epigenomics and Disease, Department of Medicine, Imperial College London, London W12 0NN, UK. [5]The Novo Nordisk Foundation Center for Basic Metabolic Research, Section for Metabolic Genetics, Faculty of Health and Medical Sciences, University of Copenhagen, 2100 Copenhagen, Denmark. [6]Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Baldiri Reixac 10-12, 08028 Barcelona, Spain. [7]Departament de Bioquímica i Biomedicina Molecular, Facultat de Biologia, Universitat de Barcelona, 08028 Barcelona, Spain. [8]Computer Sciences Department, Barcelona Supercomputing Center (BSC-CNS), 08034 Barcelona, Spain. [9]Department of Epidemiology Research, Statens Serum Institut, 2300 Copenhagen, Denmark. [10]Division of Endocrinology and Center for Basic and Translational Obesity Research, Boston Children's Hospital, Boston, MA 02116, USA. [11]Medical and Population Genetics Program, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. [12]ISGlobal, Centre for Research in Environmental Epidemiology (CREAL), 08003 Barcelona, Spain. [13]CIBER Epidemiología y Salud Pública (CIBERESP), 28029 Madrid, Spain. [14]Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain. [15]Artificial Intelligence Research Institute (IIIA), Spanish Council for Scientific Research (CSIC), 28006 Madrid, Spain. [16]Programs in Metabolism and Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA. [17]Diabetes Unit and Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA 02114, USA. [18]Division of General Internal Medicine, Massachusetts General Hospital, Boston, MA 02114, USA. [19]Department of Molecular Biology, Harvard Medical School, Boston, MA 02114, USA. [20]Research Centre for Prevention and Health, Capital Region of Denmark, DK-2600 Glostrup, Denmark. [21]Faculty of Health and Medical Sciences, University of Copenhagen, DK-2200 Copenhagen, Denmark. [22]Faculty of Medicine, University of Aalborg, DK-9220 Aalborg East, Denmark. [23]Department of Clinical Experimental Research, Rigshospitalet, Glostrup, 2100 Copenhagen, Denmark. [24]Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, DK-2200 Copenhagen, Denmark. [25]Steno Diabetes Center, 2820 Gentofte, Denmark. [26]National Institute of Public Health, Southern Denmark University, DK-5230 Odense M, Denmark. [27]Department of Public Health, Aarhus University, DK-8000 Aarhus C, Denmark. [28]Danish Diabetes Academy, DK-5000 Odense C, Denmark. [29]Medical department, Lillebaelt Hospital, 7100 Vejle, Denmark. [30]Department of Clinical Biochemistry, Lillebaelt Hospital, 7100 Vejle, Denmark. [31]Institute of Regional Health Research, University of Southern Denmark, DK-5230 Odense, Denmark. [32]MRC Epidemiology Unit, University of Cambridge School of Clinical Medicine, Cambridge Biomedical Campus, Cambridge CB2 0QQ, UK. [33]Department of Medicine, Harvard Medical School, Boston, MA 02115, USA. [34]Faculty of Health Sciences, University of Southern Denmark, DK-5230 Odense M, Denmark. [35]Institució Catalana de Recerca i Estudis Avançats (ICREA), 08010 Barcelona, Spain. Josep Maria Mercader and David Torrents jointly supervised this work.

Appendix 2. A cancer-associated polymorphism in ESCRT-III disrupts the abscission checkpoint and promotes genome instability.

Sadler JBA, Wenzel DW, Williams LK, Guindo-Martínez M, Alam SL, Mercader JM, Torrents D, Ullman KS, Sundquist WI, Martin-Serrano J. *PNAS* September 18, 2018 115 (38) E8900-E8908.

Contribution:

- Analysis of publicly available GWAS data for cancer to determine whether the *CHMP4C* rs35094336 variant is also associated with multiple cancer types.

A cancer-associated polymorphism in ESCRT-III disrupts the abscission checkpoint and promotes genome instability

Jessica B. A. Sadler[a,1], Dawn M. Wenzel[b,1], Lauren K. Williams[b,c], Marta Guindo-Martínez[d], Steven L. Alam[b], Josep M. Mercader[d,e,f,g,h], David Torrents[d,i], Katharine S. Ullman[c], Wesley I. Sundquist[b,2], and Juan Martin-Serrano[a,2]

[a]Department of Infectious Diseases, Faculty of Life Sciences and Medicine, King's College London, SE1 9RT London, United Kingdom; [b]Department of Biochemistry, University of Utah School of Medicine, Salt Lake City, UT 84112; [c]Department of Oncological Sciences, Huntsman Cancer Institute, University of Utah, Salt Lake City, UT 84112; [d]Joint Barcelona Supercomputing Center-Centre for Genomic Regulation-Institute for Research in Biomedicine Research Program in Computational Biology, Barcelona Supercomputing Center, 08034 Barcelona, Spain; [e]Program in Metabolism, Broad Institute of Harvard and MIT, Cambridge, MA 02142; [f]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA 02142; [g]Diabetes Unit, Massachusetts General Hospital, Boston, MA 02114; [h]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA 02114; and [i]Institució Catalana de Recerca i Estudis Avançats, 08010 Barcelona, Spain

Contributed by Wesley I. Sundquist, August 1, 2018 (sent for review March 30, 2018; reviewed by Stephen D. Bell and Arnaud Echard)

Cytokinetic abscission facilitates the irreversible separation of daughter cells. This process requires the endosomal-sorting complexes required for transport (ESCRT) machinery and is tightly regulated by charged multivesicular body protein 4C (CHMP4C), an ESCRT-III subunit that engages the abscission checkpoint (NoCut) in response to mitotic problems such as persisting chromatin bridges within the midbody. Importantly, a human polymorphism in CHMP4C (rs35094336, CHMP4C^{T232}) increases cancer susceptibility. Here, we explain the structural and functional basis for this cancer association: The CHMP4C^{T232} allele unwinds the C-terminal helix of CHMP4C, impairs binding to the early-acting ESCRT factor ALIX, and disrupts the abscission checkpoint. Cells expressing CHMP4C^{T232} exhibit increased levels of DNA damage and are sensitized to several conditions that increase chromosome missegregation, including DNA replication stress, inhibition of the mitotic checkpoint, and loss of p53. Our data demonstrate the biological importance of the abscission checkpoint and suggest that dysregulation of abscission by CHMP4C^{T232} may synergize with oncogene-induced mitotic stress to promote genomic instability and tumorigenesis.

abscission checkpoint | ESCRT pathway | cancer | genome instability | CHMP4C

Cytokinetic abscission is the final stage of cell division when the newly formed daughter cells are irreversibly separated. Abscission is a multistep process that culminates in the resolution of the midbody, the thin intercellular bridge that connects dividing cells following mitosis (1–3). The final membrane fission step of abscission is mediated by the endosomal-sorting complexes required for transport (ESCRT) pathway (4–7). The ESCRT machinery comprises membrane-specific adaptors and five core factors/complexes (ALIX, ESCRT-I, ESCRT-II, ESCRT-III, and VPS4), which are recruited sequentially (8–10). During cytokinesis, the midbody adaptor protein CEP55 initially recruits the early-acting ESCRT factors ALIX and ESCRT-I (4, 5, 11, 12). These factors, in turn, promote the recruitment and polymerization of essential ESCRT-III subunits, such as CHMP4B, to form filaments within the midbody. These membrane-associated filaments collaborate with the AAA ATPase VPS4 to constrict and sever the midbody (4–6, 11, 13).

Abscission is tightly coordinated with earlier stages of mitosis to ensure faithful inheritance of genetic material during cell division (14). In particular, cytokinetic abscission is temporally regulated by a conserved mechanism known as the "abscission checkpoint" (NoCut in yeast), which delays abscission in response to mitotic problems such as incomplete nuclear pore reformation or chromatin bridges within the midbody (15–19). The abscission checkpoint is governed by the master regulator, Aurora B kinase, which inhibits ESCRT-III activity in response to mitotic problems. Two key intersecting signaling nodes within this pathway are the ESCRT-III subunit CHMP4C and the regulatory ULK3 kinase. CHMP4C is a specialized ESCRT-III subunit that is dispensable for cytokinetic membrane fission, viral budding, and endosomal sorting but plays an essential role in executing the abscission checkpoint (20–22). CHMP4C is directly phosphorylated by Aurora B and is further phosphorylated by ULK3, which also phosphorylates other ESCRT-III subunits such as IST1. CHMP4C phosphorylation and ULK3 activity, together with the actions of other ESCRT-III–associated factors such as ANCHR, collectively prevent ESCRT-III polymerization and sequester VPS4 away from abscission sites, thereby delaying abscission (20, 23, 24).

Significance

The final step of cell division, abscission, is temporally regulated by the Aurora B kinase and charged multivesicular body protein 4C (CHMP4C) in a conserved pathway called the "abscission checkpoint" which arrests abscission in the presence of lingering mitotic problems. Despite extensive study, the physiological importance of this pathway to human health has remained elusive. We now demonstrate that a cancer-predisposing polymorphism in CHMP4C disrupts the abscission checkpoint and results in DNA damage accumulation. Moreover, deficits in this checkpoint synergize with p53 loss and generate aneuploidy under stress conditions that increase the frequency of chromosome missegregation. Therefore, cells expressing the cancer-associated polymorphism in CHMP4C are genetically unstable, thus suggesting an oncogenic mechanism that may involve the dysregulation of abscission.

Author contributions: J.B.A.S., D.M.W., K.S.U., W.I.S., and J.M.-S. designed research; J.B.A.S., D.M.W., L.K.W., M.G.-M., S.L.A., J.M.M., and D.T. performed research; J.B.A.S., D.M.W., L.K.W., M.G.-M., S.L.A., J.M.M., D.T., K.S.U., W.I.S., and J.M.-S. analyzed data; and J.B.A.S., D.M.W., W.I.S., and J.M.-S. wrote the paper.

Reviewers: S.D.B., Indiana University Bloomington; and A.E., Institut Pasteur.

The authors declare no conflict of interest.

Data deposition: The structures reported in this paper have been deposited in the Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank [accession codes 5V3R (ALIX Bro1-CHMP4C^{A232}) and 5WA1 (ALIX Bro1-CHMP4C^{T232})]. Atomic coordinates will be released upon publication.

[1]J.B.A.S. and D.M.W. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: wes@biochem.utah.edu or juan.martin_serrano@kcl.ac.uk.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1805504115/-/DCSupplemental.

Published online September 4, 2018.

Table 1. Significant association results for the rs35094336 genetic variant with multiple cancer types

Cancer type	OR (95% CI)	*P* value	No. cases	Cancer code
Family prostate*	1.04 (1.01–1.08)	0.007	34,359	FH1044
Male genital tract	1.12 (1.03–1.23)	0.012	3,449	1038
Skin	1.05 (1.01–1.09)	0.018	19,170	1003
Prostate	1.08 (1.01–1.16)	0.023	6,460	1044
Nonmelanoma skin	1.05 (1.01–1.10)	0.024	16,791	1060
Ovarian	1.17 (1.01–1.36)	0.037	1,208	1039

Individuals from UK Biobank (n = 337,208) were analyzed, and results show the association with the A (rs35094336) risk allele (CHMP4C^{T232}) compared with the G (CHMP4C^{A232}) reference allele. Data were obtained from Global Biobank Engine (https://biobankengine.stanford.edu/) accessed October 2017.

*Association with family history of prostate cancer.

Despite recent advances in identifying key components of the abscission checkpoint, the biological functions of the checkpoint and its contributions to human health are not yet known. Here, we have addressed these questions by analyzing the biochemical and abscission checkpoint activities of rs35094336, a human *CHMP4C* polymorphism (minor allele frequency = 0.04) associated with increased susceptibility to ovarian cancer (25). rs35094336 encodes an amino acid substitution of A232 (CHMP4C^{A232}; reference allele) to T232 (CHMP4C^{T232}; risk allele). Here, we show that the A232T substitution induces structural changes that impair ALIX binding and that cells expressing the CHMP4C^{T232} risk allele lack an abscission checkpoint and accumulate genetic damage. The CHMP4C^{T232} allele also sensitizes cells to chromosome missegregation and induces aneuploidy when the spindle-assembly checkpoint is weakened. These observations demonstrate the importance of the abscission checkpoint in maintaining genetic stability and suggest an oncogenic mechanism in which disruption of the abscission checkpoint by CHMP4C^{T232} may contribute to tumorigenesis by synergizing with oncogenic mutations that increase mitotic stress.

Results

CHMP4C^{T232} Is Associated with Multiple Cancer Types. The CHMP4C^{T232} allele was initially identified in a meta-analysis of two genome-wide association studies (GWAS) of SNPs associated with ovarian cancer (25). To test for an association of the CHMP4C^{T232} polymorphism with other cancers, we mined data from 337,208 individuals in the UK Biobank search engine (26). Our analysis of this independent cohort confirmed the previously identified association with ovarian cancer, and revealed statistically significant associations with multiple other types of cancer, including male genital tract, prostate, and skin cancers (Table 1). Although the odds ratios (ORs) for these associations are relatively modest (1.04–1.17), the association of the variant with increased risk for multiple different cancers suggests that this allele could be involved in a general pathway of genetic instability and tumorigenesis.

CHMP4C^{T232} Exhibits Reduced ALIX Binding. Position 232 is the penultimate CHMP4C residue, and A232 lies within a C-terminal helix that forms the ALIX-binding site (27). We therefore tested whether the A232T amino acid substitution affected ALIX binding and found that this substitution significantly reduced the interaction between full-length CHMP4C and ALIX (but not CHMP4C and itself) in a yeast two-hybrid assay (Fig. 1*A*). This substitution similarly inhibited the ability of a GST-fused C-terminal CHMP4C peptide (residues 216–233) to pull down endogenous ALIX from HeLa cell lysates (*SI Appendix*, Fig. S1*A*) and reduced the affinity of the terminal CHMP4C peptide for the pure recombinant ALIX Bro1 domain (residues 1–359) by 13-fold as measured in a competitive fluorescence polarization binding assay (Fig. 1*B*). In each case, we observed complete loss of ALIX binding to well-characterized control CHMP4C mutants that lacked key hydrophobic contact residues (L228A alone or with W231A) (27).

To determine the molecular basis for this reduction in ALIX-binding affinity, we determined high-resolution crystal structures of terminal CHMP4C^{A232} and CHMP4C^{T232} peptides (residues 216–233) bound to the ALIX Bro1 domain (residues 1–359) (Fig. 1*C* and *SI Appendix*, Fig. S1 *B* and *C* and Table S1) (27). Comparison of the structures revealed that although both peptides bound the same surface groove of the Bro1 domain, the A232T substitution disrupted several key ALIX interactions (Fig. 1*D*). Specifically, the A232T substitution unwound the C-terminal end of the terminal CHMP4C helix (Fig. 1*C* and *SI Appendix*, Fig. S1*D*), altered the position of CHMP4C residue W231, and disrupted intermolecular hydrogen bonds between the W231 carbonyl oxygen and indole nitrogen with ALIX residues D143 and K147, respectively (Fig. 1*D*). These structural analyses suggested that the A232T substitution might induce CHMP4C helix unwinding by introducing a beta-branched amino acid (which reduces helical propensity) and/or by prematurely capping the CHMP4C C-terminal helix (28). Indeed, both effects appeared to be operative, because mutant CHMP4C peptides that selectively retained only beta-branching (CHMP4C^{A232V}) or capping potential (CHMP4C^{A232S}) exhibited intermediate (three- to fourfold) reductions in ALIX peptide-binding affinity (Fig. 1*B*). Together, these analyses demonstrate that the CHMP4C^{T232} risk allele alters

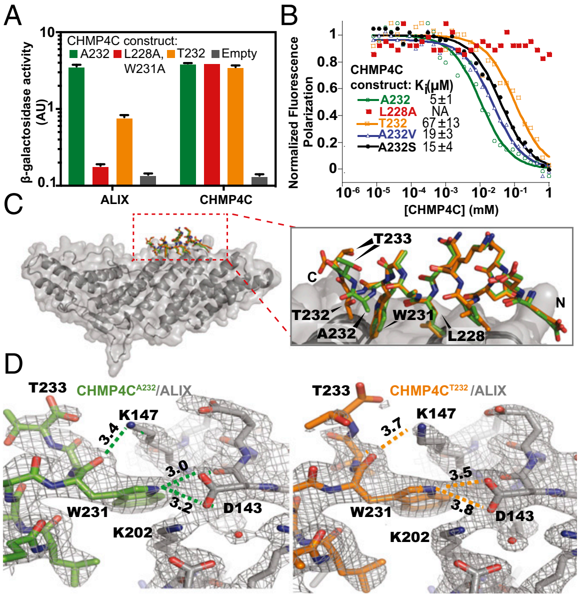


Fig. 1. The CHMP4C A232T substitution reduces ALIX binding. (*A*) β-Galactosidase activity assays of yeast cotransformed with the indicated full-length CHMP4C constructs fused to VP16 and full-length ALIX or CHMP4C fused to GAL4 (mean ± SD, n = 3). (*B*) Competitive fluorescence polarization binding assay with an ALIX construct spanning the Bro1 and V domains (residues 1–698) binding to fluorescently labeled CHMP4C^{A232} peptide (residues 216–233) competing with the indicated unlabeled CHMP4C peptides. Curves are from a representative experiment. K_i values are expressed as mean ± SD. $n \geq 7$. (*C*, *Left*) Superposition of ALIX Bro1 domain (gray) complexes with CHMP4C^{A232} (green) and CHMP4C^{T232} (orange) peptides. (*Right*) Enlarged view of the CHMP4C peptide superposition highlights the reduction in CHMP4C^{T232} helicity. (*D*) Binding interfaces between ALIX (gray) and CHMP4C^{A232} (green) (*Left*) and CHMP4C^{T232} (orange) (*Right*). Figures show equivalent intermolecular distances (in angstroms) for different atoms of CHMP4C residue W231. See also *SI Appendix*, Fig. S1 and Table S1.

the structure of the CHMP4C C-terminal helix, removes key ALIX interactions, and reduces ALIX-binding affinity by more than an order of magnitude.

ALIX–CHMP4C Interactions Are Required for Abscission Checkpoint Activity. ALIX is a key initiator of the cytokinetic abscission cascade (4, 5, 11, 29), and CHMP4C plays an essential role in maintaining the abscission checkpoint (20, 21). We therefore tested whether abscission checkpoint activity was affected by CHMP4C mutations that impaired ALIX binding, including the CHMP4C^{T232} risk allele. In these experiments, siRNA treatment was used to deplete endogenous CHMP4C from HeLa cells engineered to stably express different siRNA-resistant HA-CHMP4C proteins (Fig. 2 and *SI Appendix*, Fig. S2). Partial depletion of nuclear pore components Nup153 and Nup50 was used to activate the abscission checkpoint (30). As expected, control cells that expressed endogenous CHMP4C stalled during abscission, as indicated by elevated midbody connections, whereas cells depleted of CHMP4C did not exhibit elevated levels of midbody connections (Fig. 2*A* and *SI Appendix*, Fig. S2*A*) (20, 30). Importantly, checkpoint activity was rescued in cells that expressed HA-CHMP4C^{A232} but not in cells that expressed HA-CHMP4C^{T232}, implying that the CHMP4C^{T232} risk allele does not support the abscission checkpoint. The HA-CHMP4C^{L228A,W231A} mutant also failed to support the abscission checkpoint, further indicating that the CHMP4C–ALIX interaction is required to sustain the checkpoint. Notably, the loss of checkpoint activity for both HA-CHMP4C^{L228A,W231A} and HA-CHMP4C^{T232} was comparable to the defective response observed in cells that expressed an inactive control CHMP4C mutant lacking the amino acid insertion phosphorylated by Aurora B (HA-CHMP4C^{-INS}) (20, 21). Similarly, cells expressing only HA-CHMP4C^{T232}, HA-CHMP4C^{L228A,W231A}, or HA-CHMP4C^{-INS} also proceeded through abscission more rapidly under normal growth conditions, implying that they were insensitive to steady-state abscission checkpoint activity, likely induced by midbody tension (Fig. 2*B*, *SI Appendix*, Fig. S2*B*, and Movies S1–S6) (24, 31). Despite these defects, HA-CHMP4C^{A232}, HA-CHMP4C^{T232}, HA-CHMP4C$^{L228A, W231A}$, and HA-CHMP4C^{-INS} were all recruited to the midbody at normal levels, whether or not nucleoporins were depleted (Fig. 2 *D* and *E*). Additionally, CHMP4C ALIX-binding mutants and wild-type CHMP4C proteins were mitotically phosphorylated at comparable levels (*SI Appendix*, Fig. S2*D*), indicating that the CHMP4C mutations did not disrupt recognition by Aurora B or ULK3 kinases and that ALIX binding is not required for these activities. These observations imply that abscission checkpoint activity requires ALIX binding to CHMP4C. We find, however, that CHMP4C midbody localization does not require ALIX binding, in contrast to a previous report (29).

To test CHMP4C^{T232} activity when the abscission checkpoint was activated by a different trigger, we used live-cell imaging to measure the resolution times of intercellular chromatin bridges, as visualized using the nuclear envelope marker lamina-associated polypeptide 2β fused to YFP (YFP-LAP2β). As expected (20), chromatin bridges were resolved prematurely in CHMP4C-depleted cells compared with cells that expressed endogenous CHMP4C^{A232} (median resolution time = 250 vs. 685 min) (Fig. 2*C*, *SI Appendix*, Fig. S2*C*, and Movies S7–S12). Importantly, prolonged midbody resolution times were restored by expression of siRNA-resistant HA-CHMP4C^{A232} (800 min) but not by HA-CHMP4C^{T232} (200 min), HA-CHMP4C^{L228A,W231A} (300 min), or HA-CHMP4C^{-INS} (340 min). Thus, cells expressing the cancer-associated CHMP4C^{T232} risk allele lack an appropriate abscission checkpoint response under multiple different conditions that activate this checkpoint.

Disruption of the Abscission Checkpoint Leads to Accumulation of DNA Damage. Although the biological consequences of abscission checkpoint loss are not well understood, increased DNA damage is one possible outcome (20). We therefore compared DNA damage accumulation in cells that expressed the different CHMP4C mutants. In these experiments, CRISPR-Cas9 was used to delete the CHMP4C locus from HCT116 cells (*SI Appendix*, Fig. S3), a near diploid cell line that exhibits low levels of chromosomal instability and DNA damage (32). Genetic damage was then assessed by scoring the number of nuclear foci formed by the DNA damage response marker 53BP1. As expected, cells with low levels of DNA damage (fewer than two foci per cell) predominated in the wild-type cultures (HCT116^{WT}) (Fig. 3 *A* and *B*). In contrast, cells lacking CHMP4C (HCT116$^{\delta CHMP4C}$) exhibited heightened DNA damage (more than six foci per cell) with significantly greater frequency. Crucially, DNA damage in the HCT116$^{\delta CHMP4C}$ cells was reduced to control levels upon reexpression of HA-CHMP4C^{A232} but not HA-CHMP4C^{T232}, HA-CHMP4C^{L228A,W231A}, or HA-CHMP4C^{-INS}. Hence, the loss of the CHMP4C-dependent abscission checkpoint increases the accumulation of 53BP1-associated DNA damage foci.

CHMP4C Is Not Required for Maintenance of Nuclear Integrity, DNA Damage Responses, or Mitotic Checkpoint Signaling. To determine whether observed phenotypes were specifically due to abscission checkpoint failure, we examined other mechanisms that might underlie the elevated levels of DNA damage in HCT116$^{\delta CHMP4C}$ cells. We ruled out ESCRT-dependent loss of nuclear envelope integrity (33, 34) because nuclear envelope compartmentalization was not compromised during telophase in cells lacking CHMP4C, as assayed by nuclear morphology and retention of a GFP-NLS reporter (*SI Appendix*, Fig. S4 and Movies S13–S15). Similarly, loss of CHMP4C did not globally impair DNA damage responses because the efficiency of G2/M cell-cycle arrest in response to genotoxic stress induced by the DNA cross-linker mitomycin C was normal in HCT116$^{\delta CHMP4C}$ cells (*SI Appendix*, Fig. S5). Furthermore, in contrast to another report (35), we did not observe a failure of HCT116$^{\delta CHMP4C}$ cells to arrest in response to spindle poisons such as nocodazole, indicating that the spindle-assembly checkpoint remains largely intact in these cells (*SI Appendix*, Fig. S6). Therefore, the increased DNA damage in cells lacking CHMP4C activity and a functional abscission checkpoint does not reflect the loss of nuclear integrity, improper DNA damage responses, or defective mitotic spindle-assembly checkpoint signaling but rather a loss of the abscission checkpoint.

CHMP4C^{T232} Sensitizes Cells to Replication Stress. We next examined the possibility that cells lacking CHMP4C activities had increased levels of DNA damage because they were unable to respond properly to DNA replication stress. This is an attractive model because (*i*) a significant fraction of 53BP1 nuclear bodies originate from lesions generated by DNA replication stress (36), (*ii*) elevated replication stress triggers the abscission checkpoint in a CHMP4C-dependent manner (Fig. 3*C*) (18), and (*iii*) the abscission checkpoint plays a key role in protecting anaphase bridges that arise from replication stress (37), thereby reducing damage when they persist during cytokinetic abscission (38). In agreement with this model, inducing replication stress with ultra-low doses (30 nM) of the DNA polymerase inhibitor aphidicolin reduced the proliferation of HCT116$^{\delta CHMP4C}$ cells nearly twofold compared with HCT116^{WT} cells (Fig. 3*D*). Importantly, the HCT116$^{\delta CHMP4C}$ growth defect was again rescued by expression of HA-CHMP4C^{A232} but not by the abscission checkpoint-defective HA-CHMP4C^{T232}, HA-CHMP4C^{L228A,W231A}, or HA-CHMP4C^{-INS} mutants. Thus, the abscission checkpoint can play a protective role in cell survival when cells are subjected to increased replication stress.

CHMP4C^{T232} Sensitizes Cells to Chromosome Missegregation and Induces Aneuploidy. To examine the functions of the abscission checkpoint in the context of another mitotic stress, we tested whether a defective abscission checkpoint also sensitized cells to weakening of the spindle-assembly checkpoint, a condition that

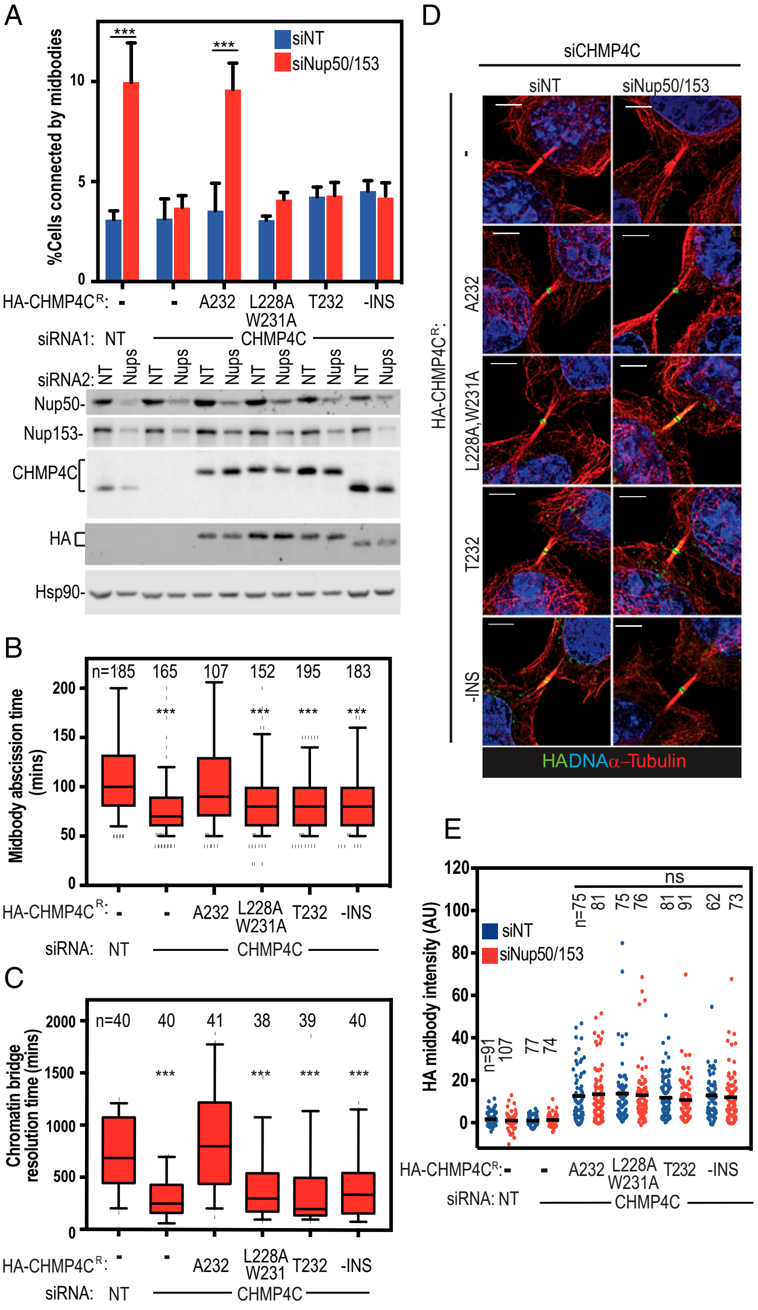


Fig. 2. CHMP4C T232 does not support the abscission checkpoint. (*A*) HeLa cells stably expressing siRNA-resistant HA-CHMP4C constructs were treated with the indicated siRNA and stained for α-tubulin. (*Upper*) Percentage of midbody-arrested cells. Data are mean ± SD from more than three separate experiments with $n > 900$ cells. (*Lower*) Representative immunoblots from *A*. (*B*) Asynchronous HeLa mCherry-tubulin cells stably expressing siRNA-resistant HA-CHMP4C constructs were transfected with the indicated siRNA, and midbody abscission times were scored from three or more separate experiments. Here and throughout, box edges mark the 25th and 75th quartiles, and whiskers mark fifth and 95th percentiles. Horizontal bars denote the median. Mean times ± SD were nontargeting (NT): 110 ± 47 min; siCHMP4C: 75 ± 26 min; HA-CHMP4C^{A232} + siCHMP4C: 108 ± 47 min; HA-CHMP4C^{L228A,W231A} + siCHMP4C: 86 ± 33 min; HA-CHMP4C^{T232} + siCHMP4C: 81 ± 27 min; and HA-CHMP4C^{-INS} + siCHMP4C: 87 ± 44 min. See also Movies S1–S6. (*C*) HeLa cells stably expressing siRNA-resistant HA-CHMP4C constructs and YFP-Lap2β were transfected with the indicated siRNA. Resolution time of Lap2β bridges were quantified from three or more separate experiments. Mean times ± SD were nontargeting (NT): 721 ± 333 min; siCHMP4C: 296 ± 187 min; HA-CHMP4C^{A232} + siCHMP4C: 853 ± 474 min; HA-CHMP4C^{L228A,W231A} + siCHMP4C: 415 ± 317 min; HA-CHMP4C^{T232} + siCHMP4C: 396 ± 394 min; and HA-CHMP4C^{-INS} + siCHMP4C: 421 ± 341 min. See also Movies S7–S12. (*D*) HA-CHMP4C recruitment to midbodies was determined by staining for DNA (DAPI, blue), α-tubulin (red), and HA (green) from three or more separate experiments. (Scale bars, 5 μm.) (*E*) Data represent the staining intensity of HA normalized to background measurements. The mean value is marked. *P* values were calculated using two-way ANOVA and Sidak's multiple comparisons test (*A* and *E*) or one-way ANOVA vs. control (HeLa siNT) (*B* and *C*); ***$P < 0.001$; ns, not significant. Immunoblots for *B* and *C* are shown in *SI Appendix*, Fig. S2.

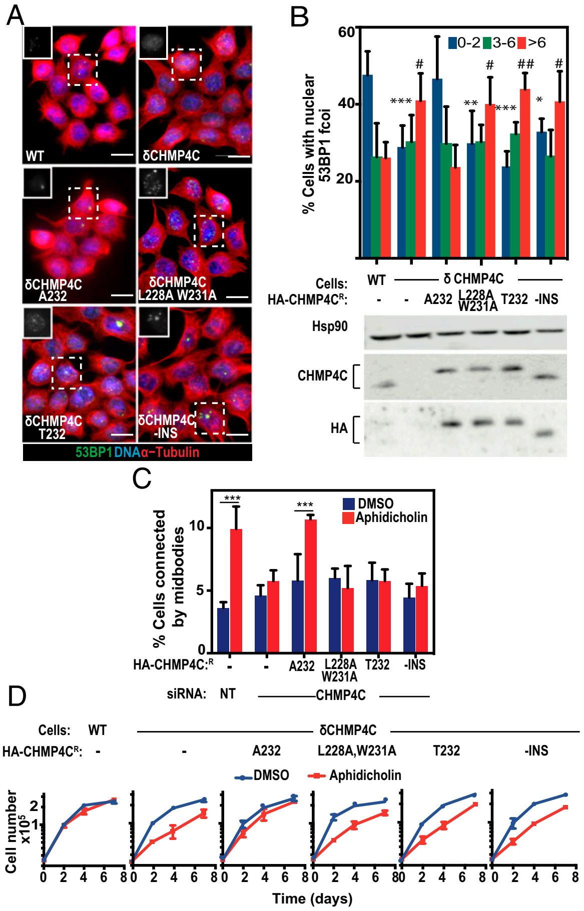


Fig. 3. Cells lacking the abscission checkpoint exhibit elevated genome damage and are sensitized to replication stress. (*A*) HCT116 cells with (WT) or without (δCHMP4C) endogenous CHMP4C expressing the indicated HA-CHMP4C construct were stained for 53BP1 (green), α-tubulin (red), and DNA (DAPI, blue). (Scale bars, 20 μm.) *Insets* show gray-scale images of boxed cells. (*B*, *Upper*) Numbers of 53BP1 foci per cell were determined from three independent experiments and binned into the designated categories. Shown are the mean ± SD from >900 cells. *P* values were calculated using two-way ANOVA and Sidak's multiple comparisons test comparing each category to control (WT); *: 0–2 foci; #: more than six foci; $^{*/\#}P < 0.05$; $^{**/\#\#} = P < 0.005$; $^{***/\#\#\#} = P < 0.001$. (*Lower*) Representative immunoblots from *B*, *Upper*. (*C*) Cells stably expressing the indicated HA-CHMP4C construct were treated with the indicated siRNA for 48 h and then with DMSO or 30 nM aphidicolin for 24 h. The number of cells connected by midbodies was scored. (*D*) HCT116^{WT} or HCT116$^{\delta CHMP4C}$ cells expressing the indicated HA-CHMP4C constructs were cultured in the continuous presence of DMSO (blue trace) or 30 nM aphidicolin (red trace), and cell numbers were determined at the indicated time points. Plotted are the mean ± SD from three independent experiments. See also *SI Appendix*, Fig. S7*A*.

induces anaphase chromosomal segregation errors. The spindle-assembly checkpoint was selectively weakened by treatment with low doses (0.1 μM) of the MPS1 kinase inhibitor reversine (39), which doubled the frequency of anaphase chromosome missegregation (Fig. 4*A*, *SI Appendix*, Fig. S7*B*, and Movies S16–S19) and activated the abscission checkpoint in a CHMP4C-dependent fashion (*SI Appendix*, Fig. S7*A*). Karyotyping of metaphase spreads revealed that HCT116^{WT} cells only rarely displayed extreme aneuploidy (1% of DMSO-treated cells had <37 or >48 chromosomes) (Fig. 4 *B* and *C*). Reversine treatment alone increased this percentage, but cells with extreme aneuploidy were still rare (7%). In contrast, DMSO-treated HCT116$^{\delta CHMP4C}$ cells exhibited a higher basal level of extreme aneuploidy (4%), and this percentage increased notably upon reversine treatment (23%). Reexpression of HA-CHMP4C^{A232}, but not HA-CHMP4C^{T232}, protected against reversine-induced increases in aneuploidy (*SI Appendix*, Fig. S8*A*). Importantly, reversine treatment in cells with a defective abscission checkpoint did not induce a multinucleation phenotype, but these cells did display a higher proportion of micronuclei (*SI Appendix*, Fig. S8 *B* and *C*).

In a complementary set of experiments, we monitored cell growth rates in reversine-treated cultures, where the high levels of chromosomal instability and aneuploidy are detrimental to cell survival

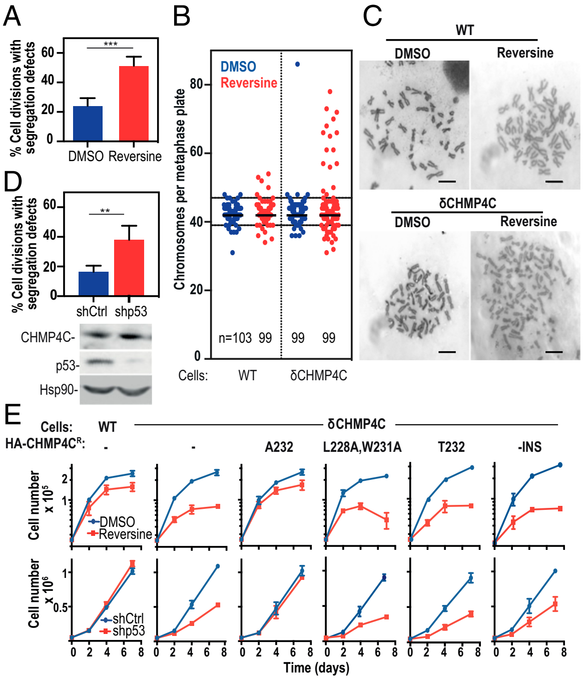


Fig. 4. Defects in chromosome segregation and the abscission checkpoint synergize to impair cell growth. (*A*) HCT116 cells expressing histone H2B-mCherry were grown in the continuous presence of DMSO (blue bar) or 0.1 μM reversine (red bar), and anaphase segregation defects were scored. Data shown are the mean ± SD from *n* > 130 cells from three separate experiments. See also *SI Appendix*, Fig. S7 *A* and *B* and Movies S16–S19. (*B* and *C*) HCT116^{WT} or HCT116$^{\delta CHMP4C}$ cells were cultured for 48 h in the continuous presence of DMSO (blue) or 0.1 μM reversine (red). Metaphases were enriched by overnight treatment with nocodazole, and chromosome number was determined. Plots show all data with medians marked. Extreme aneuploidy is defined as chromosome numbers above 48 or below 37 (solid lines). Representative metaphase spreads can be seen in *C*. (Scale bars, 10 μm.) Data were collected from more than four independent experiments. (*D*, *Upper*) HCT116 cells expressing histone H2B-mCherry were depleted of p53 by shRNA treatment (shp53), and anaphase segregation defects were scored. Data are expressed as the mean ± SD, *n* > 200, from four separate experiments across two shRNA transductions. See also *SI Appendix*, Fig. S7*C* and Movies S20–S23. (*Lower*) Representative immunoblots from *D*, *Upper*. (*E*) HCT116^{WT} or HCT116$^{\delta CHMP4C}$ cells expressing the indicated CHMP4C construct were grown in the continuous presence of either DMSO (blue traces) or 0.1 μM reversine (red traces) (*Upper Row*) or were depleted of p53 by shRNA treatment (shCtrl, blue; shp53, red) (*Lower Row*), and cell numbers were determined at the indicated time points. Data are the mean ± SD from more than three independent experiments. *P* values were calculated using two-tailed unpaired Student's *t* test; $^{**}P < 0.005$, $^{***}P < 0.001$. See also *SI Appendix*, Fig. S9*A*.

and growth (Fig. 4*E*, *Upper Row*) (40). In agreement with the karyotype analyses, reversine treatment reduced HCT116^{WT} cell growth only modestly but reduced HCT116$^{\delta CHMP4C}$ cell growth by at least 70% over a 7-d period. Robust growth in the presence of reversine was restored by the expression of HA-CHMP4C^{A232} but not HA-CHMP4C^{T232}, HA-CHMP4C^{L228A,W231A}, or HA-CHMP4C^{-INS} (Fig. 4*E*, *Upper Row*). Thus, the CHMP4C-dependent abscission checkpoint becomes more critical for cell growth under conditions that increase chromosome-segregation defects.

CHMP4C^{T232} Synergizes with p53 Loss. Our observations that CHMP4C mutations can abolish the abscission checkpoint and that these mutations synergize with increases in chromosome missegregation raised the intriguing possibility that such mutations might also synergize with genetic alterations known to be associated with ovarian cancer development. We focused our studies on TP53, the most frequently mutated gene in a wide range of cancers, including >96% of high-grade serous ovarian tumors (41). In addition to its classical role in inducing cell-cycle arrest in response to DNA damage, p53 also functions directly in DNA damage repair and chromosomal stability, and its absence is associated with increased replicative stress (42). As expected, stable depletion of p53 increased DNA damage levels as measured by increased numbers of 53BP1 foci (*SI Appendix*, Fig. S9 *B* and *C*) and increased frequencies of chromosomal segregation defects during anaphase (Fig. 4*D*, *SI Appendix*, Fig. S7*C*, and Movies S20–S23). As with reversine treatment, these defects significantly compromised HCT116 cell growth only when CHMP4C was absent (Fig. 4*E*, *Lower Row* and *SI Appendix*, Fig. S9*A*). This synergistic effect was again reversed upon reexpression of CHMP4C^{A232} but not HA-CHMP4C^{T232}, HA-CHMP4C^{L228A,W231A}, or HA-CHMP4C^{-INS} (Fig. 4*E*, *Lower Row* and *SI Appendix*, Fig. S9*A*). Thus, the loss of the CHMP4C-dependent abscission checkpoint also synergizes with the loss of functional p53. This effect can again be explained by an inability of cells to cope with the increased burden of chromosomal segregation defects, perhaps compounded further by dysfunction of the p53-mediated G1 checkpoint (43).

Discussion

Our study demonstrates that the abscission checkpoint plays a critical role in human health by protecting the genome against DNA damage and chromosomal instability. We have shown that a human polymorphism in *CHMP4C*, previously associated with increased susceptibility to ovarian cancer, is also associated with increased risk for several other cancers, thus suggesting that the CHMP4C^{T232} allele contributes to tumor development in a global fashion. Importantly, cells that express the cancer-associated CHMP4C^{T232} allele show elevated levels of 53BP1 foci, suggesting that increased DNA damage may account, at least in part, for increased cancer susceptibility in individuals who carry this allele. Furthermore, cells that express CHMP4C^{T232} are particularly sensitized to genomic instability under conditions that increase the burden of chromosomal segregation defects, such as DNA replication stress and weakening of the spindle-assembly checkpoint.

At the mechanistic level, the A232T substitution unwinds the C-terminal CHMP4C helix, impairs ALIX binding affinity, and leads to the loss of abscission checkpoint activity. Hence, in addition to the previously documented requirements for CHMP4C phosphorylation by Aurora B and ULK3 (20, 21, 24), CHMP4C must also be able to interact with ALIX (and possibly other Bro domain-containing proteins) to support the abscission checkpoint. We found, however, that altered CHMP4C midbody localization could not explain the loss of checkpoint activity because point mutations that specifically abolished ALIX binding did not reduce CHMP4C midbody localization or mitotic phosphorylation. In contrast, others have observed reduced midbody localization of a C-terminally truncated CHMP4C construct (29). It is therefore possible that the C-terminal region of CHMP4C dictates midbody localization independently of ALIX binding, perhaps through interactions with MKLP1 or the chromosomal passenger complex (20, 44). Alternatively, removal of 18 terminal CHMP4C residues could have relieved CHMP4C autoinhibition, thereby impairing midbody localization indirectly. This idea is consistent with the observation that autoinhibition of Snf7p, the yeast ortholog of CHMP4, can be relieved by removing its terminal Bro1p (ALIX)-binding helix (45–47). We do not yet know for certain why CHMP4C–ALIX binding is required to support the abscission checkpoint, but one intriguing possibility is that CHMP4C binding may competitively inhibit CHMP4B from occupying its overlapping binding site on ALIX (27), thereby sustaining the abscission checkpoint by preventing nucleation of CHMP4B-containing ESCRT-III filaments within the midbody.

Another striking finding of our study is that increasing chromosomal segregation defects in cells lacking a functional abscission checkpoint specifically induces high levels of aneuploidy and chromosomal instability. Although complete inhibition of Aurora B leads to cleavage furrow regression and binucleation when chromatin is present in the intracellular bridge (17), this mechanism does not appear to explain the increases in aneuploidy when CHMP4C is impaired. Depletion of CHMP4C (or other abscission checkpoint components downstream of Aurora B such as ULK3) induces premature resolution of chromatin bridges, not furrow regression (20, 24). Moreover, reversine treatment in cells lacking CHMP4C does not lead to multinucleation but does induce micronuclei formation. Our data are consistent with chromosomal instability resulting from chromosome breakage and refusion events and/or the failure to reincorporate lagging chromosomes into the main nuclei. It has been suggested that even very mild aneuploidy has consequences beyond chromosome gains or losses, resulting in DNA damage and replication stress which have severe effects on subsequent mitoses, and that the gain of even a single chromosome can result in further chromosomal aberrations and complex karyotypes in subsequent cell cycles (48–50). We suggest that DNA damage acquired over a number of cell cycles in cells lacking a functional abscission checkpoint may have cumulative effects that are ultimately detrimental in subsequent cell cycles. Consistent with this idea, damage acquired during mitosis can lead to p53-dependent quiescence in daughter cells (51, 52). Thus, the coordinated action of the abscission checkpoint and p53 may protect against aneuploidy. In this model, the abscission checkpoint provides additional time to retrieve, repair, and/or protect lagging chromosomes, thereby protecting cells and preventing catastrophic DNA damage during mitosis. Disruption of the abscission checkpoint could also induce aneuploidy via a mechanism that is reminiscent of the checkpoint adaptation phenomenon, in which cells continue to proceed through the cell cycle despite not having completely resolved DNA damage arising from the previous mitosis (50). Such checkpoint-adapted cells are characterized by severe chromosomal segregation defects that give rise to micronuclei containing lagging chromosomes (or chromosome fragments), which then contribute to further chromosome damage and instability in subsequent divisions.

Chromosomal segregation errors are a hallmark of many cancers and often arise in response to oncogenic mutations that increase mitotic stress (40). In particular, loss of p53 is the most common genetic abnormality in many tumor types and can lead to increased DNA damage, chromosomal instability, and increased replicative stress (41, 42, 53). These phenotypes suggest a potential mechanism by which loss of the abscission checkpoint could contribute to genetic instability and cancer development, particularly as we observed synthetic lethality between loss of p53 and the CHMP4C^{T232} risk allele. Our data suggest the possibility that homozygous germline expression of the CHMP4C^{T232} allele or the loss of heterozygosity of CHMP4C^{A232}-encoding allele expression in somatic cells where the T232-encoding allele is present could contribute to tumorigenesis by increasing genomic instability and aneuploidy, particularly when chromosome missegregation events are elevated. We speculate that although an impaired abscission checkpoint combined with genetic alterations such as p53 loss is detrimental to overall cell growth, the subset of cells that ultimately survive may accumulate further adaptations that promote tumorigenicity. These cells may nevertheless remain sensitive to further perturbations of chromosome segregation or DNA-damaging agents, and this sensitivity could, in principle, be exploited therapeutically. In this regard, it is noteworthy that CHMP4C depletion increases the effectiveness of irradiation-induced apoptosis in human lung cancer cells (54) and that many common chemotherapeutic drugs, such as paclitaxel, act, at least in part, by increasing chromosome missegregation (40). Hence, we speculate that such chemotherapeutics may be particularly effective in patients who carry the CHMP4C$^{\mathrm{T232}}$ allele.

Methods

Plasmids and Antibodies. Details of plasmids and antibodies used in this study are described in *SI Appendix*, Table S2.

Fluorescence Polarization Binding Experiments. Fluorescence polarization was measured using a BioTek Synergy Neo Multi-Mode plate reader (BioTek) with excitation at 485 nm and detection at 528 nm. For competitive binding experiments, the wild-type CHMP4C peptide (residues 216–233) was synthesized with a nonnative cysteine at the N terminus (CQRAEEEDDDIKQLAAWAT) and was labeled with Oregon Green 488 (Life Technologies/Molecular Probes 06,034) following the manufacturer's instructions. The labeled peptide was quantitated by the absorbance of Oregon Green 488 at 491 nm (extinction coefficient 83,000 cm·M^{-1} in 50 mM potassium phosphate, pH 9). Different concentrations (as determined by absorbance at 280 nm) of unlabeled N-terminally acetylated CHMP4C peptides were titrated against a CHMP4C$^{\mathrm{A232}}$-ALIX Bro1-V complex created by mixing 5 μM ALIX Bro1-V and 0.5 nM fluorescently labeled CHMP4C$^{\mathrm{A232}}$ peptide in binding buffer [20 mM sodium phosphate (pH 7.2), 150 mM NaCl, 5 mM β-mercaptoethanol, 0.01% Tween-20, and 0.2 mg/mL BSA]. IC$_{50}$s were calculated from binding curves using KaleidaGraph (Synergy Software) and were converted to K_i values (55). Competitive binding curves were measured independently seven or more times for each peptide and are expressed as mean ± SD. K_i values are reported. All peptides were synthesized by the University of Utah Peptide Synthesis Core Facility and were verified by mass spectrometry.

GST Pull-Downs. GST-fused CHMP4C peptides spanning the ALIX-binding helix (residues 216–233) from CHMP4C$^{\mathrm{A232}}$ or from the CHMP4C$^{\mathrm{L228A,W231A}}$ or CHMP4C$^{\mathrm{T232}}$ mutants were purified, immobilized on glutathione-Sepharose agarose beads, and incubated with clarified HeLa cell lysates. Bound material was analyzed by SDS/PAGE followed by immunoblotting or Coomassie staining. A detailed description is provided in *SI Appendix, Methods*.

Cell Culture. HEK293T, HeLa, and HCT116 cells were cultured and maintained in DMEM supplemented with 10% FBS and 20 μg/mL gentamycin. To generate stable cell lines, 293T cells were transfected with retroviral packaging vectors (*SI Appendix*, Table S2), MLV-GagPol/pHIV 8.1, and pHIT VSVg at a ratio of 3:2:1 for 48 h. 293T supernatant was filtered through a 0.2-μm filter and was used to transduce the indicated cell lines; antibiotic selection was carried out 48 h later. MycoAlert (Lonza) was used to screen for mycoplasma contamination. HCT116 CRISPR cell lines were generated by transfection with retroviral Cas9 expression plasmids containing specific guide RNAs targeting CHMP4C; full details are available in *SI Appendix, Methods*.

siRNA Transfections. Cells were transfected with siRNA for 72 h using DharmaFECT 1 (Dharmacon) or Lipofectamine RNAiMax (Thermo Fisher Scientific) according to the manufacturers' instructions. Cells received two transfections, one at 0 h and another at 48 h. For HeLa cells, CHMP4C and nontargeting siRNA were used at 100 nM, and Nup50 and Nup153 siRNA were used at 10 nM. For HCT116 cells, all siRNA was used at 10 nM. Cells were fixed or imaged 24 h after the second siRNA transfection. For GFP-NLS nuclear fluorescence recovery experiments HCT116 cells were imaged 8 h after the second transfection. The siRNA sequences used in this study have been described (20, 22, 24, 30, 31) and are available in *SI Appendix*, Table S2.

Immunoblotting. Cell lysates were denatured in Laemmli buffer, resolved by SDS/PAGE, and transferred to nitrocellulose membranes. Membranes were blocked with 5% skim milk in 0.1% Tween 20/Tris-buffered saline (TBS) and were incubated with primary antibodies in either 1% or 5% skim milk in blocking solution for 3 h at room temperature or overnight at 4 °C. Membranes were washed in 0.1% Tween 20/TBS, incubated with the corresponding secondary antibodies conjugated with HRP or near-infrared fluorescent dyes in blocking solution for 1 h at room temperature, and washed again. Proteins were detected and quantified using a Li-Cor Odyssey Infrared scanner and software (Li-Cor Biosciences) or Image Quant LAS 400 (GE Healthcare). Details on antibodies and dilutions can be found in *SI Appendix*, Table S2.

Immunofluorescence. Cells were grown on coverslips, washed once in PBS, and fixed for 10–20 min in ice-cold methanol. Cells were blocked with 3% FCS and 0.1% Triton X-100 in PBS for 20 min. Primary antibodies were applied for at least 1 h. After four washings with PBS, secondary antibodies were applied for 1 h, and nuclei were stained with either Hoechst or DAPI. Coverslips were mounted with ProLong Gold Antifade Reagent (Invitrogen) on a microscope slide. Images were acquired using a Leica SP8 Confocal (Fig. 2*D*) or Nikon Ti-Eclipse wide-field inverted microscope (Fig. 3 and *SI Appendix*, Figs. S2, S4, S8, and S9). Scoring was conducted blind. Where indicated, deconvolution was performed using HyVolution Pro-Automatic deconvolution software. Quantification of fluorescence staining intensity was carried out with ImageJ (NIH). The freehand selection tool was used to outline the region of interest, and staining intensity within this area was measured. Signals were background corrected using measurements from adjacent regions. Details on antibodies and dilutions can be found in *SI Appendix*, Table S2.

Live-Cell Imaging. Cells were seeded on glass-bottomed 24-well plates (MatTek) and transfected with siRNA or shRNA or were subjected to the specified drug treatments. Imaging was carried out for 24–72 h on a Nikon Ti-Eclipse wide-field inverted microscope (Nikon 40 × 0.75 N.A. dry objective lens) equipped with Perfect Focus system and housed in a 37 °C chamber (Solent Scientific) fed with 5% CO$_2$. Multiple fields of view were selected at various x and y coordinates, and three z slices were captured at 1.25-μm spacing for HeLa cells and 1.8-μm spacing for HCT116 cells. Images were acquired using a Hamamatsu Orca Flash 4.0 camera (Hamamatsu Photonics) controlled by NIS-Elements software. For abscission time measurements, images were acquired every 10 min for 48 h, and abscission time was measured as the time from midbody formation to disappearance. For resolution timing for YFP-Lap2β–positive bridges, images were acquired every 20 min for 72 h, and resolution time was measured from the time of the appearance to the disappearance of YFP-Lap2β–positive intercellular bridges. For nuclear accumulation of GFP-NLS and analysis of anaphase defects, images were acquired every 5 min for 24 h. The nuclear GFP signal was identified through colocalization with H2B-mCherry and was normalized to the cytoplasmic signal. Signals were background corrected using measurements from adjacent regions. Measurements were taken 10 frames before and 50 frames after nuclear envelope breakdown, with cells depleted of CHMP7 serving as a positive control (56).

Protein Expression and Purification. The human ALIX Bro1 domain (residues 1–359) (Fig. 1 *C* and *D* and *SI Appendix*, Fig. S1) and ALIX Bro1-V domains (residues 1–698) (Fig. 1*B*) were expressed and purified as previously described with minor modifications (27, 57). A detailed description is provided in *SI Appendix, Methods*. Plasmids for bacterial expression of ALIX proteins are available from the Addgene plasmid repository (www.addgene.org; see *SI Appendix*, Table S2 for accession numbers).

Crystallization and Data Collection. ALIX Bro1 crystallized in complex with CHMP4C$^{\mathrm{A232}}$ or CHMP4C$^{\mathrm{T232}}$ peptides (residues 216–233; N-terminally acetylated) at 20 °C from sitting drops that contained 1.2 μL of protein (200 μM ALIX Bro1 and 220 μM CHMP4C peptide) and 0.7 μL of reservoir solution [CHMP4C$^{\mathrm{A232}}$: 10% PEG 20,000, 100 mM MES, pH 6.5; CHMP4C$^{\mathrm{T232}}$: 15% PEG 8,000, 100 mM MES (pH 6.5), 200 mM sodium acetate]. Crystals were flash frozen in nylon loops in cryo-protectant composed of reservoir solutions containing 30% glycerol. Data were collected remotely (58) (0.9794-Å wavelength, 100 K) at the Stanford Synchrotron Radiation Lightsource (SSRL) on beamline 12-2 using a Dectris Pilatus 6M detector. Data were integrated and scaled using AutoXDS (59–61). Both complexes crystallized in space group C121 with one ALIX-Bro1:CHMP4C complex in the asymmetric unit. The crystals diffracted to 1.91-Å resolution (ALIX Bro1-CHMP4C$^{\mathrm{A232}}$) and 1.87-Å resolution (ALIX Bro1-CHMP4C$^{\mathrm{T232}}$). For data collection and refinement statistics, see *SI Appendix*, Table S1.

Structure Determination and Refinement. A model for molecular replacement was generated from the previously determined structure of the ALIX Bro1-CHMP4C$^{\mathrm{A232}}$ complex (Protein Data Bank ID code 3C3R) (27) by removing the coordinates for the CHMP4C peptide and using the ALIX Bro1 structure as a search model (PHASER in PHENIX) (62, 63). CHMP4C helices were built de novo into the electron density for both structures using Coot (64) and were further refined in PHENIX (65, 66) using TLS refinement strategies (67). Final models had no outliers in the Ramachandran plot. Comparison of CHMP4C backbone atoms from both structures reveal that the N termini (residues 221–229) superimpose with a 0.223-Å rmsd, whereas the C termini (residues 229–233) superimpose with a 1.14-Å rmsd. Structures were analyzed and compared using PyMOL Molecular Graphics System version 1.3. Structure coordinates have been deposited in the Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank with accession codes 5V3R (ALIX Bro1-CHMP4C^{A232}) and 5WA1 (ALIX Bro1-CHMP4C^{T232}).

Cell Growth Assays. For assays examining the effect of low-dose aphidicolin and reversine on cell growth, cells were seeded at a density of 2.5×10^4 cells per well in a 24-well plate in duplicate and in a 12-well plate, were treated 8 h later with DMSO, 0.1 μM reversine, or 30 nM aphidicolin, and were left undisturbed for 2 or 4 d (in 24-well plates) or for 7 d (in 12-well plates), except that medium was refreshed after 4 d. Cell number was determined by manual counting at the indicated time points. For analysis of cell growth following p53 depletion, cells were transduced with either control or p53 shRNA for 48 h and then were antibiotic-selected for a further 48 h. Transduced cells were seeded at a density of 2.5×10^4 cells in a single well of a 24-well plate, and cell number was determined at 2, 4, or 7 d. Cell number was determined by manual counting, and cells were reseeded into 12-well plates on day 2 or into six-well plates on day 4.

Karyotyping. HCT116^{WT} or HCT116$^{\delta CHMP4C}$ cells were cultured for 48 h in the presence of DMSO or 0.1 μM reversine and were arrested in metaphase by overnight treatment with 50 ng/mL nocodazole. Cells were harvested, washed, treated in hypotonic buffer (10% FCS in double-distilled H_2O) for 30 min at 37 °C, and were fixed in methanol:acetic acid (3:1 ratio). The fixation solution was replaced four times, and spreads were produced by dropping 16 μL of solution onto a glass slide from a height of 10 cm in a humid environment. Chromosomes from >90 metaphase-arrested cells were counted.

Mitotic Arrest Assays. To examine the role of CHMP4C in mitotic spindle function, HCT116^{WT} or HCT116$^{\delta CHMP4C}$ cells were treated with 50 ng/mL nocodazole, and all cells were collected and analyzed by flow cytometry and Western blotting. To examine the mitotic phosphorylation of CHMP4C, HeLa cells stably expressing the indicated HA-CHMP4C constructs were treated with 2 mM thymidine for 24 h, washed, and treated with 50 ng/mL nocodazole overnight. All cells were collected, and phosphorylation of HA-CHMP4C was determined by Western blotting.

Flow Cytometry Analysis. Cells treated overnight with 200 ng/mL mitomycin C, 50 ng/mL nocodazole, or DMSO were harvested, washed in PBS, and fixed in 1% paraformaldehyde. Cells were washed again and incubated with 50 μg/mL propidium iodide, 100 μg/mL RNase in PBS, and 0.1% Triton X-100. Cell-cycle analysis profiles were acquired using FACS Canto II (BD Biosciences). Twenty thousand cells were counted per condition, and data were analyzed using FlowJo (Tree Start, Inc.). All gating was applied manually. For mitotic arrest experiments all conditions were performed in duplicate, and samples were retained for immunoblotting analysis.

Analysis of Publicly Available Cancer GWAS in the UK Biobank. To determine whether the CHMP4C rs35094336 variant is also associated with other cancer types, we analyzed the publicly available data from 337,208 individuals in the UK Biobank engine. Results were obtained from Global Biobank Engine (https://biobankengine.stanford.edu/; accessed October 2017). We used the following procedure to define cases and controls for cancer GWAS. Individual level ICD-10 codes from the UK Cancer Register, Data-Field 40006, and the National Health Service, Data-Field 41202 in the UK Biobank were mapped to the self-reported cancer codes, Data-Field 20001, as described previously (68). Positive associations are displayed in Table 1. A full review of the resource is available in ref. 26.

ACKNOWLEDGMENTS. We thank Frank Whitby and Chris Hill for assistance with X-ray crystallography; Scott Endicott at the University of Utah Peptide Synthesis core for peptide synthesis; the Fluorescence Microscopy Core Facility at the University of Utah for access to imaging equipment, which was obtained using National Center for Research Resources Shared Equipment Grant 1S10RR024761-01; Leandro Ventimiglia for sharing technical advice and cell lines; Jesus Gil (Imperial College London) and Pierre Vantourout (Kings College London; KCL) for the control and p53 shRNA- and CRISPR Cas9-containing lentiviral packaging vectors, respectively; Caroline Ogilvie (KCL) and Jennifer Pratt (Viapath) for guidance with the karyotyping experiments; the Manuel Rivas laboratory for making the UK Biobank resource available; and the Nikon Imaging Centre at KCL for technical support. The UK National Institute for Health Research Comprehensive Biomedical Research Center at KCL provided an equipment grant. Funding for this project was provided by NIH Grant R01 GM112080 (to W.I.S.), by the Wellcome Trust Grant WT102871MA (to J.M.-S.), and by support in conjunction with Grant P30 CA042014 awarded to Huntsman Cancer Institute from the Cell Response and Regulation Program at the Huntsman Cancer Institute (W.I.S. and K.S.U.). D.M.W. was supported in part by American Cancer Society Postdoctoral Fellowship PF-14-102-01-CSM. Further support was provided by Severo Ochoa Program Grant SEV-2011-00067, a Sara Borrell Fellowship from the Instituto Carlos III, and a Beatriu de Pinós Fellowship from the Agency for Management of University and Research Grants (to J.M.M.). Use of the SSRL at the SLAC National Accelerator Laboratory, is supported by the US Department of Energy (DOE), Office of Science, Office of Basic Energy Sciences under Contract DE-AC02-76SF00515. The SSRL Structural Molecular Biology Program is supported by the DOE, Office of Biological and Environmental Research and by the NIH, National Institute of General Medical Sciences (NIGMS) (including Grant P41GM103393). The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official views of the NIGMS or the NIH.

1. Caballe A, Martin-Serrano J (2011) ESCRT machinery and cytokinesis: The road to daughter cell separation. *Traffic* 12:1318–1326.
2. Frémont S, Romet-Lemonne G, Houdusse A, Echard A (2017) Emerging roles of MICAL family proteins–From actin oxidation to membrane trafficking during cytokinesis. *J Cell Sci* 130:1509–1517.
3. Mierzwa B, Gerlich DW (2014) Cytokinetic abscission: Molecular mechanisms and temporal control. *Dev Cell* 31:525–538.
4. Carlton JG, Martin-Serrano J (2007) Parallels between cytokinesis and retroviral budding: A role for the ESCRT machinery. *Science* 316:1908–1912.
5. Morita E, et al. (2007) Human ESCRT and ALIX proteins interact with proteins of the midbody and function in cytokinesis. *EMBO J* 26:4215–4227.
6. Guizetti J, et al. (2011) Cortical constriction during abscission involves helices of ESCRT-III-dependent filaments. *Science* 331:1616–1620.
7. Elia N, Sougrat R, Spurlin TA, Hurley JH, Lippincott-Schwartz J (2011) Dynamics of endosomal sorting complex required for transport (ESCRT) machinery during cytokinesis and its role in abscission. *Proc Natl Acad Sci USA* 108:4846–4851.
8. Schöneberg J, Lee IH, Iwasa JH, Hurley JH (2017) Reverse-topology membrane scission by the ESCRT proteins. *Nat Rev Mol Cell Biol* 18:5–17.
9. Christ L, Raiborg C, Wenzel EM, Campsteijn C, Stenmark H (2017) Cellular functions and molecular mechanisms of the ESCRT membrane-scission machinery. *Trends Biochem Sci* 42:42–56.
10. Scourfield EJ, Martin-Serrano J (2017) Growing functions of the ESCRT machinery in cell biology and viral replication. *Biochem Soc Trans* 45:613–634.
11. Carlton JG, Agromayor M, Martin-Serrano J (2008) Differential requirements for Alix and ESCRT-III in cytokinesis and HIV-1 release. *Proc Natl Acad Sci USA* 105:10541–10546.
12. Lee HH, Elia N, Ghirlando R, Lippincott-Schwartz J, Hurley JH (2008) Midbody targeting of the ESCRT machinery by a noncanonical coiled coil in CEP55. *Science* 322: 576–580.
13. Mierzwa BE, et al. (2017) Dynamic subunit turnover in ESCRT-III assemblies is regulated by Vps4 to mediate membrane remodelling during cytokinesis. *Nat Cell Biol* 19: 787–798.
14. Agromayor M, Martin-Serrano J (2013) Knowing when to cut and run: Mechanisms that control cytokinetic abscission. *Trends Cell Biol* 23:433–441.
15. Norden C, et al. (2006) The NoCut pathway links completion of cytokinesis to spindle midzone function to prevent chromosome breakage. *Cell* 125:85–98.
16. Mendoza M, et al. (2009) A mechanism for chromosome segregation sensing by the NoCut checkpoint. *Nat Cell Biol* 11:477–483.
17. Steigemann P, et al. (2009) Aurora B-mediated abscission checkpoint protects against tetraploidization. *Cell* 136:473–484.
18. Mackay DR, Ullman KS (2015) ATR and a Chk1-Aurora B pathway coordinate post-mitotic genome surveillance with cytokinetic abscission. *Mol Biol Cell* 26:2217–2226.
19. Nähse V, Christ L, Stenmark H, Campsteijn C (2017) The abscission checkpoint: Making it to the final cut. *Trends Cell Biol* 27:1–11.
20. Carlton JG, Caballe A, Agromayor M, Kloc M, Martin-Serrano J (2012) ESCRT-III governs the Aurora B-mediated abscission checkpoint through CHMP4C. *Science* 336: 220–225.
21. Capalbo L, et al. (2012) The chromosomal passenger complex controls the function of endosomal sorting complex required for transport-III Snf7 proteins during cytokinesis. *Open Biol* 2:120070.
22. Morita E, et al. (2011) ESCRT-III protein requirements for HIV-1 budding. *Cell Host Microbe* 9:235–242.
23. Thoresen SB, et al. (2014) ANCHR mediates Aurora-B-dependent abscission checkpoint control through retention of VPS4. *Nat Cell Biol* 16:550–560.
24. Caballe A, et al. (2015) ULK3 regulates cytokinetic abscission by phosphorylating ESCRT-III proteins. *eLife* 4:e06547.
25. Pharoah PD, et al. (2013) GWAS meta-analysis and replication identifies three new susceptibility loci for ovarian cancer. *Nat Genet* 45:362–370, 370e1-2.

26. Bycroft C, et al. (2017) Genome-wide genetic data on ~500,000 UK Biobank participants. bioRxiv:10.1101/166298. Preprint, posted July 20 2017.
27. McCullough J, Fisher RD, Whitby FG, Sundquist WI, Hill CP (2008) ALIX-CHMP4 interactions in the human ESCRT pathway. *Proc Natl Acad Sci USA* 105:7687–7691.
28. Ballesteros JA, Deupi X, Olivella M, Haaksma EE, Pardo L (2000) Serine and threonine residues bend alpha-helices in the chi(1) = g(-) conformation. *Biophys J* 79:2754–2760.
29. Christ L, et al. (2016) ALIX and ESCRT-I/II function as parallel ESCRT-III recruiters in cytokinetic abscission. *J Cell Biol* 212:499–513.
30. Mackay DR, Makise M, Ullman KS (2010) Defects in nuclear pore assembly lead to activation of an Aurora B-mediated abscission checkpoint. *J Cell Biol* 191:923–931.
31. Lafaurie-Janvore J, et al. (2013) ESCRT-III assembly and cytokinetic abscission are induced by tension release in the intercellular bridge. *Science* 339:1625–1629.
32. Burrell RA, et al. (2013) Replication stress links structural and numerical cancer chromosomal instability. *Nature* 494:492–496.
33. Olmos Y, Hodgson L, Mantell J, Verkade P, Carlton JG (2015) ESCRT-III controls nuclear envelope reformation. *Nature* 522:236–239.
34. Vietri M, et al. (2015) Spastin and ESCRT-III coordinate mitotic spindle disassembly and nuclear envelope sealing. *Nature* 522:231–235.
35. Petsalaki E, Dandoulaki M, Zachos G (2018) The ESCRT protein Chmp4c regulates mitotic spindle checkpoint signaling. *J Cell Biol* 217:861–876.
36. Lukas C, et al. (2011) 53BP1 nuclear bodies form around DNA lesions generated by mitotic transmission of chromosomes under replication stress. *Nat Cell Biol* 13: 243–253.
37. Amaral N, et al. (2016) The Aurora-B-dependent NoCut checkpoint prevents damage of anaphase bridges after DNA replication stress. *Nat Cell Biol* 18:516–526.
38. Janssen A, van der Burg M, Szuhai K, Kops GJ, Medema RH (2011) Chromosome segregation errors as a cause of DNA damage and structural chromosome aberrations. *Science* 333:1895–1898.
39. Santaguida S, Tighe A, D'Alise AM, Taylor SS, Musacchio A (2010) Dissecting the role of MPS1 in chromosome biorientation and the spindle checkpoint through the small molecule inhibitor reversine. *J Cell Biol* 190:73–87.
40. Funk LC, Zasadil LM, Weaver BA (2016) Living in CIN: Mitotic infidelity and its consequences for tumor promotion and suppression. *Dev Cell* 39:638–652.
41. Cancer Genome Atlas Research Network (2011) Integrated genomic analyses of ovarian carcinoma. *Nature* 474:609–615.
42. Toledo LI, et al. (2011) A cell-based screen identifies ATR inhibitors with synthetic lethal properties for cancer-associated mutations. *Nat Struct Mol Biol* 18:721–727.
43. Agarwal ML, Agarwal A, Taylor WR, Stark GR (1995) p53 controls both the G2/M and the G1 cell cycle checkpoints and mediates reversible growth arrest in human fibroblasts. *Proc Natl Acad Sci USA* 92:8493–8497.
44. Capalbo L, et al. (2016) Coordinated regulation of the ESCRT-III component CHMP4C by the chromosomal passenger complex and centralspindlin during cytokinesis. *Open Biol* 6:160248.
45. Lata S, et al. (2009) Structure and function of ESCRT-III. *Biochem Soc Trans* 37:156–160.
46. Tang S, et al. (2016) ESCRT-III activation by parallel action of ESCRT-I/II and ESCRT-0/Bro1 during MVB biogenesis. *eLife* 5:e15507.
47. Tang S, et al. (2015) Structural basis for activation, assembly and membrane binding of ESCRT-III Snf7 filaments. *eLife* 4:e12548.
48. Passerini V, et al. (2016) The presence of extra chromosomes leads to genomic instability. *Nat Commun* 7:10754.
49. Santaguida S, et al. (2017) Chromosome mis-segregation generates cell-cycle-arrested cells with complex karyotypes that are eliminated by the immune system. *Dev Cell* 41: 638–651.e5.
50. Kalsbeek D, Golsteyn RM (2017) G2/M-Phase checkpoint adaptation and micronuclei formation as mechanisms that contribute to genomic instability in human cells. *Int J Mol Sci* 18:E2344.
51. Lezaja A, Altmeyer M (2017) Inherited DNA lesions determine G1 duration in the next cell cycle. *Cell Cycle* 17:24–32.
52. Arora M, Moser J, Phadke H, Basha AA, Spencer SL (2017) Endogenous replication stress in mother cells leads to quiescence of daughter cells. *Cell Rep* 19:1351–1364.
53. Gatz SA, Wiesmüller L (2006) p53 in recombination and repair. *Cell Death Differ* 13: 1003–1016.
54. Li K, et al. (2015) CHMP4C disruption sensitizes the human lung cancer cells to irradiation. *Int J Mol Sci* 17:E18.
55. Cer RZ, Mudunuri U, Stephens R, Lebeda FJ (2009) IC 50-to-K i: A web-based tool for converting IC 50 to K i values for inhibitors of enzyme activity and ligand binding. *Nucleic Acids Res* 37(Suppl 2):W441–W445.
56. Olmos Y, Perdrix-Rosell A, Carlton JG (2016) Membrane binding by CHMP7 coordinates ESCRT-III-dependent nuclear envelope reformation. *Curr Biol* 26:2635–2641.
57. Fisher RD, et al. (2007) Structural and biochemical studies of ALIX/AIP1 and its role in retrovirus budding. *Cell* 128:841–852.
58. Soltis SM, et al. (2008) New paradigm for macromolecular crystallography experiments at SSRL: Automated crystal screening and remote data collection. *Acta Crystallogr D Biol Crystallogr* 64:1210–1221.
59. Gonzalez A, Tsai Y (2010) A Quick XDS Tutorial for SSRL. Available at smb.slac.stanford.edu/facilities/software/xds/#autoxds_script. Accessed March 13, 2016.
60. Kabsch W (2010) XDS. *Acta Crystallogr D Biol Crystallogr* 66:125–132.
61. Kabsch W (2010) Integration, scaling, space-group assignment and post-refinement. *Acta Crystallogr D Biol Crystallogr* 66:133–144.
62. Adams PD, et al. (2010) PHENIX: A comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr* 66:213–221.
63. McCoy AJ, et al. (2007) Phaser crystallographic software. *J Appl Crystallogr* 40: 658–674.
64. Emsley P, Lohkamp B, Scott WG, Cowtan K (2010) Features and development of Coot. *Acta Crystallogr D Biol Crystallogr* 66:486–501.
65. Afonine PV, et al. (2012) Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr D Biol Crystallogr* 68:352–367.
66. Chen VB, et al. (2010) MolProbity: All-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr* 66:12–21.
67. Howlin B, Butler SA, Moss DS, Harris GW, Driessen HPC (1993) TLSANL: TLS parameter-analysis program for segmented anisotropic refinement of macromolecular structures. *J Appl Crystallogr* 26:622–624.
68. DeBoever C, et al. (2017) Medical relevance of protein-truncating variants across 337,208 individuals in the UK Biobank study. bioRxiv:10.1101/179762. Preprint, posted September 2, 2017.

Appendix 3. A functional IFN-λ4-generating DNA polymorphism could protect older asthmatic women from aeroallergen sensitization and associate with clinical features of asthma.

Chinnaswamy S, Wardzynska A, Pawelczyk M, Makowska J, Skaaby T, Mercader JM, Ahluwalia TS, Grarup N, Guindo-Martinez M, Bisgaard H, Torrents D, Linneberg A, Bønnelykke K, Kowalski ML. *Scientific Reports* 7, 10500 (2017).

Contribution:

- Replication of rs12979860 and rs8099917 variants for allergic rhinitis in GERA cohort, after imputing genotypes using 1000G phase 3, U10K and GoNL as reference panels. Only the additive model was analyzed in the association test.

www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

Received: 14 November 2016
Accepted: 10 August 2017
Published online: 05 September 2017

A functional IFN-λ4-generating DNA polymorphism could protect older asthmatic women from aeroallergen sensitization and associate with clinical features of asthma

Sreedhar Chinnaswamy[1,7], Aleksandra Wardzynska[2], Malgorzata Pawelczyk[2], Joanna Makowska[2,3], Tea Skaaby[4], Josep M. Mercader[5], Tarunveer S. Ahluwalia[6], Niels Grarup[11], Marta Guindo-Martinez[5], Hans Bisgaard[6], David Torrents[5,10], Allan Linneberg[4,8,9], Klaus Bønnelykke[6] & Marek L. Kowalski[2,7]

Lambda interferons (IFNLs) have immunomodulatory functions at epithelial barrier surfaces. IFN-λ4, a recent member of this family is expressed only in a subset of the population due to a frameshift-causing DNA polymorphism rs368234815. We examined the association of this polymorphism with atopy (aeroallergen sensitization) and asthma in a Polish hospital-based case-control cohort comprising of well-characterized adult asthmatics (n = 326) and healthy controls (n = 111). In the combined cohort, we saw no association of the polymorphism with asthma and/or atopy. However, the IFN-λ4-generating ΔG allele protected older asthmatic women (>50 yr of age) from atopic sensitization. Further, ΔG allele significantly associated with features of less-severe asthma including bronchodilator response and corticosteroid usage in older women in this Polish cohort. We tested the association of related *IFNL* locus polymorphisms (rs12979860 and rs8099917) with atopy, allergic rhinitis and presence/absence of asthma in three population-based cohorts from Europe, but saw no significant association of the polymorphisms with any of the phenotypes in older women. The polymorphisms associated marginally with lower occurrence of asthma in men/older men after meta-analysis of data from all cohorts. Functional and well-designed replication studies may reveal the true positive nature of these results.

Type III interferons (IFNs) or IFN-λs (or IFNLs) are known to play critical roles in innate and adaptive immune responses to viral infections. Several recent reports have implicated IFN-λs as the 'guardians' of the epithelial

[1]National Institute of Biomedical Genomics, PO:N.S.S, Kalyani, 741251, West Bengal, India. [2]Dept. of Immunology, Rheumatology & Allergy, Medical University of Lodz, Lodz, 92-213, Poland. [3]Department of Rheumatology, Medical University of Lodz, 92-003, Lodz, Poland. [4]Research Centre for Prevention and Health, the Capital Region of Denmark, Copenhagen, Denmark. [5]Barcelona Supercomputing Center (BSC). Joint BSC-CRG-IRB Research Program in Computational Biology, 08034, Barcelona, Spain. [6]COPSAC, Copenhagen Prospective Studies on Asthma in Childhood, Herlev and Gentofte Hospital, University of Copenhagen, Copenhagen, Denmark. [7]Chair of Clinical Immunology and Microbiology, Healthy Aging Research Center, Medical University of Lodz, 251 Pomorska Str, 92-213, Lodz, Poland. [8]Department of Clinical Experimental Research, Rigshospitalet, Denmark. [9]Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [10]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain. [11]The Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, 2100, Copenhagen, Denmark. Correspondence and requests for materials should be addressed to S.C. (email: sc2@nibmg.ac.in)

barrier surfaces that encompass large regions of the human body that include respiratory, gastrointestinal and reproductive tracts[1]. Single nucleotide polymorphisms (SNPs) at the *IFNL* locus on human chromosome 19 were discovered a few years ago to be associated with chronic hepatitis C virus (HCV) infections[2–4]. These SNPs, discovered initially by genome-wide association studies (GWAS) to be associated with treatment-induced clearance of HCV, got widely validated and were subsequently found to be associated with different human diseases, both viral and non-viral in origin (reviewed in 5). The immunomodulatory role of IFN-λs is thought to be the molecular mechanism behind these associations[5]. Among the different SNPs at the *IFNL* locus, that are in strong linkage disequilibrium (LD) in most populations, a dinucleotide polymorphism rs368234815 (TT/ΔG) has emerged as the 'causal' variant[5,6]. The ΔG allele of the variant causes a shift in the open reading frame of the *IFNL4* gene, giving rise to a new IFN-λ called IFN-λ4[6]. IFN-λ4 has undergone purifying selection during human evolution with highest frequency of the functional IFN-λ4-generating ΔG allele in the African population (0.78) and lowest in the East Asian population (0.06)[6,7]; 50-60% of the European population carries at least one copy of the functional gene[8]. IFN-λ4 is a potent antiviral cytokine with structural and functional similarity to IFN-λ1, 2 and 3[8,9]. The latter are known to modulate adaptive immunity favoring a Th1 predominant response[10–12]. Even though there is no direct evidence yet, IFN-λ4 is also expected to participate in shaping and maintaining innate and adaptive immune responses at the epithelial lining of the respiratory tract owing to its similarities with other IFN-λs. The interactions of IFN-λs with the newly discovered innate lymphoid cells (ILCs)[13] are likely to hold the key to improving our understanding of diseases like allergy and asthma[5].

Atopy, defined as increased predisposition to generate specific IgE to common allergens, and diagnosed in an individual subject by presence of aeroallergen sensitization, may lead to Th2-driven immune response to harmless allergens resulting in development of allergic diseases including asthma. Exacerbations in asthma are commonly associated with respiratory virus infections, and impaired innate immune responses have been documented in asthma patients[14,15]. Since both the development of atopic predisposition and asthma exacerbations may involve respiratory viral infections, we undertook this study with the hypothesis that these complex disorders are influenced by the IFN-λ4-generating polymorphism.

Results and Discussion

ΔG allele of rs368234815 protects older women asthmatics from atopy in a Polish hospital-based case-control cohort. We used a Polish discovery cohort that included 326 well-characterized adult bronchial asthmatics recruited from the university asthma clinic, medical university of Lodz. Cases were heterogeneous with respect to the disease control and severity. Controls comprised of 111 volunteers without any chronic respiratory disorders representing a sample of general population. Information on atopic sensitization (determined by skin prick test or SPT to a panel of inhalant allergens) was available for 384 subjects (273 asthmatics and 111 controls). The patient and control group characteristics are shown in Table 1. Owing to its functional nature we chose and genotyped the functional IFN-λ4-generating polymorphism rs368234815 in this cohort. Power calculations showed that we had enough sample size to give >80% power (at a significance level of $p = 0.05$) to detect an effect size of 1.55 and 1.6 in atopy and asthma, respectively.

The Minor Allele Frequency (MAF) in the combined population of controls and asthma patients (cases) was 0.34; the distribution of genotypes was: 38%, 53%, 9% and 44%, 44%, 12% for TT/TT, TT/ ΔG and ΔG/ΔG in controls and cases respectively. Both controls and cases were in Hardy-Weinberg equilibrium (HWE) ($P > 0.05$) either individually or as a combined group. In the combined group analysis there was no association of the polymorphism with asthma or atopy under either dominant or recessive models of inheritance (Table 2). However, by using log-linear regression we observed statistically significant interactions between the polymorphism, atopy, age and gender (interactions up to three factors; polymorphism, age and gender $p = 0.037$; polymorphism, gender and atopy $p = 0.035$). To dissect these interactions further, we stratified our atopy data into several groups based on age and gender (Table 3) and applied bonferroni correction to avoid false positives arising due to multiple testing. Since our study group had a majority of women (Table 1), we chose 50 years as a divider of age in the combined cohort for our stratified analysis as it is also the age of attainment of menopause[16] (overall median age at natural menopause in Poland is 51.25 years[17],). Only among the >50 yr sub-group of women (older women) we saw a significant association of the polymorphism with atopy after multiple testing correction under a dominant model of inheritance for the minor allele. The ΔG allele conferred protection from allergic sensitization in older asthmatic women. The significance of association was retained in both the asthmatic older women and in the combined group of asthmatic and control older women, but not in the control older women's sub-group when tested alone, in both univariate and multivariate analysis (Table 4).

Even though the association was significant in the older women's sub-group, small sample sizes may have prevented us from appreciating the effect of the polymorphism on atopy in the remaining sub-groups. In addition, a validation of the observed effect on the older women's sub-group is required from other populations and/or geographical regions. To examine this and to investigate the association of the *IFNL* locus SNPs with atopy, asthma and related illnesses, we used data from the Genetic Epidemiology Research on Adult Health and Aging (GERA, dbGaP Study Accession: phs000674.v1.p1)[18], Inter99[19], Health2006[20] and COPSAC[21,22] cohorts. The characteristics of the different study cohorts are briefly described in Table 5. The GERA cohort consisted of predominantly older participants (average age 63 yr, range 18 to over 100 yr) and hence all participants in this cohort were considered as being older (>50 yr) for analysis.

Genotype information for a related *IFNL* polymorphism rs12979860[5] was available in the other European cohorts, hence, we tested the association of atopy with rs12979860 in the Inter99, Health2006 and the COPSAC cohorts in four different sub-groups based on both age and gender and did a meta-analysis using a random-effects model on all the cohorts where atopy data was available including the Polish cohort (Fig. 1). Even though, the Polish cohort tested for rs368234815 and the other cohorts for rs12979860 the data could be compared since a strong LD ($r^2 = 0.98$) between them in the European population has been recorded[6]; furthermore, the MAFs in

	asthma, N = 326	controls, N = 111
age, years*	**59.2 ± 15.9 (18–94)**	**52.3 ± 18.3 (24–81)**
gender, women, n (%)	201 (61.7%)	72 (64.9%)
Men > 50 yr average age (range)/n	67.3 ± 9.8 (50–87)/92	65.2 ± 9.7 (50–80)/18
Men < 50 yr average age (range)/n	34.2 ± 7.5 (18–48)/33	32.9 ± 5.9 (24–48)/21
Women > 50 yr average age (range)/n	67.3 ± 9.3 (50–94)/144	67 ± 7.2 (51–81)/47
Women < 50 yr average age (range)/n*	**40 ± 6.6 (24–49)/57**	**34.1 ± 6.5 (26–49)/25**
Atopy, n/N tested (%)*	**141/273 (51.6%)**	**39/111 (35.1%)**
FEV1% pred.	75.3 ± 24.1 (17.5–129.5)	–
FEV1%/FVC	68 ± 13.1 (25.6–99.4)	–
ACT, points	17.8 ± 5.5(4–25)	–
FeNO (ppb)	30.6 ± 17.8 (2–184)	–
Asthma control (according to GINA 2016)		
Controlled, n (%)	72 (22%)	–
partly controlled, n (%)	96 (30%)	–
Uncontrolled, n (%)	156 (48%)	–
Current asthma treatment		
ICS, n (%)	240 (73.6%)	–
low dose**, n (%)	28 (8.6%)	
medium dose**, n (%)	120 (36.8%)	
high dose dose**, n (%)	92 (28.2%)	
Oral steroids, n (%)	23 (7.1%)	–
LABA, n (%)	197 (60.4%)	
Leukotriene antagonists, n (%)	64 (19.6%)	
at least 1 exacerbation/last year, n (%)	170 (52.1%)	–

Table 1. Characteristics of the study subjects in the Polish cohort. Values are presented as arithmetic means + SD, (range); *statistically significant difference between groups, $p < 0.05$; values shown in bold; **according to GINA 2016.

	Genotype (n, %)				Dominant model (TT/ΔG + ΔG/ΔG) vs TT/TT	Recessive model ΔG/ΔG vs (TT/ΔG + TT/TT)
	TT/TT	TT/ΔG	ΔG/ΔG	Total	OR (95% CI); p-value	OR (95% CI); p-value
Asthma	145, 44.5	142, 43.5	39, 12	326, 100	0.75 (0.48–1.18); 0.26	1.37 (0.66–2.85); 0.48
Controls	42, 37.8	59, 53.2	10, 9	111, 100		
Atopy +	81, 45	80, 44.4	19, 10.6	180, 100	0.72 (0.48–1.09); 0.14	0.84 (0.44–1.59); 0.63
Atopy −	76, 37.2	103, 50.5	25, 12.3	204, 100		
Asthma*						
Atopy+	64, 45.4	61, 43.3	16, 11.3	141, 100	0.75 (0.46–1.22); 0.27	0.81 (0.39–1.66); 0.58
Atopy−	51, 38.6	63, 47.7	18, 13.7	132, 100		
Controls						
Atopy+	17, 43.6	19, 48.7	3, 7.7	39, 100	0.68 (0.31–1.52); 0.41	0.77 (0.18–3.17); 1
Atopy−	25, 34.7	40, 55.5	7, 9.7	72, 100		

Table 2. Association of the functional IFN-λ4-generating polymorphism rs368234815 with asthma and atopy in the Polish cohort. *Skin prick test was not carried out for 53 asthmatics as they were under treatment with antihistamines, antidepressants or there were other contraindications.

the cohorts were similar (Table 5). Moreover, recent data shows that the SNP rs12979860 is the 'best tag-SNP' for the functional polymorphism rs368234815 due to a common underlying linkage structure at the *IFNL* locus[23]. The results show that we failed to replicate our findings from the older women's sub-group of the Polish cohort, in other cohorts (Fig. 1). Further, except for a nominal association ($p = 0.04$) with an additive genetic model in the COPSAC younger men's cohort no other sub-group in any of the remaining cohorts showed any significant association with atopy (Fig. 1). The older women's and younger men's sub-groups had significant heterogeneity in the effect of the polymorphisms ($p = 0.008$ and $p = 0.028$ respectively) between the different cohorts and no significant association was seen after meta-analysis. Even in the other sub-groups where there was no significant heterogeneity between the studies, no significant effect of the polymorphism on atopy was noted in the meta-analysis (Fig. 1). Similar results were seen using the fixed-effect model (Suppl. Figure 1)

Group	Sub-group	N	OR (95% CI)	p-value	p-value*
All	All ages	384	0.72 (0.48–1.09)	0.123	1
	>50 yr	261	0.51 (0.31–0.85)	**0.009**	0.081
	<50 yr	123	1.26 (0.59–2.68)	0.548	1
Men	All ages	145	1.28 (0.66–2.47)	0.458	1
	>50 yr	94	0.81 (0.35–1.88)	0.394	1
	<50 yr	51	1.53 (0.41–5.73)	—	—
Women	All ages	239	0.51 (0.3–0.86)	**0.012**	0.108
	>50 yr	167	0.37 (0.19–0.72)	**0.002**	**0.018**
	<50 yr	72	1.01 (0.39–2.62)	1	1

Table 3. The ΔG allele of rs368234815 associates with protection from atopy in older women in the Polish cohort. *After bonferroni correction; A dominant model of inheritance for the minor allele (TT/ΔG + ΔG/ΔG vs TT/TT) was used to compute OR.

	Dominant Model (TT/ΔG + ΔG/ΔG) vs TT/TT					
Group	OR (crude)	95% CI	p	OR (adjusted)*	95% CI	p
[1]Women (Asthma) > 50 yr N = 128; atopy, n = 55	**0.398**	**0.188–0.841**	**0.016**	**0.402**	**0.184–0.877**	**0.021**
[2]Women (Controls) > 50 yr N = 39; atopy, n = 11	0.400	0.098–1.631	0.202	0.366	0.083–1.605	0.171
[3]Women (Asthma + Controls) > 50 yr N = 167; atopy, n = 66	**0.379**	**0.198–0.724**	**0.003**	**0.409**	**0.208–0.803**	**0.016**

Table 4. Association of the functional IFN-λ4-generating polymorphism rs368234815 with atopy in older women in the Polish cohort. *[1], *[2] for age; *[3] for asthma status and age. A dominant model of inheritance is shown with significant results (p < 0.05) in bold. No significant association under any group/sub/group was seen using the recessive model of inheritance.

Next, we tested if the *IFNL* polymorphisms associated with presence of asthma in different cohorts, again by stratification analysis based on both age and gender and by meta-analysis using random-effects model (Fig. 2). No significant heterogeneity in the effect between different cohorts in the sub-groups was noted. None of the sub-groups from any of the cohorts including the Polish cohort, showed any significant association with asthma. Interestingly, the older men's sub-group showed a significant effect of the polymorphism(s) on presence/absence of asthma after meta-analysis. The minor allele (which gives rise to/linked to the allele that can express a functional IFN-λ4) had a protective effect on asthma in the older men. However, given the large sample size involved in the GERA cohort such low significance of association suggests a very small effect on the phenotype and/or phenotypic heterogeneity. Similar results were seen with a fixed-effects model (Suppl. Figure 2).

Since the GERA cohort also included data on allergic rhinitis (AR) we tested if rs12979860 and another related SNP rs8099917 associated with it (Table 6) in gender-stratified sub-groups. No significant association of either SNP with AR was evident in males or females.

We extended our meta-analysis to see if the *IFNL* polymorphisms associated with atopy and asthma in: 1) all individuals irrespective of age and gender and 2) gender-specific and age-specific strata from the other European cohorts. All the cohorts included for this meta-analysis had the same SNP rs12979860 either directly genotyped or imputed and the MAFs in each of the cohorts were similar (Table 5). All studies were homogenous for the effect of the polymorphism except for the all men's group when testing for atopy (Fig. 3A). A significant association with asthma was detected only in the sub-group of 'all men' (Fig. 3B) similar to the effect seen for asthma in the older men's sub-group in Fig. 2. Since, we did not have enough participants in the younger men's group compared to the large sample size of older men from the GERA cohort, the modifying effect of age, if any, on the association of the polymorphism with asthma could not be reliably tested.

In conclusion, the results from the Polish cohort on the association of the polymorphism(s) with atopy could not be replicated in other European cohorts while a significant association of the polymorphism(s) was seen with asthma in men/older men when data was meta-analyzed. In instances where a significant association was seen, the minor allele showed a protective phenotype.

Functional IFN-λ4-generating ΔG allele associates with less severe features of asthma in the older women's sub-group of the Polish cohort. We refocused our interest on the Polish older women's sub-group where we saw a significant association of the IFN-λ4-generating polymorphism with atopy. We observed that this sub-group which comprised of both asthmatic and control older women had a majority (75%) of asthmatics (Table 4). Therefore we were interested to examine if any of the clinical features of asthma are also associated with the polymorphism (Table 7 and Table 8) in the older asthmatic women. We saw a significant association between the polymorphism and positive bronchial reversibility (or bronchodilator response, BDR) test and usage of corticosteroids (CS) for treatment (Table 8). Older women with at least one copy of the IFN-λ4-generating ΔG allele were less likely to be treated with iCS (inhaled CS) (OR = 0.35) and oCS (oral CS) (OR = 0.16) but were more likely to show a positive BDR (OR = 2.58) by univariate analysis. To test for

Cohort	Type	Genotype* tested	MAF	Phenotype data available	Total no. of participants	Ref.	Remarks
Inter 99	Population-based	rs12979860 (g)	0.34	Atopy Asthma	5341	19	Atopy determined by serum specific IgE; self-reported doctor-diagnosed asthma
Health 2006	Population-based	rs12979860 (g)	0.34	Atopy Asthma	3134	20	
COPSAC	Case-control (hospital-based; parents of asthmatic children)	rs12979860 (g)	0.30	Atopy Asthma	551	21	All participants below 50 yr of age; Atopy determined by serum specific IgE
GERA	Population-based	rs12979860 (i); rs8099917 (i)	0.32; 0.19	Allergic rhinitis, Asthma	56637	18	Average age 63 yr (range18 to over 100 yr)
Polish	Hospital-based Case-control	rs368234815 (g)	0.34	Atopy, asthma	437	—	Atopy determined by SPT; asthma treated and monitored

Table 5. Characteristics of the different study cohorts used in the study. *g-genotyped; i-imputed; MAF-minor allele frequency.

confounders we carried out multivariate regression analysis and found that atopy could be partly mediating the association between the polymorphism and BDR (Table 8).

In summary, results from the Polish cohort demonstrated that IFN-λ4-generating ΔG allele protected a sub-population of asthmatic patients, specifically older women, from allergic sensitization. The IFN-λ4-generating ΔG allele also associates with less inhaled and oral CS usage, suggesting that it may be related with lesser disease severity in this group of severe asthmatics (78% of overall cases had either partially controlled or uncontrolled asthma, Table 1). Further, it goes on to suggest that the ΔG allele carriers possibly had reduced airway inflammation compared to the pseudogenizing TT/TT genotype carriers.

Inherent differences between the discovery and replication cohorts may have been responsible for non-replication of our results from the Polish cohort. The association of *IFNL* polymorphism with atopy in the Polish older women's cohort survived correction for multiple testing that was carried out as a measure to negate the limitation of post-hoc analysis of data (Table 3). Further, the results from the Polish study are not due to a 'cohort effect' since the association of the polymorphism with atopy is within the older women's 'cohort' and not between an older and a younger 'cohort'. The association of the polymorphism in older women was not just with atopy but also translated to association with some important determinants of severity of asthma like BDR and CS usage. Even though we did not carry out multiple testing correction for our association tests with the polymorphism and clinical features of asthma (Table 8) we feel the phenotypes are not independent of the atopic sensitization phenotype that we identified during our initial analysis (Tables 3 and 4), and therefore less likely to be false positives. For example, atopy-related eosinophilic inflammation and increased CS usage frequency among patients may be correlated[24]; this is apart from the fact that atopy was a confounder with the association involving bronchial reversibility tests (Table 8). Therefore the results from the Polish study regarding association of the polymorphism with atopy and clinical features of asthma in older women, seem to be consistent with each other.

It is important to analyze and interpret our results from the Polish study in the context that they were not replicated in other cohorts of the European population. It should be noted that the replication study had several limitations. Firstly, the findings we made in the Polish cohort were from a hospital-based case-control design while all the three replication cohorts (except COPSAC cohort) were population-based (Table 5). Secondly, 72% of the Polish cohort (among the 384 subjects with information on atopy) were asthma cases and specifically the older women's sub-group in this cohort had substantially more (75%) asthmatics than controls while the replication cohorts had 16% asthmatics among older men and women in GERA cohort (10% in AR-negative controls and 34% in AR-positives; AR-negative controls were 77% and AR -positive were 33% in the cohort); 9.6% and 11.6% of older women had asthma in the Inter99 and Heath2006 cohorts respectively. It is possible that the association with atopy in the Polish cohort had an underlying link with asthma that could not be accounted for in the replication study. In support of this, significance of association decreased when data was adjusted for asthma status in the older women's sub-group in the Polish study (Table 4). Thirdly, the Polish cohort was enriched with severe asthmatics (22% controlled, 30% partly controlled and 48% uncontrolled according to GINA 2016; up to 65% of patients were in medium or high dose iCS; 60% were on long acting β-agonists; Table 1) while this information was not available for the replication cohorts. Fourthly, similar methods were not used to test for the phenotypes in all the cohorts (Ex. SPT in the Polish cohort and serum specific IgE in the Danish cohorts to test atopy). While in the discovery cohort that had a hospital-based case-control design, we saw a significant association of the

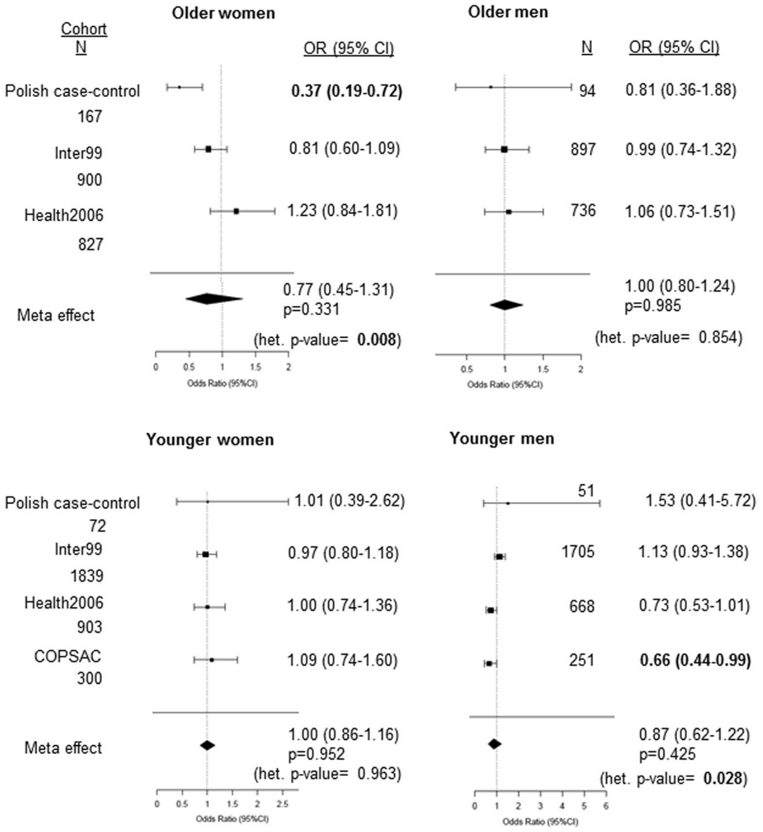


Figure 1. Forest plots showing association of *IFNL* polymorphisms and atopy in sub-groups based on age and gender in different cohorts. The tag-SNP rs12979860 was used in all other cohorts except the Polish cohort that tested for functional polymorphism rs368234815. A dominant model of inheritance for the minor allele (Ex. TT/ ΔG + ΔG/ ΔG vs TT/TT for rs368234815) was used in all cohorts except in the COPSAC cohort that used an additive model to obtain Odds Ratios (OR) shown as a forest plot. p-value < 0.05 was considered as significant and is in bold. For stratification based on age 50 yr was used as a cut-off mark. Het.-heterogeneity.

polymorphism with atopy (in an asthma-enriched background) and severe asthma features in older women, in the replication study we tested for association with atopy, allergic rhinitis and presence/absence of asthma in a population-based cohort. These discrepancies in endpoint phenotypes and the intrinsic differences in the composition of the Polish and the replication cohorts may have led to non-replication of the findings.

Our findings from the Polish cohort need to be further validated in well-designed replication studies and functional studies. It is likely that the functional IFN-λ4-generating dinucleotide variant rs368234815 may associate with a specific endotype of severe asthma in older patients. It is possible that a sustained course of inflammation of the airways that happens in older asthmatics, likely in response to viral infections over a period of time, will lead to an environment conducive for the penetrance of the genetic effect of *IFNL* variants in asthma. Since the Polish cohort had >60% women asthma patients and was enriched with older women (Table 1) we saw strong association in the older women's sub-group (Fig. 1 and Table 3). We hypothesize that presence of IFN-λ4-generating ΔG allele may be beneficial for an elderly female asthma patient by protecting the airways from increased inflammation associated with virus-induced asthma exacerbations. Although, no direct association of the allele with history of asthma exacerbations or hospitalizations was revealed (data not shown), it should be noted that the Polish asthma cohort represented a group of difficult-to-control asthmatics (only 22% had well-controlled disease according to GINA criteria; Table 1). Along these lines, association of the allele with the presence of acute airway reversibility in response to inhaled beta2- agonists may indirectly reflect less severity of the disease and/or lower airway remodeling in those patients that carry an IFN-λ4-generating ΔG allele. It remains to be established if the effect of the functional IFN-λ4-generating variant on asthma control and severity is also valid in older men. We did see a protective effect of the minor allele rs12979860 (which tags the IFN-λ4-generating allele) against asthma in older men/all men after meta-analysis of data (Figs 2 and 3). However, the significance of association is low given the large sample size tested.

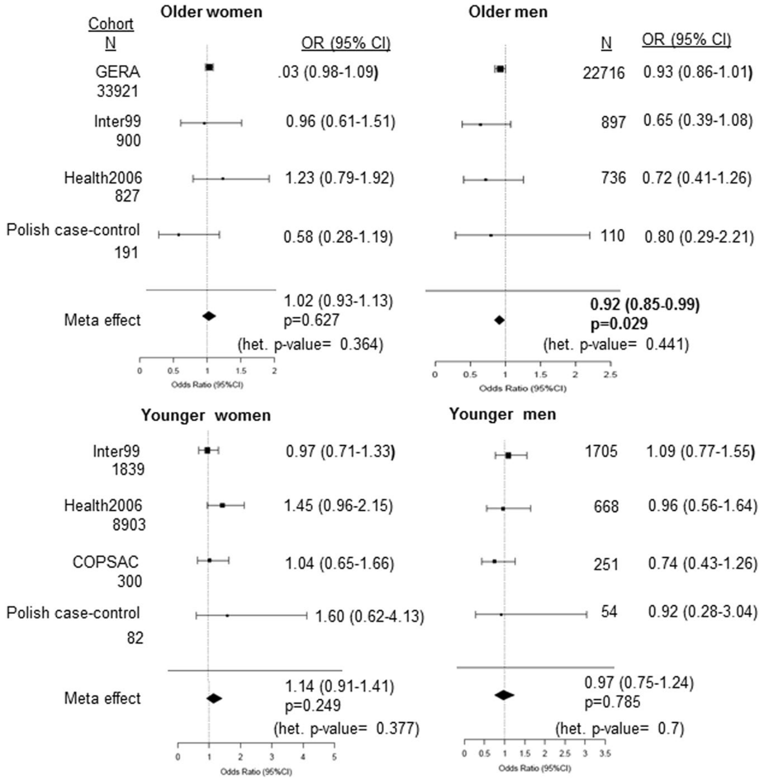


Figure 2. Forest plots showing association of *IFNL* polymorphisms and asthma in sub-groups based on age and gender in different cohorts. The polymorphisms tested are similar to Fig. 1. A dominant model of inheritance for the minor allele (Ex. TT/ ΔG + ΔG/ ΔG vs TT/TT for rs368234815) was used in all cohorts to obtain Odds Ratios (OR) shown as a forest plot. p-value < 0.05 was considered as significant and is in bold. Meta-analysis was performed using a random-effects model. For stratification based on age 50 yr was used as a cut-off mark. Het.-heterogeneity.

GERA cohort	Allergic Rhinitis				
SNP	Model tested	Group/sub-group	N, total; n, cases; n, controls	OR (95% CI)	p-value
rs12979860 (C/T)	Dominant (CT + TT vs CC)	All	56637; 13936; 42701	1.01 (0.97, 1.05)	0.62
		Men	22716; 4859; 17857	1 (0.94, 1.07)	0.96
		Women	33921; 9077; 24844	1.02 (0.97, 1.07)	0.51
rs8099917 (C/T)	Dominant (TG + GG vs TT)	All	Same as above	1 (0.96, 1.04)	0.91
		Men		1 (0.94, 1.07)	0.95
		Women		1 (0.95, 1.05)	0.93

Table 6. Association of *IFNL* locus SNPs with allergic rhinitis in the GERA cohort. GERA cohort consisted of predominantly older participants (average age 63 yr, range 18 to over 100 yr).

Further, it remains to be verified if rs368234815 is the functional variant behind this association in atopy/asthma or other *IFNL* SNPs may also contribute to the phenotype by altering the levels of IFN-λ3 expression similar to recent observations in hepatic inflammation and fibrosis[25]. Future studies aimed at understanding the functional role of IFN-λ4 in regulating inflammation of the airways are required to understand the mechanism behind the association that we identified in the Polish study. Alternately, this finding could be a false positive

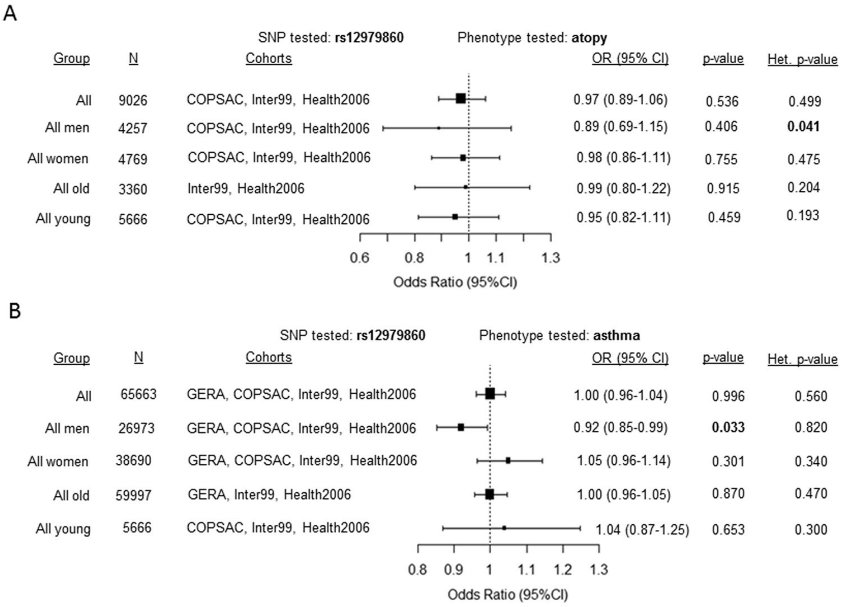


Figure 3. Forest plots showing the summary estimates obtained from all the participants or from different groups based on age and gender after meta-analysis of data from different cohorts for (**A**) atopy and (**B**) asthma. The association was tested with atopy or asthma and rs12979860 using a dominant model for the minor allele (CT + TT vs CC) in all cohorts except for atopy in the COPSAC cohort where an additive model (for minor allele T) was used. The cohorts included in each of the meta-analysis and the effect heterogeneity between different cohorts is shown. p-value < 0.05 was considered as significant and is in bold. Meta-analysis was performed using a random-effects model, while the results were similar using the fixed-effects model too. For stratification based on age 50 yr was used as a cut-off mark. Het.-heterogeneity.

Variable	Genotype		p-value
	TT/TT	TT/ΔG + ΔG/ΔG	
Age (years)	66 +/− 8,2	68,3 ± 9,9	0.156
age at asthma diagnosis (years)	47,3 ± 15,2	50,5 ± 18	0.253
asthma duration (years)	18,8 ± 14,4	18,6 ± 17,6	0.954
ACT sum (points)	17 ± 5,2	17,5 ± 5,7	0.574
MRC (points)	2,7 ± 0,9	2,6 ± 1	0.471
FeNO (ppb)	32,4 ± 27	24,7 ± 15,1	0.098
BMI	28,6 ± 5,3	27,5 ± 4,1	0.199
FEV1% pred.	76,4 ± 25	70,7 ± 25,4	0.189
FVC% pred.	89,8 ± 21,2	85,9 ± 25,4	0.345
FEV1%/FVC%	68,5 ± 14,5	67 ± 12,5	0.507

Table 7. Clinical features of asthma in older women according to rs368234815 genotypes in the Polish cohort. Values are presented as arithmetic means ± SD.

result, which also has to be confirmed by doing a well-designed replication study first in another Polish cohort and later in other populations. Nevertheless, our comprehensive analysis with sufficiently large sample sizes has established that an important candidate gene locus consisting of the immunomodulatory type III IFNs does not associate with atopy and AR in the general population (Fig. 3A and Table 6). The association could be more subtle and with specific endotypes related to inflammation of the airways in asthma. The protection to asthma seen in men/older men (Figs 2 and 3) is nominal and further studies can aim at validation of this result in more specific endotypes based on severity of asthma. A fully functional IFN-λ4 may be associated with protecting the airways from inflammation in certain endotype(s) of the disease, specifically in older asthmatics. While further studies are needed to understand this association, whether these endotypes are a result of virus-induced stimuli, also remains to be examined. A previous report documented a strong positive association of allergy[26], interestingly more pronounced in females than males, with the minor allele of rs12979860 (which is in strong LD with rs368234815[6]) in a pediatric cohort. Our results from the Polish cohort, in contrast, show that the IFN-λ4-generating ΔG minor

	BDR	**iCS**	**oCS**
TT/ΔG + ΔG/ΔG (n/N, %)	41/64, 64.1	61/83, 73.5	2/83, 2.4
TT/TT (n/N, %)	20/49, 40.8	54/61, 88.5	8/61, 13.1
OR crude	**2.583**	**0.359**	**0.163**
95% CI	**1.203–5.55**	**0.142–0.907**	**0.033–0.8**
p	**0.01**	**0.03**	**0.02**
OR adjusted*	**4.651**	**0.112**	0.146
95% CI	**1.414–15.15**	**0.02–0.619**	0.013–1.594
p	**0.01**	**0.01**	0.10
OR adjusted**	2.262	**0.248**	0.227
95% CI	0.975–5.235	**0.081–0.751**	0.041–1.237
p	0.05	**0.01**	0.08

Table 8. Association of the functional IFN-λ4-generating polymorphism rs368234815 with clinical features of asthma in women > 50 yr age sub-group in the Polish cohort. Dominant model of inheritance for the minor allele (TT/ΔG + ΔG/ΔG vs TT/TT) was tested; No significant results were seen under the recessive model. *OR adjusted for age, age at asthma diagnosis, FeNO, BMI, FEV1% and atopy; **OR adjusted for atopy. $p < 0.05$ was considered significant and is shown in bold.

allele may protect older women from atopy rather than contributing to it. The reasons for this paradox remain to be investigated but may likely involve complex epistatic effects mediated by other innate or adaptive immunity genes and age-dependent changes in Th1/Th2 balance during the transition from infancy to adulthood. Interestingly, age and gender are known to interact and influence the association of rs12979860 with another inflammatory condition of Th2-origin, fibrosis, in chronic HCV infections[27].

Material and Methods

Polish study.

Both the controls (N = 111) and asthmatics (N = 326) belonged to same ethnicity (local Polish population) and geography (residents of central Poland). Asthma control was assessed according to GINA 2016 (global initiative for asthma) guidelines and atopy was defined as presence of a positive skin response (weal diameter >3 mm) to at least one of a panel of common inhalant allergens applied as a skin-prick test (SPT). Evaluation also included a questionnaire, FeNO (fractional exhaled nitric oxide) measurement, spirometry and reversibility test with 400 µg salbutamol MDI performed in 280 patients. Genomic DNA was isolated from EDTA-treated whole blood samples using the Qiagen blood genomic DNA mini kit. A competitive allele-specific polymerase chain reaction (PCR) (KASP, LGC Genomics, UK) was used to genotype the functional IFN-λ4-generating SNP rs368234815; the assay was carried out in a StepOnePlus real-time PCR machine (Applied Biosystems, UK). The study was approved by the local Bioethics Committee (document no. RNN/121/12/KE) and all study subjects provided an informed written consent. All methods were carried out in accordance with relevant guidelines and regulations stipulated by the Medical University of Lodz and/or other relevant authorities under it. Further, the university granted approval for all the experimental protocols performed in this study.

Since the dominant and recessive models of inheritance have been reported previously for the *IFNL* SNP association with various diseases[5], we used both these models to test for association with various phenotypes in our study. Statistical analyses of data included log-linear (for testing interactions between different variables) and logistic regression (for multivariate analysis); goodness-of-fit was tested using Pearson's chi-square test or two-tailed Fischer's exact test. All statistical analyses were performed using Statistica 12.5 PL; p-value of < 0.05 was considered statistically significant unless specified.

Inter99 and Health2006 study.

The Health2006 Study took place from 2006 to 2008 and included a random sample of 7,931 Danish (Danish nationality and born in Denmark) men and women aged 18–69 years invited to participate in a health examination. The Inter99 Study is a randomised controlled trial (CT00289237, ClinicalTrials.gov) aiming to investigate the effects of a lifestyle intervention on cardiovascular disease (N = 61,301). The details of these two study cohorts on genotyping and data collection on atopy and asthma are described elsewhere[19, 20]. The Health2006 Study and the Inter99 Study were approved by the Ethics Committee of Copenhagen County and the Danish Data Protection Agency. All participants gave their informed consent, and all methods were carried out in accordance with relevant guidelines and regulations. The Health2006 and Inter99 datasets generated during and/or analysed during the current study are not publicly available due to ethical and legal reasons since public availability may compromise participant privacy, and this would not comply with Danish legislation. Requests for data should be addressed to Professor Allan Linneberg. Access to data will be provided in accordance with the Danish Data Protection Agency.

GERA study.

GERA cohort data was obtained through dbGaP under accession phs000674.v1.p1. The Resource for Genetic Epidemiology Research on Aging (GERA) Cohort was created by a RC2 "Grand Opportunity" grant that was awarded to the Kaiser Permanente Research Program on Genes, Environment, and Health (RPGEH) and the UCSF Institute for Human Genetics (AG036607). The RC2 project enabled genome-wide SNP genotyping (GWAS) to be conducted on a cohort of over 100,000 adults who are members of the Kaiser Permanente Medical Care Plan, Northern California Region (KPNC), and participating in its RPGEH. The resulting GERA cohort is composed of 42% of males, 58% of females, and ranges in age from 18 to over 100 years old with an average age of 63 years at the time of the RPGEH survey (2007). A subset of 62,281 subjects from European ancestry was quality controlled (QCed) and analyzed. A 3-step QC protocol was applied using PLINK and included 2 stages of SNP removal and an intermediate stage of sample exclusion. The exclusion criteria for genetic markers consisted on: proportion of missingness ≥0.05, HWE p-value ≤ 1 × 10–20 for all the cohort, and MAF <0.001. This protocol for genetic markers was performed twice, before and after sample exclusion. For the individuals, we considered the following exclusion criteria: gender discordance, subject relatedness (pairs with PI-HAT ≥ 0.125 from which we removed the individual with the highest proportion of missingness), variant call rates ≥0.02 and population structure showing more than 4 standard deviations within the distribution of the study population according to the first seven principal components. After the QC analysis, 56,637 subjects remained for genotype imputation and association testing.

We performed a two-stage imputation procedure, which consisted in pre-phasing the genotypes into whole chromosome haplotypes followed by imputation itself. The pre-phasing was performed using the SHAPEIT2[28] tool, IMPUTE2[29] for genotype imputation and the SNPTEST (https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html#introduction) for association testing. In this work we used 1000 G Phase 3 haplotypes (October, 2014) as a reference panel to infer ungenotyped variants. After genotype imputation, variants with info score < 0.7 and MAF < 0.001 were removed. Association testing with SNPTEST tool was performed using an additive and dominant logistic regression model adjusting by the 7 derived principal components, age distributed in 14 groups and sex. Moreover, variants with HWE p.value $\leq 1 \times 10^{-6}$ for controls were removed. The diagnostic criteria for allergic rhinitis were based on the following ICD9 codes: 477, 477.0, 477.1, 477.2, 477.8, 477.9 (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/GetPdf.cgi?id=phd004308).

COPSAC study. The Copenhagen Prospective Studies on Asthma in Childhood (COPSAC2000) is a clinical study comprising 411 children with high risk of asthma born to asthmatic mothers. Information on Doctor diagnosed asthma was available for parents of COPSAC2000 study which was used in the current analyses. The COPSAC study is described in detail elsewhere[21]. Genotyping of parents was performed on the Illumina Infinium II HumanHap550 BeadChip and has been described previously[22]. All participants gave their informed consent. The Ethics Committee for Copenhagen and the Danish Data Protection Agency approved this study.

Meta-analysis. R programing using the function 'rma' from 'metafor' R package (R Core Team, 2015) was used. Pooled data were analysed by using a random-effects model (using DerSimonian-Laird's method). The random-effects model was chosen even when effects were homogenous across cohorts since one cohort (GERA) had overwhelmingly large sample size compared to other cohorts and also the cohorts were not uniform in their characteristics (Ex. the Polish cohort was a hospital-based case-control study while the other cohorts were population-based). The significance of the pooled OR was determined by the z-test. Heterogeneity between studies was assessed using the Chi-squared based Cochran's Q-test.

References

1. Wack, A., Terczyńska-Dyla, E. & Hartmann, R. Guarding the frontiers: the biology of type III interferons. *Nature Immunology* **16**, 802–809 (2015).
2. Ge, D. *et al.* Genetic variation in IL28B predicts hepatitis C treatment-induced viral clearance. *Nature* **461**, 399–401 (2009).
3. Suppiah, V. *et al.* IL28B is associated with response to chronic hepatitis C interferon-alpha and ribavirin therapy. *Nature Genetics* **41**, 1100–1104 (2009).
4. Tanaka, Y. *et al.* Genome-wide association of IL28B with response to pegylated interferon-alpha and ribavirin therapy for chronic hepatitis C. *Nature Genetics* **41**, 1105–1109 (2009).
5. Chinnaswamy, S. Gene-disease association with human *IFNL* locus polymorphisms extends beyond hepatitis C virus infections. *Genes &. Immunity* **17**, 265–275 (2016).
6. Prokunina-Olsson, L. *et al.* A variant upstream of IFNL3 (IL28B) creating a new interferon gene IFNL4 is associated with impaired clearance of hepatitis C virus. *Nature Genetics* **45**, 164–171 (2013).
7. Key, F. M. *et al.* Selection on a variant associated with improved viral clearance drives local, adaptive pseudogenization of interferon lambda 4 (IFNL4). *PLoS Genet* **10**, e1004681 (2014).
8. O'Brien, T. R., Prokunina-Olsson, L. & Donnelly, R. P. IFN-λ4: the paradoxical new member of the interferon lambda family. *Journal of Interferon & Cytokine Research* **34**, 829–838 (2014).
9. Hamming, O. J. *et al.* Interferon lambda 4 signals via the IFNλ receptor to regulate antiviral activity against HCV and coronaviruses. *EMBO Journal* **32**, 3055–3065 (2013).
10. Jordan, W. J. *et al.* Human interferon lambda-1 (IFN-lambda1/IL-29) modulates the Th1/Th2 response. *Genes & Immunity* **8**, 254–261 (2007).
11. Srinivas, S. *et al.* Interferon-L1 (interleukin-29) preferentially down-regulates interleukin-13 over T helper type 2 cytokine responses. *Immunology* **125**, 492–502 (2008).
12. Dai, J., Megjugorac, J., Gallagher, G. E., Yu, R. Y. L. & Gallagher, G. IFN-L1 (IL29) inhibits GATA3 expression and suppresses Th2 responses in human naïve and memory T cells. *Blood* **113**, 5829–5838 (2009).
13. Artis, D. & Spits, H. The biology of innate lymphoid cells. *Nature* **517**, 293–301 (2015).
14. Wark, P. A. *et al.* Asthmatic bronchial epithelial cells have a deficient innate immune response to infection with rhinovirus. *Journal of Experimental Medicine* **201**, 937–947 (2005).
15. Contoli, M. *et al.* Role of deficient type III interferon-lambda production in asthma exacerbations. *Nature Medicine* **12**, 1023–1026 (2006).
16. Van Den Berge, M., Heijink, H. I., Van Oosterhout, A. J. M. & Postma, D. S. The role of female sex hormones in the development and severity of allergic and non-allergic asthma. *Clinical & Experimental Allergy* **39**, 1477–81 (2009).
17. Kaczmarek, M. The timing of natural menopause in Poland and associated factors. *Maturitas* **57**, 139–153 (2007).
18. Banda, Y. *et al.* Characterizing Race/Ethnicity and Genetic Ancestry for 100,000 Subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort. *Genetics* **200**, 1285–1295 (2015).
19. Jorgensen, T. *et al.* Effect of screening and lifestyle counselling on incidence of ischaemic heart disease in general population: Inter99 randomised trial. *British Medical Journal* **348**, g3617 (2014).

20. Thuesen, B. H. *et al*. Cohort Profile: the Health2006 cohort, research centre for prevention and health. *Internationa Journal of Epidemiology* **43**, 568–575 (2014).
21. Bisgaard, H. The Copenhagen Prospective Study on Asthma in Childhood (COPSAC): design, rationale, and baseline data from a longitudinal birth cohort study. *Annals of Allergy Asthma & Immnology* **93**, 381–389 (2004).
22. Okbay, A. *et al*. Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* **533**, 539–542 (2016).
23. Bhushan, A., Ghosh, S., Bhattacharjee, S., Chinnaswamy, S. Confounding by SNP rs117648444 (P70S) affects the association of IFNL locus variants with response to IFN-RBV therapy in patients with chronic genotype 3 HCV infection. *Journal of Interferon & Cytokine Research* **37**, 369-382 (2017).
24. Fahy, J. V. Eosinophilic and Neutrophilic Inflammation in Asthma. Proceedings of the American Thorasic Society **6**, 256–259 (2009).
25. Eslam, M. *et al*. IFN-λ3, not IFN-λ4, likely mediates IFNL3-IFNL4 haplotype-dependent hepatic inflammation and fibrosis. *Nature Genetics*, 10.1038/ng.3836. [Epub ahead of print] (2017)
26. Gaudieri, S. *et al*. Genetic variations in IL28B and allergic disease in children. *PLoS One* **7**, e30607 (2012).
27. Eslam, M. *et al*. Interferon-λ rs12979860 genotype and liver fibrosis in viral and non-viral chronic liver disease. Nature Communications 6:6422 (2005).
28. Delaneau, O. Marchin,i J., Zagury, J.F. A linear complexity phasing method for thousands of genomes. *Nature Methods* **9**, 179–181 (2012).
29. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature Genetics* **44**, 955–959 (2012).

Acknowledgements

Polish study: We thank Bartosz Grancow for help with statistical analysis. We thank Drs. Marcin Kurowski and Anna Lewandowska-Polak for helpful discussions. SC was supported as a visiting scientist by Healthy Ageing Research Centre, Medical University of Lodz. This study was supported by Polish National Science Centre grant no. 2013/09/B/NZ6/00746 . The authors (SC, AW, MP, JM and MLK) have been partially supported by The Healthy Ageing Research Centre Project (REGPOT-2012-2013-1, 7FP). *Inter99 and Health2006 study*: TS was supported by a grant from the Lundbeck Foundation (Grant number R165-2013-15410), the Harboe Foundation (Grant number 16152), the A.P. Møller Foundation for the Advancement of Medical Science (Grant number 15-363), Aase and Einar Danielsen's Foundation (Grant number 10-001490), and the Weimann's grant. The Novo Nordisk Foundation Center for Basic Metabolic Research is an independent Research Center at the University of Copenhagen partially funded by an unrestricted donation from the Novo Nordisk Foundation (www.metabol.ku.dk). GERA study: This work has been sponsored by the grant SEV-2011-00067 of Severo Ochoa Program, awarded by the Spanish Government. This work was supported by an EFSD/Lilly research fellowship. Josep M. Mercader was supported by Sara Borrell Fellowship from the Instituto Carlos III. Sílvia Bonàs was FI-DGR Fellowship from FI-DGR 2013 from Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR, Generalitat de Catalunya). We acknowledge PRACE for awarding us access to MareNostrum supercomputer, based in Spain at Barcelona. The technical support group from the Barcelona Supercomputing Center is gratefully acknowledged. COPSAC2000 study: We greatly acknowledge the private and public research funding allocated to COPSAC and listed on www.copsac.com, with special thanks to The Lundbeck Foundation (Grant nr. R16-A1694); Ministry of Health (Grant nr. 903516); Danish Council for Strategic Research (Grant nr.: 0603-00280B); The Danish Council for Independent Research and The Capital Region Research Foundation as core supporters. We gratefully express our gratitude to the participants of the COPSAC2000 cohort study for all their support and commitment. We also acknowledge and appreciate the unique efforts of the COPSAC research team. The authors thank Anand Bhushan, NIBMG for help with meta-analysis.

Author Contributions

Polish study: S.C. and M.L.K. conceived the study; A.W., M.P., J.M. and M.L.K. recruited patients and collected patient data and samples; S.C. performed genotyping; S.C. and A.W. performed statistical analysis; S.C., A.W. and M.L.K. wrote the paper; *Replication studies*: T.S., N.G., and A.L. were involved in the data collection, genotyping or analyses of the Inter99 and Health2006 studies. J.M.M., M.G.M. and D.T. performed the analyses of GERA cohort. T.S.A., H.B., and K.B. were involved in the data collection, genotyping or analyses of the COPSAC2000 study. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication. All authors reviewed the manuscript and approved the final version.

Additional Information

Supplementary information accompanies this paper at doi:10.1038/s41598-017-10467-y

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Appendix 4. Multitrait genome association analysis identifies new susceptibility genes for human anthropometric variation in the GCAT cohort.

Galván-Femenía I, Obón-Santacana M, Piñeyro D, Guindo-Martinez M, Duran X, Carreras A, Pluvinet R, Velasco J, Ramos L, Aussó S, Mercader JM, Puig L, Perucho M, Torrents D, Moreno V, Sumoy L, de Cid R. *J Med Genet*. 2018 Nov;55(11):765-778.

Contribution:

- Genotype imputation of the GCAT cohort using 1000G phase 3, UK10K, GoNL and HRC as reference panels.

OPEN ACCESS

ORIGINAL ARTICLE

Multitrait genome association analysis identifies new susceptibility genes for human anthropometric variation in the GCAT cohort

Iván Galván-Femenía,[1] Mireia Obón-Santacana,[1,2] David Piñeyro,[3] Marta Guindo-Martinez,[4] Xavier Duran,[1] Anna Carreras,[1] Raquel Pluvinet,[3] Juan Velasco,[1] Laia Ramos,[3] Susanna Aussó,[3] J M Mercader,[5,6] Lluis Puig,[7] Manuel Perucho,[8] David Torrents,[4,9] Victor Moreno,[2,10] Lauro Sumoy,[3] Rafael de Cid[1]

► Additional material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/jmedgenet-2018-105437).

For numbered affiliations see end of article.

Correspondence to
Dr Rafael de Cid, GCAT lab Group, Program of Predictive and Personalized Medicine of Cancer (PMPPC), Germans Trias i Pujol Research Institute (IGTP), Crta. de Can Ruti, Badalona 08916, Spain; rdecid@igtp.cat

Received 24 April 2018
Revised 19 July 2018
Accepted 21 July 2018
Published Online First 30 August 2018

Check for updates

To cite: Galván-Femenía I, Obón-Santacana M, Piñeyro D, *et al*. *J Med Genet* 2018;**55**:765–778.

ABSTRACT

Background Heritability estimates have revealed an important contribution of SNP variants for most common traits; however, SNP analysis by single-trait genome-wide association studies (GWAS) has failed to uncover their impact. In this study, we applied a multitrait GWAS approach to discover additional factor of the missing heritability of human anthropometric variation.

Methods We analysed 205 traits, including diseases identified at baseline in the GCAT cohort (Genomes For Life- Cohort study of the Genomes of Catalonia) (n=4988), a Mediterranean adult population-based cohort study from the south of Europe. We estimated SNP heritability contribution and single-trait GWAS for all traits from 15 million SNP variants. Then, we applied a multitrait-related approach to study genome-wide association to anthropometric measures in a two-stage meta-analysis with the UK Biobank cohort (n=336 107).

Results Heritability estimates (eg, skin colour, alcohol consumption, smoking habit, body mass index, educational level or height) revealed an important contribution of SNP variants, ranging from 18% to 77%. Single-trait analysis identified 1785 SNPs with genome-wide significance threshold. From these, several previously reported single-trait hits were confirmed in our sample with *LINC01432* ($p=1.9\times10^{-9}$) variants associated with male baldness, *LDLR* variants with hyperlipidaemia (ICD-9:272) ($p=9.4\times10^{-10}$) and variants in *IRF4* ($p=2.8\times10^{-57}$), *SLC45A2* ($p=2.2\times10^{-130}$), *HERC2* ($p=2.8\times10^{-176}$), *OCA2* ($p=2.4\times10^{-121}$) and *MC1R* ($p=7.7\times10^{-22}$) associated with hair, eye and skin colour, freckling, tanning capacity and sun burning sensitivity and the Fitzpatrick phototype score, all highly correlated cross-phenotypes. Multitrait meta-analysis of anthropometric variation validated 27 loci in a two-stage meta-analysis with a large British ancestry cohort, six of which are newly reported here (p value threshold $<5\times10^{-9}$) at *ZRANB2-AS2*, *PIK3R1*, *EPHA7*, *MAD1L1*, *CACUL1* and *MAP3K9*.

Conclusion Considering multiple-related genetic phenotypes improve associated genome signal detection. These results indicate the potential value of data-driven multivariate phenotyping for genetic studies in large population-based cohorts to contribute to knowledge of complex traits.

INTRODUCTION

Common disorders cause 85% of deaths in the European Union (EU).[1] The increasing incidence and prevalence of cancer, cardiovascular diseases, chronic respiratory diseases, diabetes and mental illness represent a challenge that leads to extra costs for the healthcare system. Moreover, as European population is getting older, this scenario will be heightened in the next few years. Like complex traits, many common diseases are complex inherited conditions with genetic and environmental determinants. Advancing in their understanding requires the use of multifaceted and long-term prospective approaches. Cohort analyses provide an exceptional tool for dissecting the architecture of complex diseases by contributing knowledge for evidence-based prevention, as exemplified by the Framingham Heart Study[2] or the European Prospective Investigation into Cancer and Nutrition cohort study.[3]

In the last decades, high performance DNA genotyping technology has fuelled genomic research in large cohorts, having been the most promising line in research on the aetiology of most common diseases. Genome-wide association studies (GWAS) have provided valuable information for many single conditions.[4] Despite the perception of the limitations of the GWAS analyses, efforts combining massive data deriving from whole-genome sequencing at population scale with novel conceptual and methodological analysis frameworks have been set forth to explore the last frontier of the missing heritability issue,[5] driving the field of genomic research on complex diseases to a new age.[6]Pritchard and colleagues recently proposed the breakthrough idea of the *omnigenic* character of genetic architecture of diseases and complex traits.[7] They suggested that beyond a handful of driver genes (ie, core genes) directly connected to an illness, the missing heritability could be accounted for by multiple genes (ie, peripheral genes) not clustered in functional pathways, but dispersed along the genome, explaining the pleiotropy frequently seen in most complex traits. Core genes have been already outlined by the GWAS approach, but most of the possible contributing genes have been disregarded based on methodological issues such as p value or lower minor allele frequency (MAF). Pathway disturbances have also been a landmark in the search for genetic associations,[8] but not always appear to the root of the mechanism of inheritance of complex diseases, at least for peripheral genes.[7] With this challenging vision, a multitrait genome association analysis of the whole phenome[9] becomes a more appropriate way to detect peripheral gene variation effects and new network disturbances affecting core genes. Multitrait analysis approaches are developed for research of genetically complex conditions using raw or summary-level data statistics from GWAS in order to explain the largest possible amount of the covariation between SNPs and traits.[10–15]

The contribution of total genetic variation, known as heritability (broad-sense heritability, h^2), is estimated now from genome-wide studies in large cohorts directly from SNP data (known as h^{2}SNP). However, even if most disease conditions have a strong genetic basis, it is well known that our capacity to find genetic effects depends on the overall genetic contribution of the trait. Overall estimations differed depending on the ancestry, sample ascertainment, gender and age of the population under study. Recently, data from the UK Biobank determined genetic contributions with a phenome-based approach[16] and identified a shared familial environment as a significant important factor besides genetic *heritability* values in 12 common diseases analysed.[17]

In this study, we present new data on phenotype-wide estimation of the heritability of 205 complex traits (including diseases) and new insights into the genetics of anthropometric traits in a Mediterranean Caucasian population using a two-stage meta-analysis approach with multiple-related phenotypes (MRPs).

MATERIALS AND METHODS

Population

The methodology of the GCAT study has been previously described.[18] Briefly, the subjects of the present study are part of the GCAT project, a prospective study that includes a cohort of a total of 19 267 participants recruited from the general population of Catalonia, a western Mediterranean region in the Northeast of Spain. Healthy general population volunteers between 40 and 65 years with the sole condition of being users of the Spanish National Health Service were invited to be part of the study mostly through the Blood and Tissue Bank, a public agency of the Catalan Department of Health. All eligible participants signed an informed consent agreement form and answered a comprehensive epidemiological questionnaire. Anthropometric measures and blood samples were also collected at baseline by trained healthcare personnel. The GCAT study was approved by the local ethics committee (Germans Trias University Hospital) in 2013 and started on 2014.

Study participants

This study analyses the GCATcore data, a subset of 5459 participants (3066 women) with genotype data belonging to the interim GCATdataset, August 2017 (see the URLs section). GCATcore participants were randomly selected from whole cohort based on overall demographic distribution (ie, gender, age, residence). In this study, in order to increase the robustness of heritability estimates, only Caucasian participants with a Spanish origin (based on principal component analysis (PCA) analysis, see later in this section) and with available genetic data were finally included: 4988 GCAT participants (2777 women). All samples passed genotyping quality control (QC) (see later in this section).

Phenome

Baseline variables were obtained from a self-reported epidemiological questionnaire and included biological traits, medical diagnoses, drug use, lifestyle habits and sociodemographic and socioeconomic variables.[18] Description of GCAT variables dataset is available at GCAT (see the URLs section). To keep as many as possible of the genotyped samples in the study, we imputed anthropometric missing values (<1%) from the overall distribution values using statistical approaches. Missing values (<1%) for biological and anthropometric measures (height, weight, waist and hip circumference, systolic and diastolic blood pressure and heart rate) were imputed by stratifying the whole GCAT cohort by gender and age and using multiple imputation by the fully conditional specification method, implemented in the R mice package.[19] For GWAS analysis, we retained all variables with at least five observations (n=205). For heritability estimates, only variables with at least 500 individuals per class were retained (n=96) for robustness. The description of the traits and measures included in this study is summarised in online supplementary table S1.

Genotyping, relatedness and population structure

Genotyping of the 5459 GCAT participants (GCATcore) was done using the Infinium Expanded Multi-Ethnic Genotyping Array (MEGA^{Ex}) (ILLUMINA, San Diego, California, USA). A customised cluster file was produced from the entire sample dataset and used for joint calling. We applied PCA to detect any hidden substructure and the method of moments for the estimation of identity by descent probabilities to exclude cases with cryptic relatedness. The extensive QC protocol used for cluster analysis and call filtering is accessible at GCAT (see the URLs section) and presented as supplementary material (online supplementary file S1). Briefly, GCAT participants were excluded from the analysis for different reasons, including poor call rate <0.94 (n=61), gender mismatch (n=19), duplicates (n=8), family relatedness up to second degree (n=88) and excess or loss of heterozygosity (n=52). Non-Caucasian individuals detected as outliers in the PCA plot of the European populations from the 1000 Genomes Project (n=96) and born outside of Spain (n=147) were also excluded from the study. After QC and filtering, 4988 GCAT participants and 1 652 023 genetic variants were included. Genotyping was performed at the PMPPC-IGTP High Content Genomics and Bioinformatics Unit.

Multipanel imputation

For imputation analysis, 665 592 SNPs were included (40%). Sexual and mitochondrial chromosomes were discarded as well as autosomal chromosome variants with MAF <0.01 and AT-CG sites. We followed a two-stage imputation procedure, which consists of prephasing the genotypes into whole chromosome haplotypes followed by imputation itself.[20] The prephasing was performed using SHAPEIT2, and genotype imputation was performed with IMPUTE2. As reference panels for genotype imputation, we used the 1000 Genomes Project phase 3,[21] the Genome of the Netherlands,[22] UK10K[23] and the Haplotype Reference Consortium.[24] All variants with IMPUTE2 *info* <0.7 were removed. After imputing the genotypes using each reference panel separately, we combined the results selecting the variants with a higher *info* score when they were present in more than one reference panel. The SNP dosage from IMPUTE2 was transformed to binary PLINK format by using the '-hard-call-threshold 0.1' flag from PLINK. The final core set had approximately 15 million variants with MAF>0.001 and 9.5 million variants with MAF>0.01. Imputation was performed at the Barcelona Supercomputing Center.

Heritability

Trait SNP heritability (h^2_{SNP}) was estimated from SNP/INDEL array/imputed data with the GREML-LDMS method implemented in the GCTA software.[25] Since this method is relatively unbiased regarding MAF and linkage disequilibrium (LD) parameters, we considered autosomal variants with MAF>0.001 (15 060 719 SNPs) to avoid under/overestimation of heritability due to the relatively small sample analysed in the core study. Cryptic relatedness of distant relatives was also considered, and individuals whose relatedness in the genetic relationship matrix was >0.025 were discarded (n=4717). Population stratification was controlled in the linear mixed model using the first 20 principal components of the PCA derived from population genetic structure analysis of the GCAT. Gender and age were also included as covariates in the model. The h^2_{SNP}CIs were calculated by using FIESTA.[26]

Single-trait genome-wide association analysis

We performed independent GWAs analyses for 205 selected traits (61 continuous and 144 binary). A total of 9 499 600 SNPs with MAF>0.01 were considered for this purpose. Linear regression models for continuous traits were assessed with PLINK.[27] For binary traits, given the unbalanced design of most of the traits considered, we used a scoring test with saddle point approximation included in the *SPAtest* R package.[28] This approach compensates a slight loss of power with the inclusion of uncommon and rare conditions, without affecting robustness. All the models included the first 20 PCAs, age and gender as covariates. A PCA-mixed analysis was applied to approximate the number of independent traits[29] (online supplementary figure S1). Based on these figures, Bonferroni correction for multiple traits was defined at $p<5\times10^{-10}$accounting for 100 independent traits explaining 80% of the phenome variability.

Multitrait meta-analysis for correlated traits

We applied a multitrait approach for the analysis of anthropometric traits (weight, height, body mass index (BMI) and waist and hip circumference) in a two-stage association study using individuals of British ancestry from the UK Biobank cohort (N=336 107).[30] Waist-to-hip ratio was excluded from this analysis due to its unavailability from the UK Biobank resource. UK Biobank summary-level statistics was calculated using linear regression models with the inferred gender and the first 10 PCAs as covariates, similarly to the model applied on GCAT data (see the URLs section). All SNPs with suggestive association $p<1x10^{-5}$ for any trait were retained from the GCAT GWAS analysis. Then, only SNPs intersecting with the UK Biobank resource were used for multitrait meta-analysis association testing in both samples, and $p<5x10^{-9}$was considered significant. The multitrait association testing was based on the distribution of the sum of squares of the z scores which is insensitive to the direction of the scores.[31] Briefly, let $Z = (z_1, z_2, \dots, z_k)$ be the z scores for a given SNP for *k* phenotypes. The sum of squares of the z scores, $S_{sq} = \sum_{i=1}^{k} z_i^2$, can be approximated by the χ^2 distribution (χ^2). Let Σ be the covariance matrix of the genome-wide z scores from the phenotypes under analysis. And let c_i be the eigenvalues of Σ, the distribution of S_{sq} is well approximated by $a\chi_d^2 + b$, where a, b and d depend on c_i. Then, we calculated the p value as: $p\left(\chi_d^2 > \left(S_{sq} - b\right) / a\right)$. To estimate the covariance matrix of the correlated traits, we selected independent SNPs (LD pruning in PLINK "--indep-pairwise 50 5 0.2") and filtered out SNPs with |z scores|>1.96 to avoid possible bias in the estimation of Σ because of the difference in sample size and association p values in the GCAT-UK Biobank. A summary flow chart of the methods applied in this study is shown in figure 1.

Polygenic risk score

Genetic architecture was analysed by the polygenic risk score (PRS). Polygenic risk score software (PRSice)[32] was used to predict the genetic variability of the identified loci for a given trait. PRSice plots the percentage of variance explained for a trait by using SNPs with different p value thresholds (P_T) (online supplementary figure S2). Here, we considered $P_T=0.05$.

URLs

GCAT study, http://genomesforlife.com;

National Human Genome Research Institute GWAS Catalog, http://www.genome.gov/gwastudies/ (gwas_catalog_v1.0-associations_e91_r2018-02-06);

1000 Genomes Project http://www.internationalgenome.org/ (phase 3, v5a.20130502);

Genome of Netherland http://www.nlgenome.nl/ (Release 5.4);

UK10K https://www.uk10k.org/ (Release 2012-06-02, updated on 15 Feb 2016) ;

Haplotype Reference Consortium http://www.haplotype-reference-consortium.org/(Release 1.1);

UKBiobank GWAS Results; https://sites.google.com/broadinstitute.org/ukbbgwasresults/home?authuser=0, (Manifest20170915);

GTExportal, https://www.gtexportal.org/home/. (last data accession, Release V.7, dbGaP accession phs000424. v7. P2);

RESULTS

Heritability estimates

SNP heritability estimation (h^2_{SNP}) in the GCATcore study showed values ranging from 77% to 18%, with height being the trait showing the strongest SNP contribution. The h^2_{SNP} SE for most traits was high (near 10%), with wide CIs, as expected by sample size. However, robustness of the analysis is supported by similar values to those reported elsewhere (see wide summary in Genome-wide complex trait analysis, Wikipedia. *The Free Encyclopedia*, 2018). Statistically significant h^2_{SNP} estimations for continuous and binary traits (cases >500) are shown in table 1. In particular, values for height: h^2_{SNP}=0.77, 95% CI0.56 to 0.94 and BMI: h^2_{SNP}=0.38, 95% CI0.20 to 0.59 were identical to the maxima achieved in other European populations, using comparable genomic approaches. Besides the anthropometric traits, the Fitzpatrick's phototype score, a numerical classification schema for human skin colour to measure the response of different types of skin to ultraviolet light, had a high genetic consistency in our sample (h^2_{SNP}=0.63, 95% CI 0.4 to 0.8), and concordantly all related categories (eye colour, hair colour, freckling and skin sensitivity) showed high heritability (h^2_{SNP}>0.3). It is worth noting that skin colour had the lowest value (h^2_{SNP}=0.18, 95% CI 0.02 to 0.38), which is in concordance with the blurred genetic architecture of skin colour.[33] Interestingly, other non-biological traits showed relatively high values in our study. Educational level showed the third highest heritability value (h^2_{SNP}=0.54, 95% CI 0.35 to 0.74). Lower estimates have been observed in other Caucasian populations, but this could be explained by the fact that this estimate is for educational level as a categorical variable and not as binary (higher/lower).

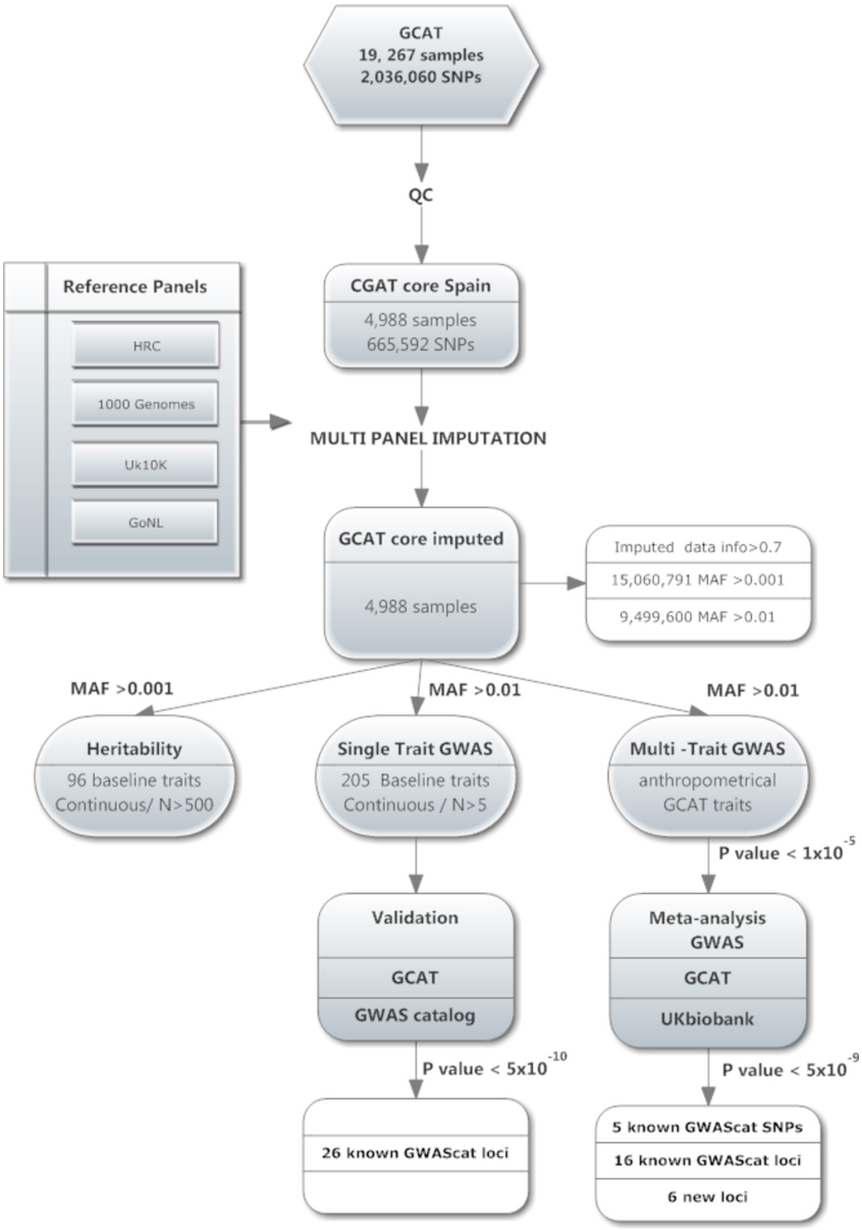


Figure 1 Flow chart of the methods and criteria used in this study. GCAT, Genomes For Life- Cohort Study of the Genomes of Catalonia; GWAS, genome-wide association studies; MAF, minor allele frequency; QC, quality control.

Self-perceived health was similar to h^2_{SNP} from recent data from a larger UK Biobank study,[16] with values around 20% (h^2_{SNP}=0.22, 95% CI 0.04 to 0.43).

Phenome analysis

GWAS identified 6820 associations in 1785 SNPs with genome-wide significance threshold $p<5\times10^{-8}$ and 29 343 associations with a suggestive association $p<1\times10^{-5}$. Here, we report 26 genome-wide association hits identified in our study which confirm results previously identified in other European ancestry samples (GWAS Catalog database (release V.1.0, e90, 27 September 2017)).[4] In table 2, we show the SNP associations with the minimum p value for each locus, the remaining SNPs are shown in online Supplementary file 5. Five genes associated with pigmentary traits were identified in the analysis with highly significant SNP associations: *SLC45A2* (rs16891982, β=−0.546, SE=0.021, $p=2.2\times10^{-130}$), *IRF4* (rs12203592, β=1.915, SE=0.118, $p=2.8\times10^{-57}$), *HERC2* (rs1667394, β=−0.608, SE=0.02, $p=2.8\times10^{-176}$), *OCA2* (rs11855019, β=−0.548, SE=0.022, $p=2.4\times10^{-121}$) and *MC1R* (rs1805007, β=3.615, SE=0.326, $p=7.7\times10^{-22}$) (online supplementary figure S3). These genes are involved in the regulation and distribution of melanin pigmentation or enzymes involved in melanogenesis itself within the melanocyte cells present in the skin, hair and eyes in Caucasian populations.[33–35] Pigmentary traits (mainly

Table 1 h^2_{SNP} of the analysed traits with $h^2_{SNP}>0$, SE <0.12, p<0.05 and $n_b>500$

Questionnaire—section	Description	Trait name	h2SNP	SE	95% CI	P values	n	n_b	NA
Anthropometric and blood pressure	Height	height_c	0.77	0.11	0.56 to 0.94	2×10^{-12}	4717	–	0
Other habits	Phototype score	phototype_ score	0.63	0.11	0.4 to 0.8	3.7×10^{-9}	4664	–	56
Demographic and socioeconomic	Educational level	education	0.54	0.10	0.35 to 0.74	1.1×10^{-8}	4698	–	19
Other habits	Fitzpatrick phototype score	phototype_score categorical	0.52	0.11	0.29 to 0.74	6.0×10^{-7}	4664	–	56
Other habits	Eye colour phototype score	eye_color_phototype_score	0.48	0.11	0.27 to 0.68	7.1×10^{-6}	4716	–	1
Other habits	Freckling (has freckles)	freckling_binary	0.47	0.11	0.26 to 0.68	8.1×10^{-6}	4713	590	4
Other habits	Hair colour phototype score	hair_color_phototype_score	0.46	0.11	0.26 to 0.68	6.7×10^{-6}	4709	–	9
Other habits	Eye colour	eye_color	0.44	0.11	0.24 to 0.65	3.4×10^{-5}	4716	–	1
Other habits	Hair colour	hair_color	0.41	0.11	0.21 to 0.63	4.1×10^{-5}	4709	–	9
Other habits	Hair colour (black)	hair_color_black	0.39	0.11	0.22 to 0.59	0.00018	4709	952	9
Anthropometric and blood pressure	BMI (kg/m^2)	bmi	0.38	0.11	0.2 to 0.59	0.00013	4717	–	0
Anthropometric and blood pressure	Weight	weight_c	0.37	0.11	0.19 to 0.57	0.00016	4717	–	0
Tobacco consumption	Smoking habit	smoking_habit	0.36	0.11	0.19 to 0.58	0.00037	4717	–	0
Tobacco consumption	Smoking packs per day	smoking_packs	0.35	0.11	0.17 to 0.55	0.00082	4717	–	0
Other habits	Skin sensitivity to sun	skin_sensitivity_to_sun	0.33	0.11	0.15 to 0.52	0.0011	4714	–	3
Anthropometric and blood pressure	Hip circumference	hip_c	0.31	0.11	0.15 to 0.51	0.0011	4717	–	0
Occupation	Working status (active)	working_status_active	0.31	0.11	0.13 to 0.54	0.0014	4696	1570	23
Other habits	Skin sensitivity to sun phototype score	skin_sensitivity_to_sun_ phototype_score	0.30	0.11	0.12 to 0.51	0.0022	4714	–	3
Anthropometric and blood pressure	BMI obesity	bmi_who_obesity	0.29	0.11	0.12 to 0.51	0.0031	4717	1388	0
Physical activity	Sleep duration	sleep_duration	0.29	0.11	0.1 to 0.49	0.0033	4645	–	79
Other habits	Freckling	freckling	0.28	0.11	0.11 to 0.5	0.0043	4713	–	4
Medical history	Mental health (MHI-5)	sadness	0.26	0.11	0.09 to 0.48	0.0053	4717	504	0
Occupation	Working last year	working_last_year	0.26	0.11	0.09 to 0.47	0.0065	4685	1190	32
Other habits	Freckling phototype score	freckling_phototype_score	0.26	0.11	0.09 to 0.46	0.0076	4713	–	4
Other habits	Eye colour (dark)	eye_color_dark	0.25	0.11	0.07 to 0.47	0.012	4716	1192	1
Other habits	Hair colour (brown)	hair_color_brown	0.24	0.11	0.07 to 0.45	0.012	4709	1229	9
Anthropometric and blood pressure	Waist circumference	waist_c	0.24	0.11	0.06 to 0.44	0.01	4717	–	0
Anthropometric and blood pressure	Waist-to-hip ratio WHO categories	whr_who	0.23	0.11	0.05 to 0.45	0.016	4717	–	0
Medical history	Self-perceived health	self_perceived_health	0.22	0.11	0.04 to 0.43	0.024	4715	–	2
Tobacco consumption	Smoking status (ever smoked)	smoking_status	0.21	0.11	0.02 to 0.42	0.026	4522	1828	204
Alcohol consumption	Current alcohol consumption	alcohol_actual	0.20	0.11	0.03 to 0.4	0.031	4713	3670	4
Diet	Predimed score	predimed_score	0.20	0.11	0.03 to 0.41	0.031	4627	–	95
Women's health	No of female children	offspring_female	0.19	0.11	0.02 to 0.4	0.028	4717	–	0
Anthropometric and blood pressure	Waist-to-hip ratio obesity	whr_who_obesity	0.19	0.11	0.04 to 0.39	0.036	4717	1512	0
Women's health	No of male children	offspring_male	0.19	0.11	0.02 to 0.41	0.036	4717	–	0
Medical history	Self-perceived health (bad)	self_perceived_health_binary	0.18	0.11	0.02 to 0.4	0.047	4715	629	2
Medical history	Certain adverse effects not classified elsewhere	icd9_code3_995	0.18	0.11	0.01 to 0.37	0.042	4717	775	0
Demographic and socioeconomic	Civil status (ever been married)	civil_status_ever_married	0.18	0.11	0.01 to 0.38	0.04	4703	523	15
Other habits	Skin colour phototype score	skin_color_phototype_score	0.18	0.11	0.02 to 0.38	0.047	4714	–	3

BMI, body mass index; h^2_{SNP}, SNP heritability estimation; MHI-5, Mental Health Inventory 5-item questionnaire; n_b, sample size of the minor category in binary traits; _c for Weight_c, height_c, hip_c and waist_c mean calculated-imputed variable.

the red hair colour phenotype) are related to the defensive capacity of the skin in response to sun exposure (UV-induced skin tanning or sun burning), and it has been established as a risk factor for sun-induced cancers (both melanoma and non-melanocytic skin cancers).[36] Other GWAS hits from the phenome-wide analysis validated previously reported findings in *CCDC141-LOC105373766* (rs79146658, β=2.359, SE=0.374, $p=3.4\times10^{-10}$), *SMARCA4-LDLR* (rs10412048, β=−0.5, SE=0.079, $p=3.2\times10^{-10}$; rs6511720, β=−0.493, SE=0.08, $p=9.4\times10^{-10}$) and *LINC01432* (rs1160312, β=0.193, SE=0.03, $p=1.9\times10^{-9}$) loci, related with cardiovascular risk (heart_rate), hyperlipidaemia (icd9_code3_272) and male pattern baldness (hair_loss_40), respectively (see table 2).

Table 2 Twenty-six genome-wide associated loci with GCAT traits and reported in the GWAS Catalog

Gene	SNP	Chr:position*	Imputed	Info	GWAS Catalog traits†	Studies	Published year	GCAT trait	β	SE	P values
CCDC141	rs151041685	2:179725237	Yes	0.998	Resting heart rate	1	2016	heart_rate_c	2.06	0.361	$1.2{\times}10^{-8}$**
CCDC141, LOC105373766	rs79146658	2:179786068	Yes	0.971	Diastolic blood pressure	1	2017	heart_rate_c	2.359	0.3749	$3.4{\times}10^{-10}$
SLC45A2	rs16891982	5:33951693	Yes	0.985	Hair colour, eye colour, black versus non-black hair colour, skin sensitivity to sun, squamous cell carcinoma, melanoma, monobrow	6	2010, 2015, 2016, 2017	skin_color	−0.546	0.021	$2.2{\times}10^{-130}$
DUSP22, IRF4	rs7773324	6:382559	Yes	0.986	Crohn's disease, inflammatory bowel disease	1	2015	freckling_phototype_score	0.281	0.045	$6.5{\times}10^{-10}$**
IRF4	rs12203592	6:396321	No	–	Black versus blond hair colour, black versus red hair colour, hair colour, eye colour, freckling, progressive supranuclear palsy, non-melanoma skin cancer, tanning, sunburns, facial pigmentation, skin colour saturation, cutaneous squamous cell carcinoma, squamous cell carcinoma, basal cell carcinoma	9	2008, 2010, 2011, 2013, 2015, 2016	hair_color_phototype_score	1.915	0.118	$2.8{\times}10^{-57}$
IRF4, LOC105374875	rs62389424	6:422631	Yes	0.882	Blond versus non-blond hair colour, brown versus non-brown hair colour, light versus dark hair colour, lung cancer in ever smokers	2	2015, 2017	freckling_phototype_score	−0.926	0.073	$1.6{\times}10^{-35}$
LOC105374875	rs12210050	6:475489	No	–	Tanning, basal cell carcinoma, schizophrenia	4	2009, 2011, 2012, 2016	hair_phototype_score	1.025	0.123	$1.7{\times}10^{-16}$
RNU2-47P, TYRP1	rs1408799	9:12672097	No	–	Blue versus green eyes, eye colour	2	2008, 2013	eye_phototype_score	0.453	0.071	$2.2{\times}10^{-10}$
BNC2-LOC105375983		9:16884586	Yes	0.991	Cutaneous squamous cell carcinoma, basal cell carcinoma	2	2		−0.089	0.016	$3.4{\times}10^{-8}$**
LOC107984363, TYR	rs1126809	11:89017961	Yes	0.993	Tanning, sunburns, cutaneous squamous cell carcinoma, squamous cell carcinoma, basal cell carcinoma	4	2013, 2016	skin_color	−1.672	0.282	$3.5{\times}10^{-9}$**
LOC105370627	rs12896399	14:92773663	No	–	Blond versus brown hair colour, blue versus green eyes, black versus blond hair colour, hair colour, eye colour	4	2007, 2008, 2010, 2013	phototype_score	0.093	0.016	$1.9{\times}2.5^{-8}$**
OCA2	rs11855019	15:28335820	No	–	Black versus blond hair colour, black versus red hair colour	1	2008	hair_color	−0.548	0.022	$2.4{\times}10^{-121}$
HERC2	rs1667394	15:28530182	No	–	Blond versus brown hair colour, blue versus green eyes, blue versus brown eyes, eye colour	2	2007, 2012	eye_color	−0.608	0.02	$2.8{\times}10^{-176}$
SPG7, RPL13	rs67689854	16:89625227	Yes	0.902	Stromal cell-derived factor 1 alpha levels	1	2016	eye_color	2.284	0.278	$7.9{\times}10^{-11}$
SPATA33	rs35063026	16:89736157	Yes	0.987	Facial pigmentation, squamous cell carcinoma	2	2015, 2016	hair_color_red	3.112	0.309	$5.4{\times}10^{-17}$
CDK10	rs258322	16:89755903	No	–	Black versus red hair colour, melanoma	5	2008, 2009, 2011, 2014, 2017	hair_color_red	2.431	0.267	$1.2{\times}10^{-13}$
FANCA	rs12931267	16:89818732	Yes	0.989	Hair colour, freckling, skin sensitivity to sun	2	2015, 2017	hair_color_red	3.218	0.311	$3.9{\times}10^{-18}$
MC1R	rs1805007	16:89986117	No	–	Freckles, blond versus brown hair colour, red versus non-red hair colour, skin sensitivity to sun, basal cell carcinoma, tanning, hair colour, sunburns, non-melanoma skin cancer, perceived skin darkness, cutaneous squamous cell, melanoma	7	2007, 2011, 2013, 2015, 2016, 2017	hair_color_red	3.615	0.326	$7.7{\times}10^{-22}$
DEF8	rs146972365	16:90022693	Yes	0.974	Red versus non-red hair colour, light versus dark hair colour, brown versus non-brown hair colour	1	2015	hair_color	0.442	0.052	$3{\times}10^{-17}$
AFG3L1P	rs8063160	16:90054709	Yes	0.988	Brown versus non-brown hair colour, light versus dark hair colour, red versus non-red hair colour	1	2015	hair_color_red	2.577	0.277	$8.9{\times}10^{-15}$
TSPAN10	rs9747347	17:79606820	Yes	0.982	Myopia	1	2016	hair_color_phototype_score	−0.526	0.087	$1.8{\times}10^{-9}$**
HMGN1P31-CDH20		18:58840518	Yes	0.953	Deep ovarian and/or rectovaginal disease with dense	1	2017	handedness	0.045	0.008	$3.9{\times}10^{-8}$**

Continued

Table 2 Continued

Gene	SNP	Chr:position*	Imputed	Info	GWAS Catalog traits†	Studies	Published year	GCAT trait	β	SE	P values
SMARCA4, LDLR		19:11193949	Yes	0.999	Cholesterol, total	1	2017	icd9_code3_272	−0.501	0.079	3.2×10^{-10}
LDLR	rs6511720	19:11202306	No	–	LDL cholesterol, carotid intima media thickness, cardiovascular disease risk factors, lipoprotein-associated phospholipase A2 activity and mass, cholesterol, total, metabolite levels, lipid metabolism phenotypes, Abdominal aortic aneurysm	12	2008, 2009, 2010, 2011, 2012, 2013	icd9_code3_272	−0.493	0.081	9.4×10^{-10}**
RPL41P1-LINC01432		20:22000281	No	–	Male-pattern baldness	1	2016	hair_loss_40	0.19	0.032	6.2×10^{-9}**
LINC01432	rs1160312	20:22050503	No	–	Male-pattern baldness	1	2008	hair_loss_40	0.193	0.032	1.9×10^{-9}**

*Chr:position based on hg19.

†GWAS Catalog traits based on GWAS Catalog database (release V.1.0, e90, 27 September 2017).

5×10^{-8} threshold for univariate GWAS and 5×10^{-10} threshold accounting for multiple phenotypes.

GWAS, genome-wide association studies; LDL, low-density lipoprotein.

_c for heart-rate_c, means calculated-imputed variable.

Multitrait meta-analysis of anthropometric traits

Anthropometric traits had a high heritability in our sample (height=77%, BMI=38%, weight=37%, hip circumference=31%and waist circumference=24%), and all were highly correlated (online supplementary figure S1). In the first stage, from single-trait GWAS, we retained 606 SNPs with suggestive association ($p<1\times10^{-5}$) (see figure 2). None of them reached the genome-wide significance threshold. In the second stage, we analysed those 476 SNPs that intersected with the UK Biobank cohort dataset. Multitrait meta-analysis identified 111 SNPs in 27 independent loci with $p<5\times10^{-9}$ (online Supplementary file 7). Table 3 shows the SNPs with the highest significance for each independent *loci* and the univariate summary statistics of the anthropometric traits in both cohorts.

We estimated the covariance matrix (Σ) for each dataset (GCAT, UK Biobank and GCAT +UK Biobank). Then, as described in the Materials and methods section, we selected those independent SNPs with |z scores|<1.96, resulting in 765 646, 630890 and 535860 being considered for the Σ estimation. Eigenvalues of Σ showed d=1.36, 1.4 and 2.72 values. Covariance matrices were similar in both GCAT and UK Biobank (online supplementary tables S4 and S5). One degree of freedom (GCAT and UK Biobank) and three (GCAT +UK Biobank) of the ^{2} distribution were considered for multitrait analysis. We identified 27 independent multitrait loci associated in GCAT and UK Biobank (table 3). We intersected these SNPs with the GWAS Catalog, and we found that 5 SNPs had previously been reported in multiple GWAS, 16 loci were reported considering a ±250000base pair window from the identified SNP and 6 were new loci involving the following genes/SNPs: *MAD1L1* (rs62444886, $p=2.3\times10^{-15}$), *PIK3R1* (rs12657050, $p=2.8\times10^{-13}$; rs695166, $p=8.4\times10^{-15}$), *ZRANB2-AS2* (rs11205277, $p=1.4\times10^{-9}$), *EPHA7* (rs143547391, $p=6.5\times10^{-10}$), *CACUL1* (rs12414412, $p=4\times10^{-13}$) and *MAP3K9* (rs7151024, $p=5.7\times10^{-10}$). Regarding *DPYD*, *DPYD-IT1* (rs140281723), *GABRG3-AS1* and *GABRG3* (rs184405367) genes/SNPs, we did not replicate association in UK Biobank samples (UKmulti p=0.035 and 1, respectively). The risk allele, frequency and functional annotation using the Variant Effect Predictor tool[37] of identified variants are shown in online Supplementary file 9.

Polygenic risk score

The skin phototype association analysis identified five loci accounting for a high predictive value (PRS of 15.6%) suggesting few main genes (oligogenic architecture) contributing to the phenotype (online supplementary figure S2). However, for anthropometric traits, 27 loci were identified in our cohort but with a lower PRS (2.3%) suggesting a polygenic architecture with multiple genes and a high environmental impact. The newly identified loci only increased PRS slightly over the corresponding single-trait analysis (2.2% to 2.5%, 2.3% to 3.3%, 2.2% to 3.5%, 2.5% to 3.7% and 1.5% to 2.6% for height, weight, BMI and hip and waist circumference, respectively) pointing towards the multitrait approach as an effective screening strategy to identify new biomarkers.

DISCUSSION

Dissecting the architecture of common diseases should incorporate multitrait approaches to understand the phenome and its genetic aetiology, including pleiotropy and the co-occurrence of multiple morbidities, correlated traits and the diseasome as targets for genomic analysis.[38] In this study, we used the GCAT study, a South-European Mediterranean population prospective

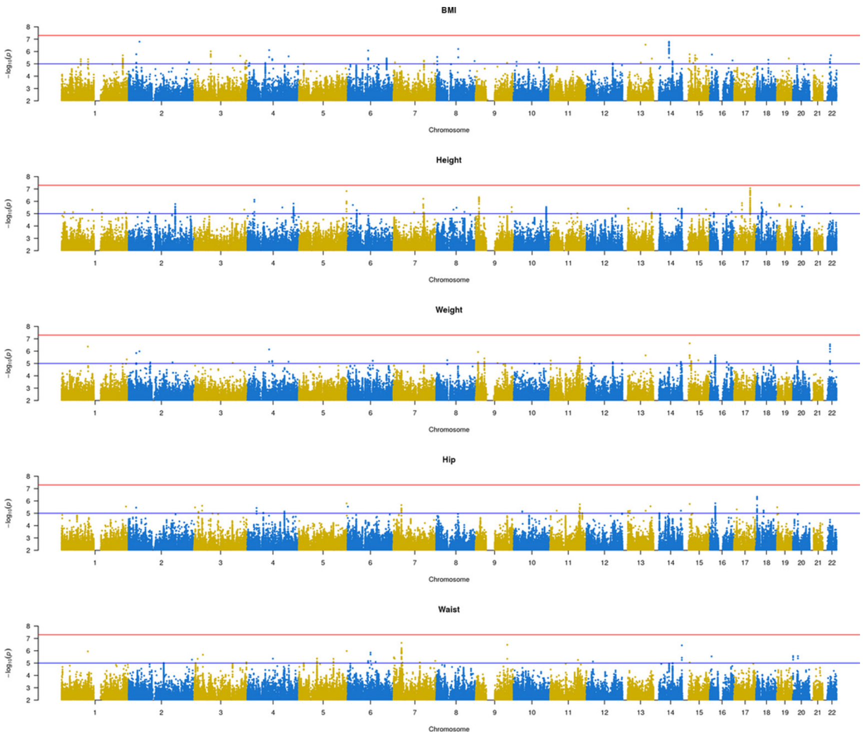


Figure 2 Manhattan plot of the anthropometric traits (BMI, height, weight and hip and waist circumference) from the GCAT. BMI, body mass index.

cohort to analyse the phenotypic variation attributable to genotype variability for 205 selected human traits (including diseases as well as biological, anthropometric and social features). Our results show that by considering genetic covariance matrices for interrelated traits, we increased the number of detected *loci* from six new *loci* for anthropometric traits, pointing to multitrait analysis as an effective strategy to gain statistical power to identify genetic association.

The relative importance of genetic and non-genetic factors varies across populations. Moreover, this is not constant in a population and changes with age.[16] Here, we have reported heritability estimates on an adult population based on SNP data. In the present study, h^2_{SNP} values move in a wide range from 18% to 77%, being anthropometric traits (height) and skin colour-related traits (Fitzpatrick's phototype score) the traits with the highest genetic determination. In our cohort, heritability of anthropometric traits, such as height and BMI, was likely estimated as a maximum, with negligible missed heritability when comparing with other reported estimates in similar populations[39] and in the same way being the observed genetic variance only a small part of their complete variance (around 3%). In the case of skin colour-related traits, the portion of the explained variance was larger, in accordance with a less complex polygenic nature of this trait, and fewer genes baring stronger predictive value (*IRF4*, *HERC2*, *OCA2*, *MC1R* and *SLC45A2*) (PRS=15.6%). The variants identified in these loci associated with skin colour-related traits are functional and have been reported elsewhere in several studies. These differences in heritability and prediction values indicate a different genomic architecture, suggesting an exposure variation, the exposome,[3] as a main actor for many polygenic traits. Higher estimates in self-perceived health heritability, and probably some other reported traits such as 'smoking_habits', 'smoking_packs', or 'sadness' (item from the Mental-Health Inventory 5-item questionnaire), reflect a pleiotropic effect[40] with multiple associated loci. In this sense, a recent meta-analysis on subjective well-being revealed new loci accounting for a polygenic model of well-being status.[41]

Single-trait GWAS analysis identified a number of genetic variants associated with skin colour-related traits (online supplementary figure S3) and other complex traits (heart rate, hyperlipidaemia or male pattern baldness); whereas failed to identify specific variants associated with any single anthropometric trait (at the $p<5\times10^{-8}$ threshold cut-off). However, we should observe that gender differences were not considered in this analysis even though it has been shown that genetic effects have a gender bias.[42] Applying multitrait analyses of anthropometric traits, we identified 27 loci, six of which had not been reported previously; *CALCUL1*, *ZRANB2-AS2*, *MAD1L1*, *EPHA7*, *PIK3R1* and *MAP3K9*. Owing to LD and the occurrence of all identified variants in non-coding regions (see online Supplementary file 9), we cannot be certain about the genes involved. Two out of six of the identified associated variants, in *CALCUL1* and *MAP3K9*, are putative expression quantitative trait loci (eQTL) (see the URLs section). Three of the variants (*ZRANB2-AS2*chr1:71702511, *EPHA7*chr6:94075927 and *MAP3K9*chr14:71268446) are specific of the GCAT sample ($p<5\times10^{-9}$) (online Supplementary files 10,11, S,12) probably due to genetic background differences between populations (ie, LD patterns) or as an expression of a particular genetic contribution of the Mediterranean populations to these polygenic traits. Identified variants implicate genes with diverse functions, involved in several pathways and processes. Some of them are involved in growth, developmental or metabolic processes.

Table 3 Loci associated with anthropometric traits in GCAT and UK Biobank cohorts

				Single-trait analysis															Multitrait analysis	
				Weight (kg)			Height (cm)			BMI (kg/m2)			Waist circumference (cm)			Hip circumference (cm)				
Loci*	Chr:position†	SNP	Cohort	β	SE	P values	β	SE	P values	β	SE	P values	β	SE	P values	β	SE	P values	P values	GWAS Catalog‡
SF3B4, SV2A	1:149892872	rs11205277	GCAT	−0.064	0.093	0.49	0.55	0.12	9.1×10^{-6}	0.29	0.27	0.27	0.2	0.24	0.4	0.2	0.19	0.29	0.00092	Reported SNP
			UK Biobank	−0.0017	0.0024	0.47	0.034	0.0017	1.1×10^{-85}	0.017	0.0021	4.3×10^{-16}	0.009	0.0022	3.3×10^{-5}	0.019	0.0024	5.1×10^{-15}	3.8×10^{-53}	
			GCAT-UK Biobank																1.5×10^{-53}	
ZRANB2-AS2	1:71702511	rs115213730	GCAT	1.7	0.37	4.2×10^{-6}	−0.77	0.49	0.12	3.8	1.1	0.00033	3.8	0.95	6.5×10^{-5}	2.8	0.76	0.0002	1.6×10^{-8}	New loci
			UK Biobank	0.021	0.0068	0.0026	0.0044	0.0049	0.37	0.019	0.006	0.0016	0.017	0.0061	0.0063	0.016	0.0068	0.017	0.00015	
			GCAT-UK Biobank																1.4×10^{-9}	
DPYD, DPYD-IT1	1:97884058	rs140281723	GCAT	1.9	0.45	2×10^{-5}	1.1	0.6	0.071	6.5	1.3	4.3×10^{-7}	5.6	1.2	1.1×10^{-6}	3.8	0.93	4.7×10^{-5}	8.6×10^{-11}	No association
			UK Biobank	0.011	0.0086	0.21	−0.015	0.0062	0.012	0.0004	0.0076	0.96	0.013	0.0077	0.1	0.0023	0.0086	0.79	0.035	
			GCAT-UK Biobank																1.9×10^{-9}	
PRELID1, RAB24, MXD3	5:176735612	rs111251222	GCAT	0.06	0.12	0.62	0.78	0.16	1×10^{-6}	0.84	0.34	0.014	0.33	0.31	0.28	0.13	0.25	0.59	0.0001	Reported loci
			UK Biobank	−0.0084	0.0028	0.0024	0.034	0.002	2.7×10^{-67}	0.012	0.0024	1.2×10^{-6}	0.0084	0.0025	0.00063	0.0007	0.0028	0.8	3.5×10^{-35}	
			GCAT-UK Biobank																2.2×10^{-36}	
LMAN2, AC146507.1	5:176772736	rs4976686	GCAT	0.049	0.1	0.63	0.71	0.13	1.5×10^{-7}	0.79	0.29	0.0071	0.29	0.26	0.26	0.14	0.21	0.5	2.8×10^{-5}	Reported loci
			UK Biobank	−0.0056	0.0026	0.034	0.028	0.0019	9.8×10^{-49}	0.01	0.0023	8.6×10^{-6}	0.0042	0.0023	0.073	0.0021	0.0026	0.42	2.7×10^{-25}	
			GCAT-UK Biobank																4.9×10^{-27}	
PIK3R1	5:67579576	rs12657050	GCAT	−0.29	0.11	0.0083	−0.33	0.15	0.022	−1	0.32	0.0011	−1.3	0.28	8.9×10^{-6}	−0.59	0.23	0.009	1.1×10^{-6}	Unreported locus
			UK Biobank	−0.0043	0.0028	0.13	−0.014	0.002	4.9×10^{-12}	−0.011	0.0025	7.7×10^{-6}	−0.009	0.0025	0.00035	−0.0067	0.0028	0.017	4.1×10^{-10}	
			GCAT-UK Biobank																2.8×10^{-13}	
	5:67604628	rs695166	GCAT	−0.27	0.1	0.011	−0.41	0.14	0.0029	−1.1	0.3	0.00043	−1.2	0.27	7.4×10^{-6}	−0.69	0.21	0.0012	1.2×10^{-7}	
			UK Biobank	−0.0031	0.0027	0.24	−0.015	0.0019	2.3×10^{-14}	−0.01	0.0024	1×10^{-5}	−0.0083	0.0024	0.00051	−0.0054	0.0027	0.044	8.7×10^{-11}	
			GCAT-UK Biobank																8.4×10^{-15}	
GMDS	6:1944345	rs62391629	GCAT	0.54	0.17	0.0017	0.52	0.23	0.023	2	0.49	7.9×10^{-5}	1.4	0.44	0.0022	1.7	0.35	2.9×10^{-6}	4.9×10^{-8}	Reported locus
			UK Biobank	0.013	0.0051	0.0085	0.0085	0.0036	0.02	0.016	0.0045	0.00045	0.014	0.0045	0.0014	0.017	0.0051	0.00068	6.4×10^{-6}	
			GCAT-UK Biobank																2×10^{-10}	
ID4, AL022068.1, –	6:19839415	rs41271299	GCAT	−0.11	0.22	0.62	1.4	0.3	2×10^{-6}	0.96	0.64	0.13	0.081	0.58	0.89	0.35	0.46	0.45	0.00048	Reported loci
			UK Biobank	−0.0032	0.0054	0.55	0.094	0.0039	1.4×10^{-129}	0.049	0.0048	1.8×10^{-24}	0.027	0.0049	1.9×10^{-8}	0.041	0.0054	6.8×10^{-14}	2.7×10^{-77}	
			GCAT-UK Biobank																4.5×10^{-78}	
GRM4, HMGA1	6:34199092	rs2780226	GCAT	0.029	0.15	0.85	0.91	0.2	5.6×10^{-6}	0.89	0.44	0.042	0.14	0.39	0.73	0.34	0.31	0.28	0.00043	Reported SNP
			UK Biobank	0.00064	0.0042	0.88	0.067	0.003	7.4×10^{-109}	0.037	0.0037	7.9×10^{-23}	0.033	0.0038	8.3×10–19	0.02	0.0042	2.7×10^{-6}	1.6×10^{-68}	
			GCAT-UK Biobank																2.7×10^{-69}	
HMGA1, SMIM29, AL354740.1	6:34214322	rs1150781	GCAT	0.029	0.15	0.85	0.9	0.2	9.8×10^{-6}	0.86	0.44	0.049	0.08	0.39	0.84	0.37	0.31	0.24	0.00059	Reported SNP
			UK Biobank	0.0023	0.0042	0.59	0.066	0.003	2.4×10^{-106}	0.037	0.0037	5.7×10^{-24}	0.034	0.0038	7.2×10^{-20}	0.021	0.0042	5.1×10^{-7}	1×10^{-68}	
			GCAT-UK Biobank																2.3×10^{-69}	
EPHA7	6:94075927	rs143547391	GCAT	1.2	0.4	0.0019	1.7	0.53	0.0013	5.2	1.1	6×10^{-6}	3.5	1	0.00063	2.8	0.82	0.00082	3.4×10^{-8}	New locus
			UK Biobank	−0.022	0.0088	0.014	−0.01	0.0063	0.1	−0.025	0.0078	0.0014	−0.027	0.0079	0.00077	−0.027	0.0088	0.0025	3.3×10^{-5}	
			GCAT-UK Biobank																6.5×10^{-10}	
AOC1, KCNH	7:150599205	rs10216051	GCAT	−0.44	0.099	9.9×10^{-6}	0.089	0.13	0.5	−1	0.28	0.00022	−1.1	0.25	2.6×10^{-5}	−0.84	0.2	3.5×10^{-5}	9.5×10^{-9}	R eported loci
			UK Biobank	0.0073	0.0025	0.0039	0.012	0.0018	2.7×10^{-11}	0.012	0.0022	2.5×10^{-8}	0.0082	0.0023	0.00029	0.0087	0.0025	0.00062	4.2×10^{-12}	
			GCAT-UK Biobank																4.8×10^{-17}	
MAD1L1, –	7:2068330	rs62444886	GCAT	−0.76	0.18	3.9×10^{-5}	−0.23	0.25	0.34	−2.3	0.53	1.6×10	−2.2	0.47	4×10^{-6}	−1.6	0.38	3.6×10^{-5}	1.9×10^{-9}	Unreported locus
			UK Biobank	−0.026	0.0052	7.3×10^{-7}	0.0051	0.0037	0.17	−0.019	0.0046	−5 4.7×10^{-5}	−0.021	0.0047	1×10^{-5}	−0.025	0.0052	2.5×10^{-6}	9.9×10^{-10}	
			GCAT-UK Biobank																2.3×10^{-15}	

Continued

Table 3 Continued

				Single-trait analysis																Multitrait analysis	
				Weight (kg)			Height (cm)			BMI (kg/m2)			Waist circumference (cm)			Hip circumference (cm)				
Loci*	Chr:position†	SNP	Cohort	β	SE	P values	β	SE	P values	β	SE	P values	β	SE	P values	β	SE	P values	P values	GWAS Catalog‡
FUBP3	9:133482006	rs11792294	GCAT	−0.075	0.099	0.45	−0.59	0.13	7.1×10^{-6}	−0.69	0.29	0.016	−0.079	0.26	0.76	−0.2	0.21	0.32	0.00029	Reported locus
			UK Biobank	0.0021	0.0025	0.39	−0.02	0.0018	2×10^{-29}	−0.0092	0.0022	2.5×10^{-5}	−0.0024	0.0022	0.27	−0.0032	0.0025	0.2	5.5×10^{-16}	
			GCAT-UK Biobank																6.3×10^{-17}	
CALCUL1	10:120465796	rs12414412	GCAT	0.054	0.17	0.75	1	0.23	5.9×10^{-6}	1.1	0.5	0.029	0.15	0.44	0.73	0.45	0.35	0.2	0.00033	Unreported locus
			UK Biobank	0.022	0.0043	4.3×10^{-7}	0.0074	0.0031	0.017	0.022	0.038	6.4×10^{-9}	0.015	0.0039	6.6×10^{-5}	0.023	0.0043	1×10^{-7}	3.7×10^{-12}	
			GCAT-UK Biobank																4×10^{-13}	
INS-IGF2, IGF2-AS, –	11:2172830	rs7948458	GCAT	−0.37	0.11	0.00044	−0.42	0.14	0.0027	−1.4	0.31	5.8×10^{-6}	−0.91	0.27	0.00091	−0.94	0.22	1.7×10^{-5}	4.5×10^{-9}	Reported loci
			UK Biobank	−0.002	0.0031	0.5	−0.022	0.0022	4.2×10^{-24}	-0.014	0.0027	3.6×10^{-7}	−0.0044	0.0027	0.1	−0.014	0.0031	5.8×10^{-6}	2.3×10^{-16}	
			GCAT-UK Biobank																1.5×10^{-21}	
MAP3K9	14:71268446	rs7151024	GCAT	−0.49	0.11	6.3×10^{-6}	0.34	0.14	0.017	−1	0.31	0.0013	−1.1	0.28	4.4×10^{-5}	−0.93	0.22	3.1×10^{-5}	5.7×10^{-9}	New locus
			UK Biobank	−0.0097	0.0029	0.00073	0.0054	0.0021	0.0084	−0.0051	0.0025	0.042	−0.0065	0.0026	0.012	−0.0058	0.0029	0.044	0.00015	
			GCAT-UK Biobank																5.7×10^{-10}	
GABRG3-AS1, GABRG3	15:27398499	rs184405367	GCAT	1.5	0.32	1.7×10^{-6}	0.67	0.42	0.11	4.7	0.91	2.4×10^{-7}	3.6	0.81	9.1×10^{-6}	3.1	0.65	1.8×10^{-6}	1.3×10^{-11}	No association
			UK Biobank	−0.0027	0.016	0.86	−0.0015	0.011	0.89	−0.0041	0.014	0.77	−0.005	0.014	0.72	0.0024	0.016	0.88	1	
			GCAT-UK Biobank																3.5×10^{-9}	
SEMA6D	15:47923520	rs10220751	GCAT	−0.44	0.093	2×10^{-6}	0.21	0.12	0.086	−1	0.27	0.00015	−0.8	0.24	0.00086	−0.59	0.19	0.0022	7.1×10^{-8}	Reported locus
			UK Biobank	−0.011	0.0024	2.7×10^{-6}	0.0043	0.0018	0.014	-0.0071	0.0022	0.001	−0.0045	0.0022	0.039	−0.011	0.0024	1.1×10^{-5}	1.6×10^{-7}	
			GCAT-UK Biobank																8×10^{-12}	
GPRC5B-GPR139	16:19988852	rs9940317	GCAT	0.43	0.12	0.00033	0.42	0.16	0.0085	1.6	0.35	6.3×10^{-6}	1.4	0.31	1.2×10^{-5}	1.2	0.25	2.7×10^{-6}	3.7×10^{-10}	Reported loci
			UK Biobank	0.012	0.0029	3.8×10^{-5}	0.0068	0.0021	0.00096	0.013	0.0025	1.4×10^{-7}	0.0079	0.0026	0.0022	0.011	0.0029	0.00012	3×10^{-9}	
			GCAT-UK Biobank																1.6×10^{-15}	
GPR139	16:20046115	rs2045457	GCAT	0.38	0.1	0.00016	0.24	0.13	0.069	1.3	0.29	1.4×10^{-5}	1.2	0.26	3.3×10^{-6}	0.93	0.21	8.6×10^{-6}	9×10^{-10}	Reported locus
			UK Biobank	0.013	0.0026	7.6×10^{-7}	0.0057	0.0019	0.0024	0.014	0.0023	2×10^{-9}	0.0068	0.0023	0.0038	0.011	0.0026	5.4×10^{-5}	1.1×10^{-10}	
			GCAT-UK Biobank																1.4×10^{-16}	
ECI1, AC009065.8	16:2296197	rs77407216	GCAT	−0.39	0.13	0.0036	−0.7	0.18	9.6×10^{-5}	−1.7	0.39	6.9×10^{-6}	−0.82	0.35	0.018	−0.97	0.28	0.00053	5.4×10^{-8}	Reported loci
			UK Biobank	−0.0021	0.0035	0.55	−0.016	0.0025	4.4×10^{-10}	−0.01	0.0031	0.0011	−0.0076	0.0031	0.016	−0.0068	0.0031	0.051	3.1×10^{-7}	
			GCAT-UK Biobank																1.2×10^{-11}	
ATAD5, AC130324.2	17:29165934	rs9890032	GCAT	−0.098	0.095	0.3	−0.61	0.13	1.4×10^{-6}	−0.83	0.27	0.0023	−0.49	0.24	0.044	−0.41	0.2	0.035	7×10^{-6}	Reported SNP
			UK Biobank	-6.7×10^{-5}	0.0025	0.98	−0.032	0.0018	1.8×10^{-71}	-0.017	0.0022	1.6×10^{-15}	−0.011	0.0022	2.2×10^{-6}	−0.013	0.0025	3.8×10^{-7}	1.4×10^{-43}	
			GCAT-UK Biobank																8.2×10^{-46}	
TBX2	17:59498052	rs7214743	GCAT	0.084	0.095	0.37	0.66	0.13	2×10^{-7}	0.82	0.27	0.0028	0.43	0.24	0.082	0.47	0.2	0.016	3×10^{-6}	Reported locus
			UK Biobank	−0.0093	0.0026	0.00027	0.034	0.0018	1.3×10^{-78}	0.011	0.0023	3.8×10^{-7}	0.0064	0.0023	0.0053	0.00026	0.0026	0.92	1.9×10^{-40}	
			GCAT-UK Biobank																5.1×10^{-43}	
CABLES1	18:20758310	rs34302357	GCAT	0.22	0.12	0.071	−0.75	0.16	2.9×10^{-6}	−0.13	0.35	0.7	0.21	0.31	0.49	−0.037	0.25	0.88	0.00048	Reported locus
			UK Biobank	−0.0024	0.0031	0.43	−0.042	0.0022	3.3×10^{-80}	−0.024	0.0027	3.8×10^{-19}	−0.018	0.0028	2.7×10^{-10}	−0.017	0.0031	3.7×10^{-8}	2.9×10^{-51}	
			GCAT-UK Biobank																6.4×10^{-52}	
RIOK3, Y RNA	18:21039393	rs9954741	GCAT	0.2	0.1	0.05	−0.6	0.14	8.7×10^{-6}	−0.025	0.29	0.93	0.14	0.26	0.58	−0.032	0.21	0.88	0.00076	Reported loci
			UK Biobank	−0.01	0.0025	7.1×10^{-5}	−0.013	0.0018	1.1×10^{-13}	−0.016	0.0022	2.2×10^{-12}	−0.015	0.0023	1.1×10^{-11}	−0.014	0.0025	2.8×10^{-8}	8.9×10^{-21}	
			GCAT-UK Biobank																2.7×10^{-21}	
ADAMTS10	19:8670147	rs62621197	GCAT	0.48	0.19	0.011	−1.2	0.25	2.3×10^{-6}	0.29	0.55	0.59	0.37	0.49	0.45	0.2	0.39	0.61	0.00016	Reported locus
			UK Biobank	0.014	0.0067	0.036	−0.11	0.0048	5.4×10^{-121}	−0.05	0.0059	2.6×10^{-17}	−0.036	0.006	2.4×10^{-9}	−0.042	0.0067	4×10^{-10}	1.9×10^{-69}	
			GCAT-UK Biobank																1.3×10^{-70}	

Continued

MAP3K9, mitogen-activated protein kinase 9, has been associated to some rare cancers (ie, retroperitoneum carcinoma and retroperitoneum neuroblastoma), and GWAS studies have identified variants associated with reasoning ability.[43] Based on GTEx database (see URL section) we identified *rs7151024* as an eQTL, expressed in subcutaneous adipose tissue ($p=1.4\times10^{-8}$, eQTL effect size (es)$=-0.38$) that may affect fat distribution and anthropometric traits. *ZRANB2-AS2* is a non-coding RNA, and GWAS studies have identified variants in *ZRANB2-AS2* associated with facial morphology,[44] and also with general cognitive function,[45] traits which are genetically correlated with a wide range of physical variables. *EPHA7* belongs to the ephrin receptor subfamily of protein-tyrosine kinase, implicated in mediating developmental events, particularly in the nervous system. *EPHA7* has been implicated in neurodevelopment processes[46] as well being as a tumour suppressor gene in cancer.[47]*CACUL1*, CDK2-associated cullin domain 1, is a cell cycle-dependent kinase binding protein capable of promoting cell progression. In the GWAS Catalog, any of the anthropometric traits analysed here have been associated with variants in *CACUL1* (online Supplementary file 13). However, the associated rs12414412, reported as an eQTL expressed in skeletal muscle ($p=1.4\times10^{-7}$, eQTL es$=-0.31$), may affect body constitution. *CACUL1* suppresses androgen receptor (AR) transcriptional activity, impairing LSD-mediated activation of the AR,[48] whose genetic variation is associated with longitudinal height in young boys.[49]*MAD1L1,* mitotic arrest deficient 1-like protein 1, is a component of the mitotic spindle-assembly checkpoint, and some cancers (prostate and gastric) have been associated to *MAD1L1* dysfunction.[50] Our study identified BMI, weight and hip and waist circumference single-trait association ($p<10^{-5}$) with the intronic variant *rs62444886* in the *MAD1L1* locus, as well as a significant multitrait association in meta-analysis (table 3, online Supplementary file 14). GWAS analysis identified *MAD1L1* as a susceptibility gene for bipolar disorder and schizophrenia, involved in reward system functions in healthy adults,[51] but until now, no other study has identified it as a genetic contributor to weight. The higher prevalence of obesity and related disorders such as diabetes in schizophrenia patients could reflect a possible underlying common genetic contribution. In this sense, we observed also GWAS significant signals in *INS-IGF2* (GCAT-UKmulti $p=1.5\times10^{-21}$), an analogue of the *INS* gene (previously associated with diabetes type I and type II disorders).[52] Additionally, epigenome-wide association studies in adults[53] and children[54] support a role for *MAD1L1* in BMI–methylation association, with differentially methylated CpG patterns in CD4+ and CD8+ T cells between obese and non-obese women. *PIK3R1,* phosphoinositide-3-kinase regulatory subunit 1, plays a role in the metabolic actions of insulin, and a mutation in this gene has been associated with insulin resistance. Moreover, common variants are associated with lower body fat percentage as well as the control of peripheral adipose tissue mobilisation.[55] Genetic variation in the GWAS Catalog is also associated with cartilage thickness[56] and mineral bone density,[57] both related to anthropometric traits. Diseases associated with *PIK3R1* include SHORT syndrome,[58] characterised by individuals with short stature and a restricted intrauterine growth, in addition to multiple anomalies. Our study identified the intronic variant (*rs695166*) associated with waist circumference association in single-trait analysis ($p<10^{-6}$), but not in the UKdataset, which associates with height ($p=2.3\times10^{-14}$). However, analysis of the UKBiobank data supported a similar peak profile overlapping the

Table 3 Continued

				Single-trait analysis															Multitrait analysis	
				Weight (kg)			Height (cm)			BMI (kg/m2)			Waist circumference (cm)			Hip circumference (cm)				
Loci*	Chr:position†	SNP	Cohort	β	SE	P values	β	SE	P values	β	SE	P values	β	SE	P values	β	SE	P values	P values	GWAS Catalog‡
GDF5, GDF5OS	20:34025756	rs143384	GCAT	0.072	0.094	0.45	0.59	0.13	2.7×10^{-6}	0.74	0.27	0.0065	0.31	0.24	0.21	0.55	0.19	0.0052	1.4×10^{-5}	Reported SNP
			UK Biobank	–0.0014	0.0024	0.58	0.064	0.0018	8.8×10^{-292}	0.033	0.0022	1.9×10^{-53}	0.0071	0.0022	0.0013	0.028	0.0024	1.6×10^{-30}	8.3×10^{-168}	
			GCAT-UK Biobank																1.3×10^{-170}	
HORMAN, LIF	22:30610546	rs9608851	GCAT	0.45	0.095	2×10^{-6}	–0.029	0.13	0.022	1	0.27	0.00027	0.82	0.24	0.00086	0.66	0.2	0.00075	3.2×10^{-8}	Reported loci
			UK Biobank	0.005	0.0024	0.038	0.0062	0.0017	0.00037	0.0077	0.0021	0.00032	0.0058	0.0022	0.007	0.0054	0.0024	0.025	1.7×10^{-5}	
			GCAT-UK Biobank																3.3×10^{-10}	

*Loci, a locus was considered as the ±250 000 base pair window flanking the identified SNP.
†Chr:position, coordinates on hg19.
‡GWAS Catalog traits, data from GWAS Catalog database (release V.1.0, e90, 27 September 2017).
Single-trait and multi trait results are presented. Concordant significant results are marked in violet
BMI, body mass index; GWAS, genome-wide association studies.

gene region (see online Supplementary file 12) and multitrait analysis association (GCAT-UK multi $p=8.4\times10^{-15}$) (table 3).

Multiple approaches for multitrait analysis using GWAS data have been successfully applied in the research of genetically complex conditions using raw data or summary-level data statistics. Using raw data, Ferreira and Purcell[11] used a test based on the Wilk's lambda derived from a canonical correlation analysis. Korte *et al*[13] implemented a mixed-model approach accounting for correlation structure and the kinship relatedness matrix. O'Reilly *et al*[14] proposed an inverted regression model for each SNP as the response and all the traits as covariates. Regarding the use of GWAS summary-level data statistics, Cotsapas *et al*[10] developed a statistic for cross-phenotype analysis based on an asymptotic ^{2} distribution derived from p values of the SNP associations. Zhu *et al*[15] implemented CPASSOC that accounts for the genetic correlation structure of the traits and the sample size for each cohort. Kim *et al*[12] proposed an adaptive association test for multiple traits that uses Monte Carlo simulations to approximate its null distribution. Recently, Bayes factor approaches[59] have been proposed for studying multitrait genetic associations. Here, for meta-analysis purposes, we chose the multitrait analysis described by Yang and Wang.[31] This test, based on the ^{2} distribution with 'd' df, depends on the genetic covariance structure of the traits and considers the distribution of the sum square of the z scores which is insensitive to the heterogeneous effect of the SNP. Nevertheless, this approach doesn't allow allele effect estimation. In this sense, maximum likelihood methods have been recently proposed to deal with this limitation[41] by accounting for different measures of the same phenotypic trait with different levels of heritability.

In complex diseases research, MRPs are the common observation in genome-wide association analysis of large cohorts, and over simplification of extreme phenotypes or the use of standardised phenotypes for meta-analysis reduces the power to detect the underlying genetic contribution to complex traits. As an alternative, multitrait analyses help to detect additional loci that are missing by applying a conventional meta-analysis. Our results highlight the potential value of data-driven multivariate phenotyping for genetic studies in large complex cohorts.

Author affiliations
[1]GenomesForLife-GCAT Lab Group, Program of Predictive and Personalized Medicine of Cancer (PMPPC), Germans Trias i Pujol Research Institute (IGTP), Crta. de Can Ruti, Badalona, Catalunya, Spain
[2]Unit of Biomarkers and Susceptibility, Cancer Prevention and Control Program, Catalan Institute of Oncology (ICO), IDIBELL and CIBERESP, Barcelona, Spain
[3]High Content Genomics and Bioinformatics Unit, Program of Predictive and Personalized Medicine of Cancer (PMPPC), Germans Trias i Pujol Research Institute (IGTP), Badalona, Catalunya, Spain
[4]Life Sciences - Computational Genomics, Barcelona Supercomputing Center (BSC-CNS), Joint BSC-CRG-IRB Research Program in Computational Biology, Barcelona, Spain
[5]Programs in Metabolism and Medical & Population Genetics, Broad Institute of Harvard and MIT, Cambridge, Massachusetts, US
[6]Diabetes Unit and Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts, US
[7]Blood Division, Banc de Sang i Teixits, Barcelona, Spain
[8]Cancer Genetics and Epigenetics Group, Program of Predictive and Personalized Medicine of Cancer (PMPPC), Germans Trias i Pujol Research Institute (IGTP), Badalona, Catalunya, Spain
[9]ICREA, Catalan Institution for Research and Advanced Studies, Barcelona, Catalunya, Spain
[10]Department of Clinical Sciences, Faculty of Medicine, University of Barcelona, Barcelona, Spain

Correction notice This article has been corrected since it was published online first. JMM has been added to the authors list and to the 'Contributors' section.

Acknowledgements The authors thank all the GCAT participants and all BST members for generously helping with this research.

Contributors All authors contributed to the feedback of the manuscript and played an important role in implementing the study. IG-F, MP, VM and RdC conceived the study. IG-F and RdC planned the study. LP coordinated the cohort recruitment. AC, JV and XD prepared the samples. MO-S and XD curated the epidemiological data variables. DP, RP, LR, SA and LS conducted the genotyping. IG-F, DP and LS analysed the clustering analysis. IG-F, MG-M, JMM and DT conducted the imputation analysis. IG-F and RdC conducted and supervised the genetic analysis. IG-F, MO-S and RdC wrote the manuscript. RdC submitted and supervised the study.

Funding This work was supported in part by the Spanish Ministerio de Economía y Competitividad (MINECO) project ADE 10/00026, by the Catalan Departament de Salut and by the Departament d'Empresa i Coneixement de la Generalitat de Catalunya, the Agència de Gestió d'Estudis Universitaris i de Recerca (AGAUR) (SGR 1269, SGR 1589 and SGR 647). RdC is the recipient of a Ramon y Cajal grant (RYC-2011-07822). The Project GCAT is coordinated by the Germans Trias i Pujol Research Institute (IGTP), in collaboration with the Catalan Institute of Oncology (ICO), and in partnership with the Blood and Tissue Bank of Catalonia (BST). IGTP is part of the CERCA Programme/Generalitat de Catalunya.

Competing interests None declared.

Patient consent Obtained.

Ethics approval http://www.ceicgermanstrias.cat/.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Eurostat Statistics Explained. Mortality and life expectancy statistics, 2016. http://ec.europa.eu/eurostat/statistics-explained/index.php/Mortality_and_life_expectancy_statistics
2 Dawber TR, Meadors GF, Moore FE. Epidemiological approaches to heart disease: the Framingham Study. *Am J Public Health Nations Health* 1951;41:279–86.
3 Riboli E, Hunt KJ, Slimani N, Ferrari P, Norat T, Fahey M, Charrondière UR, Hémon B, Casagrande C, Vignat J, Overvad K, Tjønneland A, Clavel-Chapelon F, Thiébaut A, Wahrendorf J, Boeing H, Trichopoulos D, Trichopoulou A, Vineis P, Palli D, Bueno-De-Mesquita HB, Peeters PH, Lund E, Engeset D, González CA, Barricarte A, Berglund G, Hallmans G, Day NE, Key TJ, Kaaks R, Saracci R. European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. *Public Health Nutr* 2002;5:1113–24.
4 Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, Parkinson H. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* 2014;42:D1001–6.
5 Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM. Finding the missing heritability of complex diseases. *Nature* 2009;461:747–53.
6 Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J. 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet* 2017;101:5–22.
7 Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* 2017;169:1177–86.
8 Chakravarti A, Turner TN. Revealing rate-limiting steps in complex disease biology: The crucial importance of studying rare, extreme-phenotype families. *Bioessays* 2016;38:578–86.
9 Freimer N, Sabatti C. The human phenome project. *Nat Genet* 2003;34:15–21.
10 Cotsapas C, Voight BF, Rossin E, Lage K, Neale BM, Wallace C, Abecasis GR, Barrett JC, Behrens T, Cho J, De Jager PL, Elder JT, Graham RR, Gregersen P, Klareskog L, Siminovitch KA, van Heel DA, Wijmenga C, Worthington J, Todd JA, Hafler DA, Rich SS, Daly MJ. FOCiS Network of Consortia. Pervasive sharing of genetic effects in autoimmune disease. *PLoS Genet* 2011;7:e1002254.
11 Ferreira MAR, Purcell SM. A multivariate test of association. *Bioinformatics* 2009;25:132–3.
12 Kim J, Bai Y, Pan W. An Adaptive Association Test for Multiple Phenotypes with GWAS Summary Statistics. *Genet Epidemiol* 2015;39:651–63.
13 Korte A, Vilhjálmsson BJ, Segura V, Platt A, Long Q, Nordborg M. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nat Genet* 2012;44:1066–71.

14 O'Reilly PF, Hoggart CJ, Pomyen Y, Calboli FC, Elliott P, Jarvelin MR, Coin LJ. MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. *PLoS One* 2012;7:e34861.

15 Zhu X, Feng T, Tayo BO, Liang J, Young JH, Franceschini N, Smith JA, Yanek LR, Sun YV, Edwards TL, Chen W, Nalls M, Fox E, Sale M, Bottinger E, Rotimi C, Liu Y, McKnight B, Liu K, Arnett DK, Chakravati A, Cooper RS, Redline S; COGENT BP Consortium. Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension. *Am J Hum Genet* 2015;96:21–36.

16 Ge T, Chen CY, Neale BM, Sabuncu MR, Smoller JW. Phenome-wide heritability analysis of the UK Biobank. *PLoS Genet* 2017;13:e1006711.

17 Muñoz M, Pong-Wong R, Canela-Xandri O, Rawlik K, Haley CS, Tenesa A. Evaluating the contribution of genetics and familial shared environment to common disease using the UK Biobank. *Nat Genet* 2016;48:980–3.

18 Obón-Santacana M, Vilardell M, Carreras A, Duran X, Velasco J, Galván-Femenía I, Alonso T, Puig L, Sumoy L, Duell EJ, Perucho M, Moreno V, de Cid R. GCAT|Genomes for life: a prospective cohort study of the genomes of Catalonia. *BMJ Open* 2018;8:e018324.

19 Liu Y, De A. Multiple Imputation by Fully Conditional Specification for Dealing with Missing Data in a Large Epidemiologic Study. *Int J Stat Med Res* 2015;4:287–95.

20 Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet* 2012;44:955–9.

21 Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR; 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* 2015;526:68–74.

22 Deelen P, Menelaou A, van Leeuwen EM, Kanterakis A, van Dijk F, Medina-Gomez C, Francioli LC, Hottenga JJ, Karssen LC, Estrada K, Kreiner-Møller E, Rivadeneira F, van Setten J, Gutierrez-Achury J, Westra HJ, Franke L, van Enckevort D, Dijkstra M, Byelas H, van Duijn CM, de Bakker PI, Wijmenga C, Swertz MA; Genome of Netherlands Consortium. Improved imputation quality of low-frequency and rare variants in European samples using the 'Genome of The Netherlands'. *Eur J Hum Genet* 2014;22:1321–6.

23 Huang J, Howie B, McCarthy S, Memari Y, Walter K, Min JL, Danecek P, Malerba G, Trabetti E, Zheng HF, Gambaro G, Richards JB, Durbin R, Timpson NJ, Marchini J, Soranzo N, Turki SA, Amuzu A, Anderson CA, Anney R, Antony D, Artigas MS, Ayub M, Bala S, Barrett JC, Barroso I, Beales P, Benn M, Bentham J, Bhattacharya S, Birney E, Blackwood D, Bobrow M, Bochukova E, Bolton PF, Bounds R, Boustred C, Breen G, Calissano M, Carss K, Casas JP, Chambers JC, Charlton R, Chatterjee K, Chen L, Ciampi A, Cirak S, Clapham P, Clement G, Coates G, Cocca M, Collier DA, Cosgrove C, Cox T, Craddock N, Crooks L, Curran S, Curtis D, Daly A, Inm D, Day-Williams A, Dedoussis G, Down T, Du Y, van DCM, Dunham I, Edkins S, Ekong R, Ellis P, Evans DM, Farooqi IS, Fitzpatrick DR, Flicek P, Floyd J, Foley AR, Franklin CS, Futema M, Gallagher L, Gasparini P, Gaunt TR, Geihs M, Geschwind D, Greenwood C, Griffin H, Grozeva D, Guo X, Guo X, Gurling H, Hart D, Hendricks AE, Holmans P, Huang L, Hubbard T, Humphries SE, Hurles ME, Hysi P, Iotchkova V, Isaacs A, Jackson DK, Jamshidi Y, Johnson J, Joyce C, Karczewski KJ, Kaye J, Keane T, Kemp JP, Kennedy K, Kent A, Keogh J, Khawaja F, Kleber ME, van KM, Kolb-Kokocinski A, Kooner JS, Lachance G, Langenberg C, Langford C, Lawson D, Lee I, van LEM, Lek M, Li R, Li Y, Liang J, Lin H, Liu R, Lönnqvist J, Lopes LR, Lopes M, Luan J, MacArthur DG, Mangino M, Marenne G, März W, Maslen J, Matchan A, Mathieson I, McGuffin P, McIntosh AM, McKechanie AG, McQuillin A, Metrustry S, Migone N, Mitchison HM, Moayyeri A, Morris J, Morris R, Muddyman D, Muntoni F, Nordestgaard BG, Northstone K, O'Donovan MC, O'Rahilly S, Onoufriadis A, Oualkacha K, Owen MJ, Palotie A, Panoutsopoulou K, Parker V, Parr JR, Paternoster L, Paunio T, Payne F, Payne SJ, Perry JRB, Pietilainen O, Plagnol V, Pollitt RC, Povey S, Quail MA, Quaye L, Raymond L, Rehnström K, Ridout CK, Ring S, Ritchie GRS, Roberts N, Robinson RL, Savage DB, Scambler P, Schiffels S, Schmidts M, Schoenmakers N, Scott RH, Scott RA, Semple RK, Serra E, Sharp SI, Shaw A, Shihab HA, Shin S-Y, Skuse D, Small KS, Smee C, Smith GD, Southam L, Spasic-Boskovic O, Spector TD, Clair DS, Pourcain BS, Stalker J, Stevens E, Sun J, Surdulescu G, Suvisaari J, Syrris P, Tachmazidou I, Taylor R, Tian J, Tobin MD, Toniolo D, Traglia M, Tybjaerg-Hansen A, Valdes AM, Vandersteen AM, Varbo A, Vijayarangakannan P, Visscher PM, Wain LV, Walters JTR, Wang G, Wang J, Wang Y, Ward K, Wheeler E, Whincup P, Whyte T, Williams HJ, Williamson KA, Wilson C, Wilson SG, Wong K, Xu C, Yang J, Zaza G, Zeggini E, Zhang F, Zhang P, Zhang W; UK10K Consortium. Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat Commun* 2015;6:8111.

24 McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, Kang HM, Fuchsberger C, Danecek P, Sharp K, Luo Y, Sidore C, Kwong A, Timpson N, Koskinen S, Vrieze S, Scott LJ, Zhang H, Mahajan A, Veldink J, Peters U, Pato C, van Duijn CM, Gillies CE, Gandin I, Mezzavilla M, Gilly A, Cocca M, Traglia M, Angius A, Barrett JC, Boomsma D, Branham K, Breen G, Brummett CM, Busonero F, Campbell H, Chan A, Chen S, Chew E, Collins FS, Corbin LJ, Smith GD, Dedoussis G, Dorr M, Farmaki AE, Ferrucci L, Forer L, Fraser RM, Gabriel S, Levy S, Groop L, Harrison T, Hattersley A, Holmen OL, Hveem K, Kretzler M, Lee JC, McGue M, Meitinger T, Melzer D, Min JL, Mohlke KL, Vincent JB, Nauck M, Nickerson D, Palotie A, Pato M, Pirastu N, McInnis M, Richards JB, Sala C, Salomaa V, Schlessinger D, Schoenherr S, Slagboom PE, Small K, Spector T, Stambolian D, Tuke M, Tuomilehto J, Van den Berg LH, Van Rheenen W, Volker U, Wijmenga C, Toniolo D, Zeggini E, Gasparini P, Sampson MG, Wilson JF, Frayling T, de Bakker PI, Swertz MA, McCarroll S, Kooperberg C, Dekker A, Altshuler D, Willer C, Iacono W, Ripatti S, Soranzo N, Walter K, Swaroop A, Cucca F, Anderson CA, Myers RM, Boehnke M, McCarthy MI, Durbin R; Haplotype Reference Consortium. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* 2016;48:1279–83.

25 Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* 2011;88:76–82.

26 Schweiger R, Fisher E, Rahmani E, Shenhav L, Rosset S, Halperin E. Using Stochastic Approximation Techniques to Efficiently Construct Confidence Intervals for Heritability: In. *Research in Computational Molecular Biology*. Cham: Springer, 2017:241–56.

27 Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 2015;4:7.

28 Dey R, Schmidt EM, Abecasis GR, Lee S. A Fast and Accurate Algorithm to Test for Binary Phenotypes and Its Application to PheWAS. *Am J Hum Genet* 2017;101:37–49.

29 Chavent M, Kuentz-Simonet V, Labenne A, Saracco J. *Multivariate analysis of mixed type data: The PCAmixdata R package*, 2014.

30 Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, Liu B, Matthews P, Ong G, Pell J, Silman A, Young A, Sprosen T, Peakman T, Collins R. 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:e1001779.

31 Yang Q, Wang Y. Methods for Analyzing Multivariate Phenotypes in Genetic Association Studies. *J Probab Stat* 2012;2012:1–13.

32 Euesden J, Lewis CM, O'Reilly PF. PRSice: Polygenic Risk Score software. *Bioinformatics* 2015;31:1466–8.

33 McEvoy B, Beleza S, Shriver MD. The genetic architecture of normal variation in human pigmentation: an evolutionary perspective and model. *Hum Mol Genet* 2006;15:R176–81.

34 Liu F, Visser M, Duffy DL, Hysi PG, Jacobs LC, Lao O, Zhong K, Walsh S, Chaitanya L, Wollstein A, Zhu G, Montgomery GW, Henders AK, Mangino M, Glass D, Bataille V, Sturm RA, Rivadeneira F, Hofman A, van IJcken WF, Uitterlinden AG, Palstra RJ, Spector TD, Martin NG, Nijsten TE, Kayser M. Genetics of skin color variation in Europeans: genome-wide association studies with functional follow-up. *Hum Genet* 2015;134:823–35.

35 Robles-Espinoza CD, Roberts ND, Chen S, Leacy FP, Alexandrov LB, Pornputtapong N, Halaban R, Krauthammer M, Cui R, Timothy Bishop D, Adams DJ. Germline MC1R status influences somatic mutation burden in melanoma. *Nat Commun* 2016;7:12064.

36 Sturm RA. Skin colour and skin cancer - MC1R, the genetic link. *Melanoma Res* 2002;12:405–16.

37 McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. The Ensembl Variant Effect Predictor. *Genome Biol* 2016;17:122.

38 Wysocki K, Ritter L. Diseasome: an approach to understanding gene-disease interactions. *Annu Rev Nurs Res* 2011;29:55–72.

39 Yang J, Bakshi A, Zhu Z, Hemani G, Vinkhuyzen AA, Lee SH, Robinson MR, Perry JR, Nolte IM, van Vliet-Ostaptchouk JV, Snieder H, Esko T, Milani L, Mägi R, Metspalu A, Hamsten A, Magnusson PK, Pedersen NL, Ingelsson E, Soranzo N, Keller MC, Wray NR, Goddard ME, Visscher PM; LifeLines Cohort Study. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat Genet* 2015;47:1114–20.

40 Krapohl E, Rimfeld K, Shakeshaft NG, Trzaskowski M, McMillan A, Pingault JB, Asbury K, Harlaar N, Kovas Y, Dale PS, Plomin R. The high heritability of educational achievement reflects many genetically influenced traits, not just intelligence. *Proc Natl Acad Sci U S A* 2014;111:15273–8.

41 Turley P, Walters RK, Maghzian O, Okbay A, Lee JJ, Fontana MA, Nguyen-Viet TA, Wedow R, Zacher M, Furlotte NA, Magnusson P, Oskarsson S, Johannesson M, Visscher PM, Laibson D, Cesarini D, Neale BM, Benjamin DJ; 23andMe Research Team, Social Science Genetic Association Consortium. Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat Genet* 2018;50:229–37.

42 Weedon MN, Lango H, Lindgren CM, Wallace C, Evans DM, Mangino M, Freathy RM, Perry JR, Stevens S, Hall AS, Samani NJ, Shields B, Prokopenko I, Farrall M, Dominiczak A, Johnson T, Bergmann S, Beckmann JS, Vollenweider P, Waterworth DM, Mooser V, Palmer CN, Morris AD, Ouwehand WH, Zhao JH, Li S, Loos RJ, Barroso I, Deloukas P, Sandhu MS, Wheeler E, Soranzo N, Inouye M, Wareham NJ, Caulfield M, Munroe PB, Hattersley AT, McCarthy MI, Frayling TM; Diabetes Genetics Initiative, Wellcome Trust Case Control Consortium, Cambridge GEM Consortium. Genome-wide association analysis identifies 20 loci that influence adult height. *Nat Genet* 2008;40:575–83.

43 McClay JL, Adkins DE, Åberg K, Bukszár J, Khachane AN, Keefe RSE, Perkins DO, McEvoy JP, Stroup TS, Vann RE, Beardsley PM, Lieberman JA, Sullivan PF, van den Oord EJCG. Genome-wide pharmacogenomic study of neurocognition as an indicator of antipsychotic treatment response in schizophrenia. *Neuropsychopharmacology* 2011;36:616–26.

44 Lee MK, Shaffer JR, Leslie EJ, Orlova E, Carlson JC, Feingold E, Marazita ML, Weinberg SM. Genome-wide association study of facial morphology reveals novel associations with FREM1 and PARK2. *PLoS One* 2017;12:e0176566.

45 Hill WD, Marioni RE, Maghzian O, Ritchie SJ, Hagenaars SP, McIntosh AM, Gale CR, Davies G, Deary IJ. A combined analysis of genetically correlated traits identifies 187 loci and a role for neurogenesis and myelination in intelligence. *Mol Psychiatry*;15.

46 Wang X, Sun J, Li C, Mao B. EphA7 modulates apical constriction of hindbrain neuroepithelium during neurulation in Xenopus. *Biochem Biophys Res Commun* 2016;479:759–65.

47 Prost G, Braun S, Hertwig F, Winkler M, Jagemann L, Nolbrant S, Leefa IV, Offen N, Miharada K, Lang S, Artner I, Nuber UA. The putative tumor suppressor gene EphA7 is a novel BMI-1 target. *Oncotarget* 2016;7:58203–17.

48 Choi H, Lee SH, Um SJ, Kim EJ. CACUL1 functions as a negative regulator of androgen receptor in prostate cancer cells. *Cancer Lett* 2016;376:360–6.

49 Voorhoeve PG, van Mechelen W, Uitterlinden AG, Delemarre-van de Waal HA, Lamberts SW. Androgen receptor gene CAG repeat polymorphism in longitudinal height and body composition in children and adolescents. *Clin Endocrinol* 2011;74:732–5.

50 Tsukasaki K, Miller CW, Greenspun E, Eshaghian S, Kawabata H, Fujimoto T, Tomonaga M, Sawyers C, Said JW, Koeffler HP. Mutations in the mitotic check point gene, MAD1L1, in human cancers. *Oncogene* 2001;20:3301–5.

51 Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium. Genome-wide association study identifies five new schizophrenia loci. *Nat Genet* 2011;43:969–76.

52 Ng MC, Shriner D, Chen BH, Li J, Chen WM, Guo X, Liu J, Bielinski SJ, Yanek LR, Nalls MA, Comeau ME, Rasmussen-Torvik LJ, Jensen RA, Evans DS, Sun YV, An P, Patel SR, Lu Y, Long J, Armstrong LL, Wagenknecht L, Yang L, Snively BM, Palmer ND, Mudgal P, Langefeld CD, Keene KL, Freedman BI, Mychaleckyj JC, Nayak U, Raffel LJ, Goodarzi MO, Chen YD, Taylor HA, Correa A, Sims M, Couper D, Pankow JS, Boerwinkle E, Adeyemo A, Doumatey A, Chen G, Mathias RA, Vaidya D, Singleton AB, Zonderman AB, Igo RP, Sedor JR, Kabagambe EK, Siscovick DS, McKnight B, Rice K, Liu Y, Hsueh WC, Zhao W, Bielak LF, Kraja A, Province MA, Bottinger EP, Gottesman O, Cai Q, Zheng W, Blot WJ, Lowe WL, Pacheco JA, Crawford DC, Grundberg E, Rich SS, Hayes MG, Shu XO, Loos RJ, Borecki IB, Peyser PA, Cummings SR, Psaty BM, Fornage M, Iyengar SK, Evans MK, Becker DM, Kao WH, Wilson JG, Rotter JI, Sale MM, Liu S, Rotimi CN, Bowden DW. FIND Consortium eMERGE Consortium DIAGRAM Consortium MuTHER Consortium MEta-analysis of type 2 DIabetes in African Americans Consortium. Meta-analysis of genome-wide association studies in African Americans provides insights into the genetic architecture of type 2 diabetes. *PLoS Genet* 2014;10:e1004517.

53 Demerath EW, Guan W, Grove ML, Aslibekyan S, Mendelson M, Zhou YH, Hedman ÅK, Sandling JK, Li LA, Irvin MR, Zhi D, Deloukas P, Liang L, Liu C, Bressler J, Spector TD, North K, Li Y, Absher DM, Levy D, Arnett DK, Fornage M, Pankow JS, Boerwinkle E, ÅK H, Li L-A IMR. Epigenome-wide association study (EWAS) of BMI, BMI change and waist circumference in African American adults identifies multiple replicated loci. *Hum Mol Genet* 2015;24:4464–79.

54 Rzehak P, Covic M, Saffery R, Reischl E, Wahl S, Grote V, Weber M, Xhonneux A, Langhendries JP, Ferre N, Closa-Monasterolo R, Escribano J, Verduci E, Riva E, Socha P, Gruszfeld D, Koletzko B. DNA-Methylation and Body Composition in Preschool Children: Epigenome-Wide-Analysis in the European Childhood Obesity Project (CHOP)-Study. *Sci Rep* 2017;7:14349.

55 Lotta LA, Gulati P, Day FR, Payne F, Ongen H, van de Bunt M, Gaulton KJ, Eicher JD, Sharp SJ, Luan J, De Lucia Rolfe E, Stewart ID, Wheeler E, Willems SM, Adams C, Yaghootkar H, Forouhi NG, Khaw KT, Johnson AD, Semple RK, Frayling T, Perry JR, Dermitzakis E, McCarthy MI, Barroso I, Wareham NJ, Savage DB, Langenberg C, O'Rahilly S, Scott RA; EPIC-InterAct Consortium Cambridge FPLD1 Consortium. Integrative genomic analysis implicates limited peripheral adipose storage capacity in the pathogenesis of human insulin resistance. *Nat Genet* 2017;49:17–26.

56 Castaño-Betancourt MC, Evans DS, Ramos YF, Boer CG, Metrustry S, Liu Y, den Hollander W, van Rooij J, Kraus VB, Yau MS, Mitchell BD, Muir K, Hofman A, Doherty M, Doherty S, Zhang W, Kraaij R, Rivadeneira F, Barrett-Connor E, Maciewicz RA, Arden N, Nelissen RG, Kloppenburg M, Jordan JM, Nevitt MC, Slagboom EP, Hart DJ, Lafeber F, Styrkarsdottir U, Zeggini E, Evangelou E, Spector TD, Uitterlinden AG, Lane NE, Meulenbelt I, Valdes AM, van Meurs JB. Novel Genetic Variants for Cartilage Thickness and Hip Osteoarthritis. *PLoS Genet* 2016;12:e1006260.

57 Mullin BH, Walsh JP, Zheng HF, Brown SJ, Surdulescu GL, Curtis C, Breen G, Dudbridge F, Richards JB, Spector TD, Wilson SG. Genome-wide association study using family-based cohorts identifies the WLS and CCDC170/ESR1 loci as associated with bone mineral density. *BMC Genomics* 2016;17:136.

58 Dyment DA, Smith AC, Alcantara D, Schwartzentruber JA, Basel-Vanagaite L, Curry CJ, Temple IK, Reardon W, Mansour S, Haq MR, Gilbert R, Lehmann OJ, Vanstone MR, Beaulieu CL, Majewski J, Bulman DE, O'Driscoll M, Boycott KM, Innes AM; FORGE Canada Consortium. Mutations in PIK3R1 cause SHORT syndrome. *Am J Hum Genet* 2013;93:158–66.

59 Majumdar A, Haldar T, Bhattacharya S, Witte JS. An efficient Bayesian meta-analysis approach for studying cross-phenotype genetic associations. *PLoS Genet* 2018;14:e1007139.

Appendix 5. Genomic profiling in advanced stage non-small-cell lung cancer patients with platinum-based chemotherapy identifies germline variants with prognostic value in SMYD2.

Galván-Femenía I, Guindo M, Duran X, Calabuig-Fariñas S, Mercader JM, Ramirez JL, Rosell R, Torrents D, Carreras A, Kohno T, Jantus-Lewintre E, Camps C, Perucho M, Sumoy L, Yokota J, de Cid R. *Cancer Treat Res Commun*. 2018;15:21-31.

Contribution:

- Genotype imputation using 1000G phase 3, UK10K and GoNL as reference panels.
- Statistical analysis and interpretation of data.

Cancer Treatment and Research Communications 15 (2018) 21–31

Contents lists available at ScienceDirect

Cancer Treatment and Research Communications

journal homepage: www.elsevier.com/locate/ctarc

Genomic profiling in advanced stage non-small-cell lung cancer patients with platinum-based chemotherapy identifies germline variants with prognostic value in *SMYD2*

Iván Galván-Femenía[a], Marta Guindo[b], Xavier Duran[a], Sílvia Calabuig-Fariñas[c,d,e], Josep Maria Mercader[b,1,2], Jose Luis Ramirez[f], Rafael Rosell[f], David Torrents[b,g], Anna Carreras[a], Takashi Kohno[h], Eloisa Jantus-Lewintre[d,e,i], Carlos Camps[c,d,j,k], Manuel Perucho[f], Lauro Sumoy[l], Jun Yokota[f], Rafael de Cid[a,*]

[a] *Genomes For life-GCAT Lab. Program of Predictive and Personalized Medicine of Cancer (PMPPC), Institute for Health Science Research Germans Trias i Pujol (IGTP), Can Ruti Biomedical Campus, Crta de Can Ruti, Camí de les Escoles S/N, 08916 Badalona, Barcelona, Spain*
[b] *Barcelona Supercomputing Center (BSC-CNS), Joint BSC-CRG-IRB Research Program in Computational Biology, Carrer de Jordi Girona, 29-31, 08034 Barcelona, Spain*
[c] *Department of Medical Oncology, Hospital General Universitario de Valencia, Avenida Tres Cruces, 2, 46014, València, Spain*
[d] *Molecular Oncology Laboratory, Fundación Hospital General Universitario de Valencia, Avda. Tres Cruces s/n 46014 València, Spain*
[e] *Department of Pathology, Universitat de València, Av. de Blasco Ibáñez, 13, 46010 València, Spain*
[f] *Program of Predictive and Personalized Medicine of Cancer (PMPPC), Institute for Health Science Research Germans Trias i Pujol (IGTP), Can Ruti Biomedical Campus, Crta de Can Ruti, Camí de les Escoles S/N, 08916 Badalona, Barcelona, Spain*
[g] *ICREA, Catalan Institution for Research and Advanced Studies, Spain*
[h] *Division of Genome Biology, National Cancer Center Research Institute, 5-1-1 Tsukiji, Chuo-ku, Tokyo 104-0045, Japan*
[i] *Molecular Oncology Laboratory, Fundación Hospital General Universitario de València, Avda. Tres Cruces s/n, 46014 València*
[j] *Department of Biotechnology, Universitat Politècnica de València, Camí de Vera, s/n, 46022 València, Spain*
[k] *Department of Medicine, Universitat de València, Av. de Blasco Ibáñez, 13, 46010 València, Spain*
[l] *Genomics and Bioinformatics. Program of Predictive and Personalized Medicine of Cancer (PMPPC), Institute for Health Science Research Germans Trias i Pujol (IGTP), Can Ruti Biomedical Campus, Crta de Can Ruti, Camí de les Escoles S/N, 08916 Badalona, Barcelona, Spain*

ARTICLE INFO

Keywords:
Lung cancer
NSCLC
Advanced stage
Prognostic factors
Genome-Wide-Association Studies

ABSTRACT

Objective: The aim of the study was to investigate the relationship between germline variations as a prognosis biomarker in patients with advanced Non-Small-Cell-Lung-Cancer (NSCLC) subjected to first-line platinum-based treatment.

Materials and Methods: We carried out a two-stage genome-wide-association study in non-small-cell lung cancer patients with platinum-based chemotherapy in an exploratory sample of 181 NSCLC patients from Caucasian origin, followed by a validation on 356 NSCLC patients from the same ancestry (Valencia, Spain).

Results: We identified germline variants in *SMYD2* as a prognostic factor for survival in patients with advanced NSCLC receiving chemotherapy. *SMYD2* alleles are associated to a decreased overall survival and with a reduced Time to Progression. In addition, enrichment pathway analysis identified 361 variants in 40 genes to be involved in poorer outcome in advanced-stage NSCLC patients.

Conclusion: Germline *SMYD2* alleles are associated with bad clinical outcome of first-line platinum-based treatment in advanced NSCLC patients. This result supports the role of *SMYD2* in the carcinogenic process, and might be used as prognostic signature directing patient stratification and the choice of therapy.

Microabstract: A two-Stage Genome wide association study in Caucasian population reveals germline genetic variation in *SMYD2* associated to progression disease in first-line platinum-based treatment in advanced NSCLC

* Corresponding author.
E-mail addresses: igalvan@igtp.cat (I. Galván-Femenía), marta.guindomartinez@bsc.es (M. Guindo), xduran@igtp.cat (X. Duran), calabuix_sil@gva.es (S. Calabuig-Fariñas), mercader@broadinstitute.org (J.M. Mercader), jramirez@iconcologia.net (J.L. Ramirez), rrosell@iconcologia.net (R. Rosell), david.torrents@bsc.es (D. Torrents), acarreras@igtp.cat (A. Carreras), tkkohno@ncc.go.jp (T. Kohno), jantus_elo@gva.es (E. Jantus-Lewintre), camps_car@gva.es (C. Camps), mperucho@igtp.cat (M. Perucho), lsumoy@igtp.cat (L. Sumoy), jyokota@igtp.cat (J. Yokota), rdecid@igtp.cat (R. de Cid).

[1] Programs in Metabolism and Medical & Population Genetics, Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA.

[2] Diabetes Unit and Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts, USA.

https://doi.org/10.1016/j.ctarc.2018.02.003
Received 24 July 2017; Received in revised form 26 January 2018; Accepted 19 February 2018

Table 1
Clinical and pathological characteristics of the discovery (BREC, n = 178) and validation sample (n = 323).

		BREC		**Disease progression**					**Validation Sample**		**Disease progression**					**BREC + Validation Sample**		
				NO		**YES**					**NO**		**YES**					
		N	**%**	**N**	**%**	**N**	**%**	**p-value$_1$**	**N**	**%**	**N**	**%**	**N**	**%**	**p-value$_2$**	**NBREC**	**NValidation**	**p-value$_3$**
Gender								0.29							0.55			0.16
	Male	139	78	104	76	35	85		270	83	171	83	99	85		139	270	
	Female	39	22	33	24	6	15		53	17	36	17	17	15		39	53	
Smoker								0.16							-			
	Yes	167	94	126	92	41	100											
	No	10	6	10	7	0	0											
	NA	1	0	1	1	0	0											
ECOG								0.73							**0.002**			0.23
	0	59	33	45	33	14	34		95	29	74	36	21	18		59	95	
	1	114	64	88	65	26	64		220	68	128	61	92	80		114	220	
	2	2	1	2	1	0	0		7	2	4	2	3	2		2	7	
	NA	3	2	2	1	1	2		1	1	1	1	0	0		3	1	
Histology								**0.006**							0.83			**0.002**
	ADCA	99	56	83	61	16	39		164	51	105	51	59	51		99	164	
	SCC	64	36	44	32	20	49		101	31	67	32	34	29		64	101	
	LCC	6	3	6	4	0	0		45	14	28	13	17	15		6	45	
	Others	9	5	4	3	5	12		13	4	7	4	6	5		9	13	
Treatment								0.66							1			-
	doce/cis	123	69	93	68	30	73		323	100	207	100	116	100				
	gemci/cis	44	25	36	26	8	20											
	doce	11	6	8	6	3	7											
Stage								1							0.70			**0.0001**
	III	7	4	5	4	2	5		52	15	35	17	17	15		7	52	
	IV	164	92	127	92	37	90		271	85	172	83	99	85		164	271	
	NA	7	4	5	4	2	5											
RECIST																		**6.6×10^{-9}**
	PD	41	23						123	37						41	123	
	SD	56	31						113	34						56	113	
	PR	58	32						93	28						58	93	
	CR	23	14						3	1						23	3	

ECOG, performance status; doce, docetaxel; cis, cisplatin; gemci, gemcitabine; RECIST, response evaluation criteria in solid tumors; PD, progression disease; SD, stable disease; PR, partial response; CR, complete response; NA, not available.

The **p-value$_1$** and **p-value$_2$** columns show the difference between progression disease regarding each clinical variable in BREC and validation sample respectively.

The **p-value$_3$** column show the difference between the number of patients in BREC and validation sample for each clinical variable.

patients. *SMYD2* profiling might have prognostic / predictive value directing choice of therapy and enlighten current knowledge on pathways involved in human carcinogenesis as well in resistance to chemotherapy.

1. Introduction

Lung cancer is the most common cancer in the world, and the leading cause of mortality among cancer-related deaths. The Non-Small-Cell-Lung-Cancer (NSCLC), being the most common form, has an overall 5-years survival of less than 15% [15]. NSCLC is a histological diverse group of tumors, with major classes being squamous (SCC), adenocarcinoma (ADC), and large cell carcinoma (LCC). Despite the enormous heterogeneity, these tumors have been treated homogeneously for a long time with cytotoxic chemotherapy as the choice treatment [17].

Platinum-based chemotherapy is still widely used for treatment of the vast majority of NSCLC patients with advanced-stage disease, with the exception of cases bearing E*GFR, BRAF*, *ROS1* or *EML4-ALK tumor* mutations. The latter have greatly benefited from the advances achieved in the last ten years in targeted therapy based on somatic genetic/molecular profiling. While chemotherapy provides palliation, advanced NSCLC remains incurable in most cases since acquired resistance is common, response rates are only 15%–30%, and median OS is less than 12 months. Resistance can arise from a several causes (drug delivery, altered target, tolerance to damage, etc…) [16] and differences observed in therapy efficacy could be explained by the impact of host genotype variants on target/resistance factors.

Customization of chemotherapy has relied on tumor cell expression profiles of specific genes. Candidate gene studies have indicated possible association to response of genetic variants in genes of the platinum pathway (reviewed by Hildebrandt et al. [20]) and DNA-repair genes [32]. Genome-wide association studies (GWAS) have been used successfully to identify germline genetic variants associated with an increased risk of developing lung cancer including NSCLC, and have been applied to identify prognostic biomarkers [22,27,49,50,54,55] as well as to identify genetic variability associated to adverse effects to chemotherapy [6,7]. Furthermore, re-positioning of GWAS-derived germline predisposition markers as prognostic markers, have been successfully reported in other cancer forms.

The aim of this study was to investigate the relationship between germline variants to identify prognosis biomarkers for clinical outcome in patients with advanced NSCLC subjected to first-line platinum-based treatment. In this study we report a genome-wide scan study in two independent samples from the same ancestry (Spain).

We present evidence of germline variation in the *SMYD2* affecting the clinical outcome of advanced NSCLC patients. *SMYD2* profiling might have prognostic/predictive value directing choice of therapy and enlighten current knowledge on pathways involved in human carcinogenesis as well in resistance to chemotherapy.

2. Material and Methods

2.1. Patients

This study was approved by the institutional review board of the IGTP. The recruitment of NSCLC patients in the discovery phase and validation phase was approved by the institutional review board of each participating institution.

2.2. Discovery sample

Patients included in the study were selected from the BREC clinical trial study (Multicenter, Predictive, Prospective, Phase III, Open, Randomized, Pharmacogenomic Study in Patients with Advanced Lung Carcinoma (BREC)) [35]. BREC patients with advanced NSCLC who had not received treatment for the disease at the time they entered the study and had a good performance status (ECOG 0–1) and measurable disease (at least one target lesion according to the RECIST (response evaluation criteria in solid tumors), received six to eight chemotherapy cycles. The 94% received cisplatin 75 mg/m2 combined with Docetaxel 75 mg/m2 (73%) (group 1) or Gemcitabine 1250 mg/m2 (27%) (group 2), both at day 1, in 21-day cycles. Remaining 6% received Docetaxel 75 mg/m2 (group 3), on day 1 every 3 weeks. See complete description at clinicaltrials.gov/show/NCT00617656.

A total of 178 patients were included in the genetic analysis. All considered patients were *EGFR*-WT. The median age was 62 years (range: 39–82), 78% males, and 92% stage IV of the disease. ADC was the most common histological subtype (56% of patients) of NSCLC, followed by squamous cell carcinoma (SCC) (36%) and large cell carcinoma (LCC) (3%), 5% were grouped in other categories. The most frequent ECOG score was 1 (64%) (33% and 1% for 0 and 2 status). Overall progression free survival (PFS) (calculated from the date of randomization to progression or death) was 5.3 months (95% CI 4.71–5.88), and survival time (Overall Survival OS; calculated from the date of randomization to death) was 10.16 months (95% CI 8.32–12.01).

2.3. Validation sample

Patients included in the validation cohort were from a Multicenter study coordinated by the Spanish Lung Cancer Group. All patients were with advanced NSCLC, from Caucasian ancestry and the same geographical region (Valencia, Spain) [24,25,47]. Blood samples were recollected from 356 NSCLC stage IIIB with pleural effusion or stage IV, who received cisplatin (75 mg/m^2) and docetaxel (75 mg/m^2) on day 1 every 3 weeks. Among 356 patients, 323 with fulfilled response data were considered for the analysis.

The median age was 59 years (range: 31–80), 83% males. 15% of patients had stage III and 85% stage IV of the disease. Like in BREC patients, ADC was the most common histological subtype (51%) of NSCLC, followed by SCC (31%) and LCC (14%), 4% were grouped in other categories. The most frequent ECOG score was 1 (68%) (29% and 2% for 0 and 2 status). TTP was 5.53 months (95% CI 4.93–6.33) and OS 9.9 months (95% CI 9.17–11.07).

According to the study objectives, the clinical outcomes were diagnosis of NSCLC and response to treatment (according to the criteria established in the RECIST). Patients were categorized as progressing disease if its RECIST was assessed as PD (PD) (23% BREC, 37% validation sample) and as non-progressing if their RECIST was complete (CR) or partial response (PR) (14%, 1% and 32%, 28%) or stable disease (31%, 34%)(SD), in both exploratory and replication cohorts, respectively. The main clinical and pathological characteristics of the discovery and validation samples are shown in Table 1.

2.4. Genome scan

2.4.1. Genotyping

Genome-wide genotypes were generated for the discovery sample using SNP-array technology. The Infinium® HTS Assay automated protocol, was used on HumanCoreExome-24v1-0 BeadChips scanned with a HiScan confocal scanner (ILLUMINA, San Diego, CA). Genotyping was performed at the Genomic Units of the PMPPC-IGTP. Genome Studio version 2011.1 was used for raw data analysis. All illumina internal system controls were fulfilled. Before the genetic association analysis, we conducted systematic quality control on the raw genotyping data to filter both unqualified samples and SNPs. Overall call rate was 99.89%. Samples were excluded if they failed genotyping in more than 10% of variants. Variants were excluded if they failed genotyping in more than 10% of samples, were non-polymorphic, or showed departure from Hardy-Weinberg Equilibrium (HWE) (p value > 0.0001). 40% of genotyped markers were monomorphic in our sample. Gender control detected a mismatch in one sample that was included in the study after database consultation. After these quality control steps, 181 cases with 325,762 SNPs were considered. PLINK 1.9 version [9,43] was used to perform the quality control analysis. Genotyping of candidate SNPS in the replication sample was done at the Spanish National Centre of Genotyping (ISCIII-CEGEN-Santiago Node) facility by using the iPLEX Sequenom MassARRAY platform (Sequenom Inc., San Diego, CA, USA) and at PMPPC-IGTP by Real-Time PCR, using TaqMan^{TM} (ILLUMINA, San Diego, CA) when do not fit Sequenom's basics.

2.4.2. *In silico genotyping*

IMPUTE2 [21] was used to impute untyped SNPs from sequence-based reference panels (1000Genomes, UK10K, GoNL). SHAPEIT [11] was used for haplotype estimation prior to imputation procedures. Imputed genotypes with IMPUTE2info lower than 0.7 were discarded for association analysis. The best IMPUTE information score was used for those SNPs present in more than one reference panel. Finally, from 24,873,940 imputed SNPs we considered 10,307,177 unique SNPs for association analysis.

2.5. *Population structure and relatedness*

All patients in the discovery sample were used to detect population substructure and independence. Principal Component Analysis (PCA) was applied to detect any hidden substructure, and method of moments (MoM) for the estimation of identity by descent probabilities was applied to exclude cases with cryptic relatedness. A Spanish population based cohort GCAT (genomesforlife.com) and public databases (HapMap) were used to test ancestry homogeneity before imputation analysis [2].

2.6. *Statistical analysis*

We perform a multivariate logistic regression model, under an additive model, adjusted by gender, smoking (yes/no), tumor histology (i.e. ADCA, SCC, LCC, other), pretreatment performance status (ECOG score) (0, 1, > 2), chemotherapy treatment group (docetaxel/cisplatin, gemcitabine/cisplatin, docetaxel) and the first seven principal components (PC) as covariates. Genomic control inflation for the association results was calculated from observed and imputed data ($\lambda = 1.12$). All p-values were corrected for genomic inflation factor.

For the replication analysis we considered those SNPs with corrected p-values $< 1 \times 10^{-5}$ (Fig. 1). For validation purposes, due to the relative small sample size and the inflated or deflated size effect for SNPs with MAF < 0.01 generated from imputation methods, we considered those with an effect size (OR) in the [0.05–20] range. From suggestive signals, alternative SNPs were selected with LDlink [30] and FINEMAP software [3]. Both tools were used for exploring possible functional variants via linkage disequilibrium and a shotgun stochastic search algorithm. Selected candidates are shown in Fig. 2.

We analyze all candidates SNPs in the validation sample under the same model assumptions, but excluding smoking, since was not relevant in the BREC analysis, and was not available in the replication sample. Then a joint analysis was performed. Since differences were evident among cohorts (Table 1), a heterogeneity analysis was performed and I^2 measure was estimated. Heterogeneity source was investigated by a multiple correspondence analysis [26] to detect any data structuring within BREC and Valencia sample regarding gender, histology, stage and ECOG categorical variables (Supplementary Fig. 1). For replication, we performed a matched analysis with resampling. Each Valencia's individual was matched with BREC cohort by disease progression and stage to preserve the same clinical features before association analysis. Then, we resampled 10,000 times and p-values were derived and ranked. A p-value representing the 5% percentile of the p-values distribution [13] was considered for each SNP.

Cox proportional hazard regression models were used to evaluate survival outcomes (TTP and OS) in the validation cohort, and multivariate analysis was performed adjusting the Cox models by age, gender, ECOG and disease progression status. No individual data was available from the BREC cohort. Survival curves were computed with the Kaplan–Meier estimator. Hazard ratios (HRs) and their 95% confidence intervals were assessed to evaluate the risk of death.

We used SNPtest software [31] for association analysis in the discovery sample, and PLINK 1.9 version for association analysis in the validation sample. SNPtest allows worked seamlessly with imputation data. R software (version 3.3.1, R Core Team, 2016) was used for data visualization (Manhattan plot, QQ plot, Kaplan-Meier and ROC curves) and statistical analyses. Data visualization was made with LocusZoom, for plotting chromosomal regions.

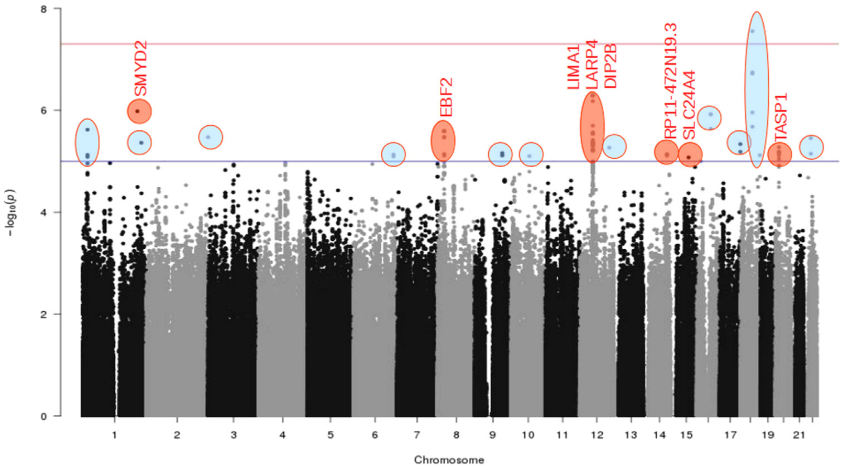


Fig. 1. Manhattan plot for genome-wide association results in the BREC discovery sample. Association p-values are expressed as -log10(p). P-values comes from multivariate models accounting for gender, smoking status, histology, ECOG performance status, chemotherapy treatment and the first seven principal components. Red circles are the selected peaks used for replication purposes (MAF > 0.01 and 0.05 > OR < 20). The blue and red lines indicate the p-value threshold for the candidate genes at -log10(10^5) and -log10(5×10^8) respectively.

CHR:BP	Alleles	Freq.	Info	Proxy	R2	P−val	OR	CI95% OR
1:214481630	G/A*	0.47	1.00	FINEMAP	0.59	8×10^{-5}	0.23	[0.11−0.5]
1:214495703	C/T*	0.47	1.00	BI_p	0.59	1.3×10^{-4}	0.25	[0.12−0.52]
1:214502898	C/G*	0.31	0.99	BI_p	0.65	1.9×10^{-3}	3.30	[1.52−7.15]
1:214503489	T*/G	0.41	0.99	BI	−	1×10^{-6}	7.02	[2.88−17.12]
8:25643577	G/A*	0.50	1.00	FINEMAP		8×10^{-5}	0.22	[0.1−0.49]
8:25647711	A/G*	0.32	1.00	BI_p		1.3×10^{-5}	5.81	[2.43−13.86]
8:25652367	T/C*	0.31	1.00	BI_p		2.6×10^{-6}	7.40	[2.89−18.9]
8:25662655	G/T*	0.31	0.99	FINEMAP		3.3×10^{-6}	7.36	[2.87−18.89]
12:50589836	A/G*	0.20	1.4×10^{-4}	BO_p	0.77	4.6×10^{-4}	5.25	[1.92−14.32]
12:50645471	A/G*	0.21	1.00	BO	−	5.1×10^{-7}	17.10	[4.23−69.14]
12:50824232	G/A*	0.20	0.97	BI	−	3.8×10^{-6}	14.42	[3.55−58.58]
12:50881148	C/A*	0.26	0.99	BI_p	0.61	1.2×10^{-5}	6.96	[2.62−18.49]
12:50947729	G*/T	0.29	1.00	FINEMAP	0.85	3×10^{-5}	5.30	[2.22−12.64]
12:50959426	G/A*	0.25	1.00	BO	−	4.6×10^{-6}	8.30	[2.93−23.55]
12:51015509	A/G*	0.26	1.00	BO_p	0.97	1.3×10^{-5}	6.94	[2.6−18.54]
12:51026878	G/T*	0.17	1.00	FINEMAP	0.07	3.7×10^{-5}	0.20	[0.09−0.44]
14:92726738	A*/G	0.30	1.00	BO_p	0.37	6.9×10^{-3}	0.39	[0.2−0.76]
14:92726813	C*/G	0.22	0.99	FINEMAP	0.61	1.4×10^{-4}	0.23	[0.11−0.5]
14:92729757	C/A*	0.27	1.00	BO_p	1	7.3×10^{-6}	0.19	[0.09−0.41]
14:92729907	G/T*	0.27	1.00	BO	−	7.2×10^{-6}	0.19	[0.09−0.41]
15:70089618	A*/G	0.47	0.96	FINEMAP	0.56	3.3×10^{-4}	0.30	[0.15−0.59]
15:70095755	T/C*	0.21	1.00	BO_p	0.27	6.2×10^{-3}	0.38	[0.2−0.74]
15:70098476	A/G*	0.48	0.96	BO_p	0.68	4.2×10^{-5}	0.24	[0.12−0.5]
15:70104713	C*/T	0.46	1.00	BO	−	8.4×10^{-6}	0.23	[0.11−0.46]
18:46538022	A*/G	0.12	0.76	BI	−	2.8×10^{-8}	0.01	[0.002−0.11]
20:13582476	C/A*	0.35	1.00	BO	−	1.2×10^{-5}	5.15	[2.36−11.26]
20:13582879	T/G*	0.33	1.00	BO_p	0.94	6.9×10^{-6}	5.52	[2.48−12.32]
20:13612157	G/T*	0.40	1.00	BO_p	0.73	1×10^{-3}	3.19	[1.58−6.43]

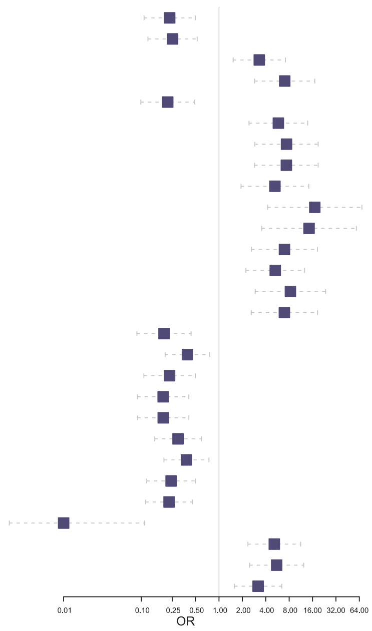


Fig. 2. Forest plot diagram of the association results of the BREC discovery dataset used for replication analysis. The variants are listed by chromosome and position (CHR:BP) showing the IMPUTE information measure (Info) and the effect size (OR) regarding the first allele of the Alleles column. BO, best observed; BI, best imputed; BO_p, best observed proxy; BI_p, best imputed proxy.

2.7. Pathway enrichment analysis

In order to provide biological hypotheses from our GWAS results we performed a pathway analysis to highlight enriched pathways based on genes in associated loci. All genes with at least one variant at p-value $< 1\times10^{-4}$ were included in the analysis. We used the seq2-pathway R package [51] to select the subset of the most significant genes within a search radius of 150 kbps from the SNPs with an association p-value below 1×10^{-4}. Pathway enrichment analysis of the 889 selected genes was performed against Gene Ontology and Reactome annotation data with both seq2pathway and PANTHER Overrepresentation Test tool (release 20160715) [34]. The significance of the GO terms was estimated through the adjusted p-values based on the binomial testing with Bonferroni correction for multiple hypotheses.

2.8. Fine-mapping and functional annotation

Variant Effect Predictor (VEP) tool [33] was used for the functional characterization of identified variants (hg19). The VEP application determines the effect of variants (SNPs, insertions, deletions, CNVs or structural variants) on genes, transcripts, protein and regulatory regions.

3. Results

3.1. Clinical and pathological characteristics of the two-stage used cohorts

Bivariate analysis of the clinical and pathological characteristics shows differences in tumor histological type (p = 0.002), stage (p = 0.0001) and progression disease ($p = 6.6\times10^{-9}$), with more cases of LC, stage III and PD in the validation sample than in BREC. No other differences in gender, age, and pretreatment performance status were statistically significant. Regarding PD, we observed differences in tumor histological type, slightly different in the discovery sample, but not in the replication sample, and ECOG, related to PD in the Valencia sample but not in BREC. Concerning chemotherapy treatment group, no significant differences were observed in BREC. All statistically significant differences were considered as covariates in further analyses. The clinical and pathological characteristics of the study population are shown in Table 1.

3.2. Twenty genomic regions show association with disease progression outcome in the discovery sample

PCA analysis indicated that the BREC as an ethnically homogenous Caucasian. All patients except three overlapped with the CEU ancestry reference panel from HapMap and with the geographically matched sample from the Spanish GCAT cohort (genomesforlife.com). The three genetically distant patients were discarded for the genomic analysis. The first seven PCA dimensions were incorporated in the association analysis as covariates. No cryptic relatedness was found by estimating identity by descent (IBD) probabilities.

Association analysis was made with observed and imputed data recovered from three public reference panels. In the discovery phase we observed one SNP with p-value $< 1\times10^{-8}$, two SNPs with p-value $< 1\times10^{-7}$, 22 SNPs with p-value $< 1\times10^{-6}$, 147 SNPs with p-value $< 1\times10^{-5}$, 864 SNPs with $< 1\times10^{-4}$, 8,674 SNPS with $< 1\times10^{-3}$ and 90,826 SNPs with p-value $< 1\times10^{-2}$, associated with PD. Resulting genome-wide association results are shown by the Manhattan plot in Fig. 1. Top hits with a p-value $< 1\times10^{-5}$, and (OR) [0.05–20] were selected for replication in the Valencia sample. None of the retained SNPs reached the genome-wide threshold. Further, as single SNP analysis results could be misleading, we plotted genotypes 500Kb around the peak together with along additional annotation from the GWAS catalogue, recombination rates, LD measures with genotyped or imputed SNPs in the region, and functional annotation for each SNP. After visualization, eight regions were retained (Supplementary Fig. 2). Observed genotype was preferentially retained when imputed signals were also present; three derived from *in silico* genotyping (imputation). We selected additional SNPs as proxies ($r^2 = 1–0.6$ on average) for individual genotyping, by using FINEMAP and LDlink tools. In addition to this selection, the genome-wide associated SNP (p-value $= 2.8 \times 10^{-8}$) at Chr18, was included in the replication step (Supplementary Fig. 2). A total of 28 SNPs in nine chromosomal regions were chosen for replication testing in the Valencia cohort. All of them were in Hardy-Weinberg equilibrium (p-value > 0.001). Results of the association analysis and minor allele effect sizes for selected SNPs are shown in Fig. 2.

3.3. *SMYD2 replicated the association in an independent sample*

Five variants out of 28 analyzed were associated with a PD in the validation cohort, overlapping with the *SMYD2*, *LARP4*, *RP11–472N19.3.* The observed variant effect size was in the same direction in both discovery and validation samples, except for the variant in *LARP*4 (Fig. 3). Variants in *SMYD2* and *RP11-472N19.3* were statistically significant. *SMYD2* carry two variants, chr1:214502898-rs4655246 and chr1:214503489-rs2291830, associated with a poor outcome in both cohorts. Minor alleles at two positions (p-value $= 1.9 \times 10^{-3}$, freq. = 0.31 and p-value $= 1.0 \times 10^{-6}$, freq. = 0.41 for BREC; p-value = 0.016, freq. = 0.32 and p-value = 0.038, freq. = 0.42 for validation sample) were associated with PD in BREC and validation cohort. The rs4655246-C allele variant, and the rs2291830-T allele variant showed a strong effect towards progressing disease; OR = 3.33 and OR = 1.47 for C-allele, and OR = 7.02 and OR = 1.26 for T-allele, for BREC and the validation cohort respectively. In *RP11–472N19.3*, two variants, chr14:92726738-rs7142050 and chr14:92726813-rs4904853, show a protective effect (i.e. favoring non progressing disease) by the minor allele for both cohorts (p-value $= 6.9 \times 10^{-3}$, freq. = 0.3 and p-value = 1.4×10^{-4}, freq. = 0.22 for BREC; p-value = 0.045, freq. = 0.34 and p-value = 0.035, freq. = 0.23 for validation sample); for the rs7142050-A allele variant the effect size was OR = 0.39 and OR = 0.81, and for rs4904853-C, OR = 0.23 and OR = 0.75, for BREC and the validation cohort respectively.

Table 2

Results from survival analysis for overall survival (OS) and time to progression (TTP) of significant variants in the validation sample.

	Gene	Variant	HR (95% CI)	p-value
(OS)	*SMYD2*	chr1:214481630-rs6665343-A	1.370 (1.050, 1.788	0.020
		chr1:214495703-rs11120295-T	1.368 (1.047, 1.787)	0.022
		chr1:214503489-rs2291830-T	1.289 (1.017, 1.633)	0.036
(TTP)		chr1:214495703-rs11120295-T	1.331 (1.020, 1.737)	0.035

Variant, chromosome position in GRch37/hg19, rs identifier and the allele effect; HR (95% CI), hazard ratio; 95% confidence interval of the hazard ratio; p-value of the variant calculated from the Cox regression model with gender, age, ECOG, progression disease and stage as covariates.

3.4. *SMYD2 variants have an impact on survival endpoints in the validation cohort*

Survival analysis was assessed in the replicated regions, and only reach statistical significance for *SMYD2*. Impact on survival outcomes was analyzed for overall changes in survival (OS) as well as in time to progression (TTP). We then stratified survival analysis by outcome (i.e. disease progression) to test the impact on other aspects of survival outcomes. Median OS and TTP was lower in the PD patients from those with response and stable disease (non PD); OS = 6 months (CI95% = [5.1,7.1]) and 13.4 (CI95% = [12.1,15.7]) and TTP = 2.8 months (CI95% = [2.6,3.1]) and 7.9 months (CI95% = [7.5,8.4]). Summary results for *SMYD2* variants are presented in Table 2.

In *SMYD2*, three analyzed variants (rs6665343 G/A, rs11120295 C/T, rs2291830 T/G) were associated with a reduced survival time. OS was shorter for the rs6665343-A, rs11120295-T, rs2291830-T allele carriers, showing a dominant effect. Allele variant rs6665343-A carriers had a shorter OS; 12.8 to 9.7 months (p-value = 0.020, HR = 1.370 95% (1.047–1.787)), allele variant rs11120295-T shows similar reduction of OS (12.5 to 9.7 months, p-value = 0.022, HR = 1.368 95% (1.047–1.787)), and rs2291830-T shows the lower effect, with a slight reduction (10.4 to 9.8 months, p-value = 0.036, HR = 1.289 95%

Gene	CHR:BP−rs−Allele	Freq.	P−val	OR	CI95% OR
SMYD2	1:214502898−rs4655246−C	0.31	1.9×10^{-3}	3.30	[1.52−7.15]
		0.32	0.016	1.47	[1.16−1.8]
	1:214503489−rs2291830−T	0.41	1×10^{-6}	7.02	[2.88−17.12]
		0.44	0.038	1.26	[1.03−1.53]
LARP4	12:50824232−rs11612002−G	0.20	3.8×10^{-6}	14.42	[3.55−58.58]
		0.19	0.036	0.76	[0.58−0.95]
RP11−472N19.3	14:92726738−rs7142050−A	0.30	6.9×10^{-3}	0.39	[0.2−0.76]
		0.34	0.045	0.81	[0.67−0.99]
	14:92726813−rs4904853−C	0.22	1.4×10^{-4}	0.23	[0.11−0.5]
		0.23	0.035	0.75	[0.57−0.95]

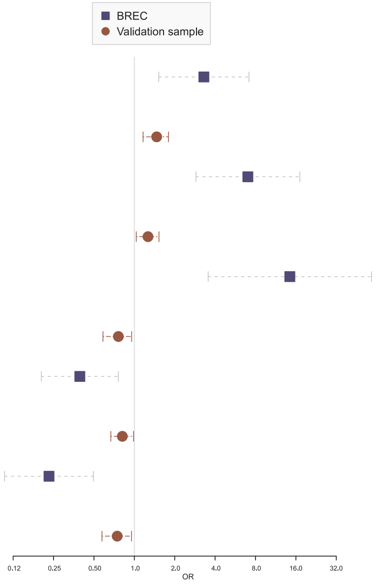


Fig. 3. Forest plot diagram of the replicated variants in the discovery and validation sample. Variants in *SMYD2* (chr1:214502898; chr1: 214503489), *RP11–472N19.3* (chr14:92726738; chr1492726813) show the same effect direction, but it is discordant in LARP4.

(1.017–1.633). When considered survival, only rs11120295-T allele was associated with shorter TTP with a dominant effect for the common allele (freq. = 0.53) with a reduction in 1.6 months (from 6.9 months to 5.3 months, p-value = 0.035, HR = 1.331 95% (1.020–1.737)) (Fig. 4). In the BREC cohort, individual survival data was not available but in concordance, rs11120295-T allele was associated with a PD outcome (OR = 0.25, CI95% = [0.12, 0.52], p-value = 1.3×10^{-4}).

3.5. Pathway analysis

From the filtered raw association signals shown in Fig. 1, we performed the functional characterization of the 889 selected genes overlapping with genome scan signals with a nominal p-value < 1×10^{-4}. Nine GO pathways were significantly enriched with the overlapping genes (Table 3). The sequence-specific DNA binding pathway (GO: 0043565) (OR = 5.32, p value = 0.0050) with nominal values overlapping *ATF1*, *PAX7*, *TBX3*, *IRX5*, *IRX3* and *CERS5,* and the cAMP-mediated signaling pathway (GO:0019933) (OR = 13.60, p value = 0.0054) highlighted by the *ADM*, *EIF4, EBP2*, *PDE4D*, *RAPGEF2*, *PCLO* genes were the most significant ones. None of the Reactome pathways reached significant level after multiple-testing correction.

4. Discussion

To better understand the germline genetic factors modulating disease progression in advanced NSCLC with first-line platinum-based treatment we performed a genome wide analysis in a two-stage approach, including two independent populations with the same ethnic ancestry. Our results provide evidence for implication in disease progression and overall survival of germline genetic variants in *SMYD2.*

In our study, the *SMYD2* variant chr1:214503489-rs2291830 T/G*,* is associated with poor clinical outcome for treated patients. The effect size observed for the rs2291830-T allele is the highest *SMYD2* signal observed in our study; OR = 7.02, CI 95% = [2.88 − 17.12]. Furthermore, survival analysis shows that rs2291830-T carriers have a reduction in the survival time (10.4 to 9.8 months, p-value = 0.036) in the validation cohort. *SMYD2* (SET and MYND domain containing 2) encode for one of the SMYD methyltransferase family proteins (SMYD1–5) [18], some of which have already been reported as candidate targets for anticancer drugs [48]. *SMYD2* is overexpressed in multiple cancer cells [10], and in addition to histones, methylates other protein substrates, including RB1 and p53, leading to loss of its tumor suppressive function [23]. There are also interesting observations, showing that depletion of SMYD2 is linked to cancer chemotherapy improvement, through the reduction of PARP1 activity, which is involved in DNA repair, chromatin modification, transcriptional regulation and genomic stability [40]. Concordantly, genetic variants in *PARP1* have been associated to a better response to platinum-based chemotherapy in NSCLC [46]. Furthermore, a prognostic value has been proposed for this protein, but there is contradictory data on functionality, while SMYD2 overexpression has been reported as a bad prognostic factor in leukemia, esophageal squamous cell carcinoma and gastric carcinoma, low expression levels in renal tumors have been associated with worse disease-specific survival and disease-free survival [41]. Supporting the role in the carcinogenic process, Nakamura's Group recently reported SMYD2-mediated ALK methylation as a new mechanisms regulating cell growth in NSCLC ALK-fused gene cell lines [52].

The other SMYD2 variants in close LD (rs6665343, rs4655246, rs11120295, rs2291830, $r^2 > 0.60$) were concordant with the observed SMYD2 association (Fig. 2), however, none of the variants had any clinical significance. No variation effect on protein function was observed using SIFT and Polyphen analysis. All variants were intronic. Expression quantitative trait loci (eQTL) analysis was performed, a significant cis-eQTL, on SMYD2 expression for rs2291830-T allele *(pvalue = 7.30 x10*$^{-7}$*, eQTL effect size (es) = −0.31), as well for* rs6665343-A, rs4655246-C and rs11120295-T alleles. *(pvalue = 3.3 ×* 10^{-5}*, es-0.16, pvalue = 3.7 pvalue = 3,7 ×* 10^{6}*, es = 0.17, and pvalue = 4.3 ×* 10^{-5}*, es = −0.17)* was present in transformed samples (fibroblasts) on the GTEx database (Release V6p (dbGaP Accession phs000424.v6.p1), and non-transformed samples (peripheral blood cells) (*pvalue* = 3.4 10^{-6}, *pvalue* = 9.2 × 10^{-11}, *pvalue* = 2.11 × 10^{-9}, *pvalue* = 2.6 × 10^{-11}) from Westra et al. [53] but not in lung tissues. However, a trans-eQTL, was observed when consider lung tissue samples on KCNK2 *(potassium two pore domain channel subfamily K member 2)* expression; rs6665343-A, rs11120295-T, and rs2291830-T alleles were correlated with a higher expression of *KCNK2 (pvalue = 1.1 ×* 10^{-2}*, es = 0.18)*. KCNK2 belongs to the two-pore-domain background potassium channel protein family, and interestingly overexpression of the channel protein, in prostate cancer, has been

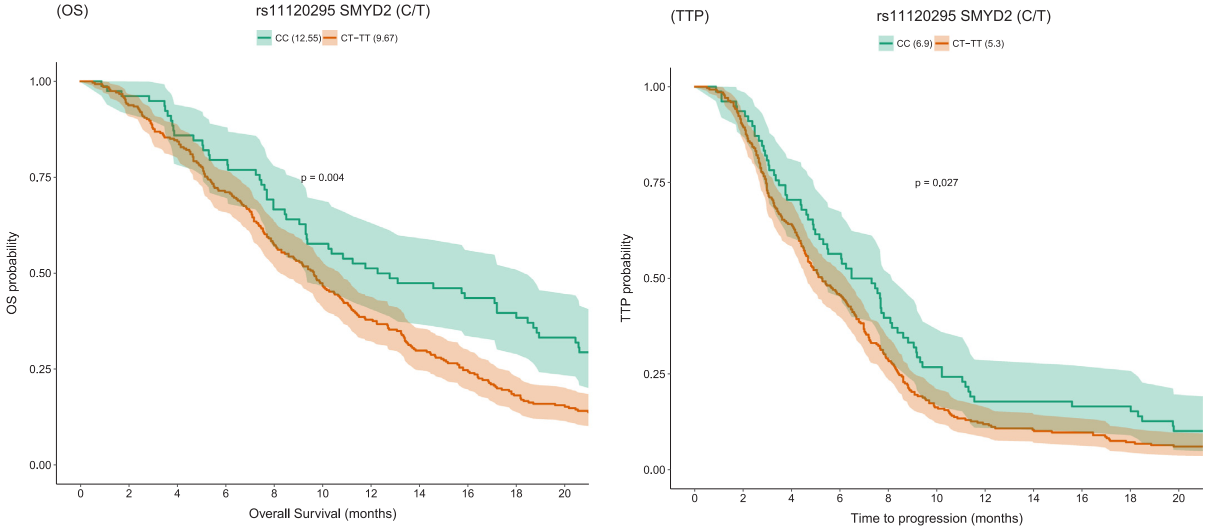


Fig. 4. Kaplan Meier plot for the validation sample: overall survival (OS) and time to progression (TTP) of patients with risk (CT-TT) and non-risk genotypes (CC) for the rs11120295 SMYD2 variant.

Table 3
Summary of the pathway enrichment analysis results in the discovery sample.

Method	GWAs	GO:ID	Description	Corrected p-value	OR	Intersect count	GO count	Intersect genes
PANTHER	−4	GO:0019864	IgG binding	0.016	20.46	5	12	*FCGR2A FCGR3B FCGR2C FCGR2B FCGR3A*
PANTHER	−4	GO:0060986	Endocrine hormone secretion	0.032	22.32	5	11	*GATA3 CGA GHRL TBX3* ***FZD4***
seq2pathway	−4	GO:0019933	cAMP-mediated signaling	0.005	13.60	5	22	*ADM EIF4* ***EBP2 PDE4D*** *RAPGEF2 PCLO*
seq2pathway	−4	GO:0010595	Positive Regulation Of Endothelial Cell Migration	0.026	7.70	5	35	*AGT ANGPT1 GATA3* ***PROX1*** *NRP1*
seq2pathway	−5	GO:0006351	Transcription, DNA-templated	0.036	2.41	10	1766	***PTPN14*** *TBX3 IRX5 IRX3 SALL3 ESF1* ***SMYD2 EBF2*** *ING5 TLE3*
seq2pathway	−5	GO:0045893	Positive regulation of transcription, DNA-templated	0.03	4.15	5	487	***PROX1*** *TBX3* ***TASP1 EBF2*** *ING5*
seq2pathway	−4	GO:0016055	Wnt signaling pathway	0.03	3.35	10	150	*HHEX PITX2 TLE3 TLE4* ***FZD4*** *PYGO1 WWOX CXXC4 NKD2 RSPO2*
seq2pathway	−5	GO:0003700	Sequence-specific DNA binding transcription factor activity	0.018	3.11	7	990	***ATF1*** *PAX7 TBX3 IRX5 IRX3 CERS5* ***PROX1***
seq2pathway	−5	GO:0043565	Sequence-specific DNA binding	0.005	5.32	6	500	***ATF1*** *PAX7 TBX3 IRX5 IRX3 CERS5*

Methods, PANTHER and seq. 2pathway overrepresentation methods; GWAS, p-value below 10^{-4} and 10^{-5} threshold for SNP inclusión; Corrected p-values on seq. 2pathway correspond to FDR while corrected p-values on PANTHER overrepresentation tests are adjusted with Bonferroni correction.

associated with a reduced survival, while knockdown inhibits cell proliferation in vivo [56].

In order to identify possible functional haplotypes, we estimated genewide haplotype structure of *SMYD2*, and haploblocks were inferred with the CI method as implemented in Haploview. All four variants were in the same block, the largest conserved block in the 3′-terminal region, but interestingly rs4655246-C/G differentiate two different haplotypes; ATCT (freq = 0.315) and ATGT (freq = 0.093), suggesting a functional role for ATCG / ATCT haplotype carriers.

Genes frequently methylated in lung cancer cells and associated with oncogenic growth of cancer cells could be targets of SMYD2, which is over-expressed in most cancer types. All of the validated methylated substrates of SMYD2 are implicated in stress responses and cellular checkpoints, it is possible that overexpression and dysregulated methylation activity could lead to compromised chemotherapy response and reduced overall survival [12,44]. Nowadays, 20 published non-histone proteins have been reported as validated targets of SMYD2 [1]. In concrete, some authors have reported that SMYD2-methylation mediated of RB1, HSP90, PTEN, PARP1 has a critical roles in tumorigenesis [10,19,36,40], and confirm, as a possible common mechanism for SMYD2 cancer progression, a SMYD2-mediated methylation causing the nuclear translocation of b-catenin and activation of Wnt/b catenin signaling pathway [12], a hallmark of a large proportion of human cancers. A higher methylation activity leads to an increased nuclear translocation activity for b-catenin, then to a high activation of the Wnt/b-catenin pathway and cancer cell progression. However lower activity could produce the contrary effect, leading to cancer cell death apoptosis, hence a higher resistance to the cisplatine action.

Identified genetic polymorphism show neighborhood enrichments of chromatin functional annotations in rs4655246 with enhancer and promoter functions (i.e., 11_TxEnh3, H3K4me3_Pro, H3K27ac_Enh) (Roadmap Epigenomics Consortium, 2015). Even out of the promoter regions, this could suggest a cryptic promoter region modulating the expression of alternative regulatory transcripts, but to date only one alternative transcript has been described in placenta tissues.

We do not have any available data for somatic mutations and methylation in cancer cells of those patients, and further studies will be needed to clarify the significance of SMYD2 polymorphisms.

Another interesting finding from our study is *RP11–472N19.3.*, a long non-coding RNA (lncRNA) locus. LncRNAs are normally found as endogenous cellular RNAs, larger than 200nt, and lacking an open reading frame of significant length. They are functional RNA elements, expressed at low levels in a tissue-specific and time-restricted manner. *RP11–472N19.3.* is transcribed in several tissues, including lung, but to date no phenotype, functional annotation or eQTL have been reported in this locus. The uncommon rs7142050-C allele was associated to a better prognosis, suggesting RP11–472N19.3 as a possible new candidate therapeutic target for lung cancer treatment. Based on several evidences (score 2b RegulomeDB, Version 1.1.) the variant rs7142050, is likely to affect binding of several transcription factors such as IRF4 (*Interferon Regulatory Factor 4*), SPI1 (*Spi-1* Proto-Oncogene), and ATF2 (*Activating transcription factor 2*). ATF2 is a transcription factor involved in stress and DNA damage which has been recently involved in cisplatin resistance in non-small cell lung cancer. LncRNAs are regarded with increasing interest as new targets for cancer therapy. Dysregulation of lncRNA expression has been implicated in lung cancer etiology, oncogenic or tumor suppressive. Zhou et al., proposed a eight-lncRNA signature as an effective independent prognostic molecular biomarker in the prediction of NSCLC patient survival [57]. Recent studies, using RNAi experiments to inhibit *HOTAIR* (Hot Transcription Antisense RNA), have reported a decreased migration, invasion and metastasis in NSCLC cells along with reduced expression of genes involving and antisense RNA inhibitory process. Similar results were reported for *MALAT1 (Metastasis Associated Lung Adenocarcinoma Transcript 1)* in mouse lung cancer models [14].

In addition to the single analysis, we performed a pathway enrichment analysis to analyze all excluded signals (pvalue $> 1\times10^{-5}$) from the replication phase. With this analysis we highlighted several pathways involved in differential clinical outcome. Some of the identified signals were in primary retained regions with a suggestive profile but were discarded prior to the replication phase (*PAX7*, *IRX5*, or *ATF1).* *SMYD2* has been identified in the pathway enrichment analysis belonging to one of the statistically significant overrepresented pathways; GO:0006351 (OR = 2.4, p-value = 0.036), a wide functional category that includes transcription regulator activity genes. Furthermore, it is interesting to note two of the enriched pathways. The cAMP-mediated signaling pathway (GO:0019933) is the second most significantly associated term (OR = 13.60, p value = 0.0054), with 5 genes out 22 associated to clinical outcome (*ADM*, *EIF4, EBP2*, *PDE4D*, *RAPGEF2*, *PCLO).* Among them, EBP2 (EBNA1-binding protein (homolog)) and PDE4D (Phosphodiesterase 4) are relevant as therapeutic targets for lung cancer therapy. EBP2 has been reported as a novel binding partner of c-Myc, regulating the function of nucleolar c-Myc, cell proliferation and tumorigenesis [28], and PDE4D has been reported as a promoter of proliferation and angiogenesis of lung cancer [42]. Moreover, the Wnt signaling pathway (GO:0016055) was overrepresented, with 10 out 150 genes (*HHEX, PITX2, TLE3, TLE4, FZD4, PYGO1, WWOX, CXXC4, NKD2, RSPO2H*)(OR = 3.35 p value = 0.03). In NSCLC it has been reported that Wnt ligand and Fzd are overexpressed and that Wnt antagonists are downregulated [37]. The same authors suggest that elevation of the β-catenin pathway is a common mechanism for conferring resistance to cancer treatment, not only to EGFR tyrosine kinase inhibitors (TKIs), but also to other types of treatment, including chemotherapy and radiotherapy. In NSCLC, a study reported inherited genetic variation in the Wnt signaling pathway contributing to variable clinical outcomes for patients with early-stage disease [8]. The involvement in NSCLC, but in different stage could indicate a common mechanism related to resistance in both phases of the disease.

In the last years, genetic analysis of somatic variation has yielded valuable profiles for lung cancer subtype classification and prediction of response to treatment [4,24]. Individual germline genetic configuration could help to improve disease management and guide treatment choice decisions. GWAS has been used successfully to identify susceptibility genes to lung cancer, has also been used to identify prognostic and predictive biomarkers to response in early [50,55] or advanced NSCLC patients [22,27,45,54], as well as to analyze adverse effects of drug treatment [6,7,49]. Any of the genes uncovered in our study has been previously reported in advanced NSCLC patients treated with chemotherapy. Most of the reported GWAS (GWAS catalog) are from Asian ancestry populations (11 out 12), and, until now, only one study is from European ancestry patients [54]. Other study using mixed ancestry data come from a different approach using cell lines in the discovery phase. As seen for susceptibility factors, ethnic differences could account for these inconsistencies.

Here, using a genomewide screening approach, we have identified a gene with potential clinical value in advanced NSCLC patients treated with chemotherapy. It is noteworthy that our approach takes advantage of massive variation information collected in deep sequencing derived public panels to empower the study. Identified SMYD2 variant have been genotyped by imputation, and inferred genotypes predicted by IMPUTE2 info shown a highly concordance (average for all inferred variants 96.9%) with genotyping. As widely reported elsewhere, these results corroborate the power of SNP imputation using sequencing derived panels for improving genome scanning results.

We identified 20 regions in the exploratory sample, and even if those signals did not reach genomewide significance, we have replicated one region in an independent sample. All signals remained significant in the joint analysis, however, heterogeneity analysis for replicated variants precluded any joint meta-analysis interpretation (median, mean $I^2 = 92.2\%$, 84.9%). We can discard a genetic bias from different ethnicity since both cohorts are from the same wide-geographical area (Spain), and share the same ancestry; or from genotyping platform, or imputation, since a high correlation was observed in our study among imputation panels. But slight differences were present regarding stage, histological type, and ECOG status that could account for these heterogeneous values. Even if clinical regimens are standardized we cannot underscore the effect of these differences between cohorts. Moreover, the treatment choices in both cohorts were slightly different, and therefore even if we account for these differences in the analysis, in the BREC cohort, we cannot overcome if present the distinct effect of cisplatin and the other dual combination chemotherapies (cisplatin-gemcitabine, cisplatin-docetaxel) in the genetic variant effects. Cisplatin enters cells via multiple pathways, and forms DNA-platinum adducts initiating a cellular self-defense system resulting in cancer cell destruction. Since resistance is supposed to be pleiotropic, these differences do not invalidate the identified signals. In the same way, a pleiotropy of alterations could be related to natural or acquired resistance [16].

Data dimensionality in genome wide analyses is a major concern when applied to clinical cohort series, generally composed by a small number of patients. In order to increase the robustness of the results, our study only considered signals with a reasonable effect in a two-stage design. The large effect size observed for *SMYD2* alleles in the BREC cohort should be considered with caution, since overestimation of the initial effect size could be present. In addition, we cannot discard that other genomic mutations, further than EGFR mutations, could be confounding the results.

5. Conclusion

In conclusion, our study identified germline genetic variation in *SMYD2* associated to bad clinical outcome (PD) in first-line platinum-based treatment in advanced NSCLC patients. These results support the biological significance of methylation process in human carcinogenesis, and open up new drug targeting possibilities and patient stratification in lung cancer therapy based on germline profiling. *SMYD2* profiling could represent an additional prognostic biomarker to better tailor multidisciplinary treatment of patients.

Clinical practice points

- What is already known about this subject?
  Tumor genomic profiling of advanced NSCLC patients determines an increase in the overall survival rates when matched therapies are provided compared with cytotoxic chemotherapy
  In advanced NSCLC patients under first-line cytotoxic chemotherapy, tumor profiling is always a tardy option.
  Furthermore, repeat tissue biopsies should be avoided and sometimes genomic profiling is precluded due to exhausted sample.
  Alternative, germline variants are identified as a valuable prognostic marker in those patients (e.g. DNA-repair genes, *CTNNB1* or *CMKLR1*).
- What are the new findings?
  In this article, we have show that genetic variation on SMYD2 is a biomarker for a bad outcome and reduced overall survival of advanced NSCLC patients when risk alleles are carried at germinal level.
  Multivariate survival analysis showed that genetic variants were independent prognostic factors.
  We report evidences of SMYD2 genetic variation impact on its own expression, and support the biological significance of methylation process of SMYD2 in human carcinogenesis.
- How might it impact on clinical practice in foreseeable future?

Evidences for SMYD2 genetic variation lead to new drug targeting possibilities.

SMYD2 alleles could be used as a biomarker for patient stratification in lung cancer therapy prior to tumor genomic profiling.

Acknowledgement

This study makes use of data generated by the GCAT |Genomes for Life Project. Cohort study of the Genomes of Catalonia", PMPPC-IGTP. *Obón-Santacana et al 2018 (BMJ Open2018;0:e018324. doi:10.1136/bmjopen-2017-018324). A full list of the investigators who contributed to the generation of the data is available from www.genomesforlife.com.* We are grateful to F. J. Gómez-Santonja for giving help on the statistical analysis. The authors also thank Laia Ramos and Raquel Pluvinet from Genomics and Bioinformatics Unit for genotyping support.

Authors' contributions

RdC, JY, and IGF conceived and designed the study. RdC, IGF, JY, SC, EJ, CC, AC, JLR, RR and LS contributed to the generation, collection and assembly. RdC, MG, JMM, XD, SC and IGF contributed to the analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis). RdC, JY, LS, JMM, MP, TK, SC, EJ and IGF contributed to writing, review, and/or revision of the manuscript. All authors approved the final version of the manuscript.

Funding

This study was funded, by the Ministerio de Economía y Competitividad (ADE10/00026) and Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR) (SGR 1269). R de Cid is granted by the Ramón y Cajal (RYC) Program (RYC-2011–07822). IGTP is part of the CERCA Program / Generalitat de Catalunya.

Conflicts of interest

None.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at https://dx.doi.org/10.1016/j.ctarc.2018.02.003.

References

[1] H. Ahmed, S. Duan, C.H. Arrowsmith, D. Barsyte-Lovejoy, M. Schapira, An integrative proteomic approach identifies novel cellular SMYD2 substrates, J Proteome Res 15 (2016) 2052–2059.

[2] A. Auton, L.D. Brooks, R.M. Durbin, E.P. Garrison, H.M. Kang, J.O. Korbel, J.L. Marchini, S. McCarthy, G.A. McVean, G.R. Abecasis, A global reference for human genetic variation, Nature 526 (2015) 68–74.

[3] C. Benner, C.C. Spencer, A.S. Havulinna, V. Salomaa, S. Ripatti, M. Pirinen, FINEMAP: efficient variable selection using summary data from genome-wide association studies, Bioinformatics 32 (2016) 1493–1501.

[4] L. Bonanno, C. Costa, M. Majem, J.J. Sanchez, A. Gimenez-Capitan, I. Rodriguez, A. Vergnenegre, B. Massuti, A. Favaretto, M. Rugge, et al., The predictive value of 53BP1 and BRCA1 mRNA expression in advanced non-small-cell lung cancer patients treated with first-line platinum-based chemotherapy, Oncotarget 4 (2013) 1572–1581.

[6] S. Cao, C. Wang, H. Ma, R. Yin, M. Zhu, W. Shen, J. Dai, Y. Shu, L. Xu, Z. Hu, H. Shen, Genome-wide association study on platinum-induced hepatotoxicity in non-small cell lung cancer patients, Sci Rep 5 (2015) 11556.

[7] S. Cao, S. Wang, H. Ma, S. Tang, C. Sun, J. Dai, C. Wang, Y. Shu, L. Xu, R. Yin, et al., Genome-wide association study of myelosuppression in non-small-cell lung cancer patients with platinum-based chemotherapy, Pharmacogenomics J 16 (2016) 41–46.

[8] A. Coscio, D.W. Chang, J.A. Roth, Y. Ye, J. Gu, P. Yang, X. Wu, Genetic variants of the Wnt signaling pathway as predictors of recurrence and survival in early-stage non-small cell lung cancer patients, Carcinogenesis 35 (2014) 1284–1291.

[9] C.C. Chang, C.C. Chow, L.C. Tellier, S. Vattikuti, S.M. Purcell, J.J. Lee, Second-generation PLINK: rising to the challenge of larger and richer datasets, Gigascience 4 (2015) 7.

[10] H.S. Cho, S. Hayami, G. Toyokawa, K. Maejima, Y. Yamane, T. Suzuki, N. Dohmae, M. Kogure, D. Kang, D.E. Neal, et al., RB1 methylation by SMYD2 enhances cell cycle progression through an increase of RB1 phosphorylation, Neoplasia 14 (2012) 476–486.

[11] O. Delaneau, B. Howie, A.J. Cox, J.F. Zagury, J. Marchini, Haplotype estimation using sequencing reads, Am J Hum Genet 93 (2013) 687–696.

[12] X. Deng, R. Hamamoto, T. Vougiouklakis, R. Wang, Y. Yoshioka, T. Suzuki, N. Dohmae, Y. Matsuo, J.H. Park, Y. Nakamura, Critical roles of SMYD2-mediated beta-catenin methylation for nuclear translocation and activation of Wnt signaling, Oncotarget 8 (2017) 55837–55847.

[13] F. Dudbridge, A. Gusnanto, Estimation of significance thresholds for genomewide association scans, Genet Epidemiol 32 (2008) 227–234.

[14] M. Eissmann, T. Gutschner, M. Hammerle, S. Gunther, M. Caudron-Herger, M. Gross, P. Schirmacher, K. Rippe, T. Braun, M. Zornig, S. Diederichs, Loss of the abundant nuclear non-coding RNA MALAT1 is compatible with life and development, RNA Biol 9 (2012) 1076–1087.

[15] J. Ferlay, I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D.M. Parkin, D. Forman, F. Bray, Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012, Int J Cancer 136 (2015) E359–E386.

[16] L. Galluzzi, L. Senovilla, I. Vitale, J. Michels, I. Martins, O. Kepp, M. Castedo, G. Kroemer, Molecular mechanisms of cisplatin resistance, Oncogene 31 (2012) 1869–1883.

[17] J. Goffin, C. Lacchetti, P.M. Ellis, Y.C. Ung, W.K. Evans, First-line systemic chemotherapy in the treatment of advanced non-small cell lung cancer: a systematic review, J Thorac Oncol 5 (2010) 260–274.

[18] P.D. Gottlieb, S.A. Pierce, R.J. Sims, H. Yamagishi, E.K. Weihe, J.V. Harriss, S.D. Maika, W.A. Kuziel, H.L. King, E.N. Olson, et al., Bop encodes a muscle-restricted protein containing MYND and SET domains and is essential for cardiac differentiation and morphogenesis, Nat Genet 31 (2002) 25–32.

[19] R. Hamamoto, G. Toyokawa, M. Nakakido, K. Ueda, Y. Nakamura, SMYD2-dependent HSP90 methylation promotes cancer cell proliferation by regulating the chaperone complex formation, Cancer Lett 351 (2014) 126–133.

[20] M.A. Hildebrandt, J. Gu, X. Wu, Pharmacogenomics of platinum-based chemotherapy in NSCLC, Expert Opin Drug Metab Toxicol 5 (2009) 745–755.

[21] B.N. Howie, P. Donnelly, J. Marchini, A flexible and accurate genotype imputation method for the next generation of genome-wide association studies, PLoS Genet 5 (2009) e1000529.

[22] L. Hu, C. Wu, X. Zhao, R. Heist, L. Su, Y. Zhao, B. Han, S. Cao, M. Chu, J. Dai, et al., Genome-wide association study of prognosis in advanced non-small cell lung cancer patients receiving platinum-based chemotherapy, Clin Cancer Res 18 (2012) 5507–5514.

[23] J. Huang, L. Perez-Burgos, B.J. Placek, R. Sengupta, M. Richter, J.A. Dorsey, S. Kubicek, S. Opravil, T. Jenuwein, S.L. Berger, Repression of p53 activity by Smyd2-mediated methylation, Nature 444 (2006) 629–632.

[24] E. Jantus-Lewintre, E. Sanmartin, R. Sirera, A. Blasco, J.J. Sanchez, M. Taron, R. Rosell, C. Camps, Combined VEGF-A and VEGFR-2 concentrations in plasma: diagnostic and prognostic implications in patients with advanced NSCLC, Lung Cancer 74 (2011) 326–331.

[25] E. Jantus-Lewintre, R. Sirera, A. Cabrera, A. Blasco, C. Caballero, V. Iranzo, R. Rosell, C. Camps, Analysis of the prognostic value of soluble epidermal growth factor receptor plasma concentration in advanced non-small-cell lung cancer patients, Clin Lung Cancer 12 (2011) 320–327.

[26] S. Lê, J. Josse, F. Husson, FactoMineR: an R Package for Muktivariate Analysis, Journal of Statistical Software 25 (2008) 1–18.

[27] Y. Lee, K.A. Yoon, J. Joo, D. Lee, K. Bae, J.Y. Han, J.S. Lee, Prognostic implications of genetic variants in advanced non-small cell lung cancer: a genome-wide association study, Carcinogenesis 34 (2013) 307–313.

[28] P. Liao, W. Wang, M. Shen, W. Pan, K. Zhang, R. Wang, T. Chen, Y. Chen, H. Chen, P. Wang, A positive feedback loop between EBP2 and c-Myc regulates rDNA transcription, cell proliferation, and tumorigenesis, Cell Death Dis 5 (2014) e1032.

[30] M.J. Machiela, S.J. Chanock, LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants, Bioinformatics 31 (2015) 3555–3557.

[31] J. Marchini, B. Howie, Genotype imputation for genome-wide association studies, Nat Rev Genet 11 (2010) 499–511.

[32] A. Matakidou, R. el Galta, E.L. Webb, M.F. Rudd, H. Bridle, T. Eisen, R.S. Houlston, Genetic variation in the DNA repair genes is predictive of outcome in lung cancer, Hum Mol Genet 16 (2007) 2333–2340.

[33] W. McLaren, L. Gil, S.E. Hunt, H.S. Riat, G.R. Ritchie, A. Thormann, P. Flicek, F. Cunningham, The Ensembl Variant Effect Predictor, Genome Biol 17 (2016) 122.

[34] H. Mi, A. Muruganujan, J.T. Casagrande, P.D. Thomas, Large-scale gene function analysis with the PANTHER classification system, Nat Protoc 8 (2013) 1551–1566.

[35] T. Moran, J. Wei, M. Cobo, X. Qian, M. Domine, Z. Zou, I. Bover, L. Wang, M. Provencio, L. Yu, et al., Two biomarker-directed randomized trials in European and Chinese patients with nonsmall-cell lung cancer: the BRCA1-RAP80 Expression Customization (BREC) studies, Ann Oncol 25 (2014) 2147–2155.

[36] M. Nakakido, Z. Deng, T. Suzuki, N. Dohmae, Y. Nakamura, R. Hamamoto, Dysregulation of AKT pathway by SMYD2-mediated lysine methylation on PTEN, Neoplasia 17 (2015) 367–373.

[37] A. Nakata, R. Yoshida, R. Yamaguchi, M. Yamauchi, Y. Tamada, A. Fujita, T. Shimamura, S. Imoto, T. Higuchi, M. Nomura, et al., Elevated β-catenin pathway as a novel target for patients with resistance to EGF receptor targeting drugs, Scientific Reports 5 (2015) 13076.

[40] L. Piao, D. Kang, T. Suzuki, A. Masuda, N. Dohmae, Y. Nakamura, R. Hamamoto, The histone methyltransferase SMYD2 methylates PARP1 and promotes poly(ADP-ribosyl)ation activity in cancer cells, Neoplasia 16 (2014) 257–264 (264 e252).

[41] A.S. Pires-Luis, M. Vieira-Coimbra, F.Q. Vieira, P. Costa-Pinheiro, R. Silva-Santos, P.C. Dias, L. Antunes, F. Lobo, J. Oliveira, C.S. Goncalves, et al., Expression of histone methyltransferases as novel biomarkers for renal cell tumor diagnosis and prognostication, Epigenetics 10 (2015) 1033–1043.

[42] S.S. Pullamsetti, G.A. Banat, A. Schmall, M. Szibor, D. Pomagruk, J. Hanze, E. Kolosionek, J. Wilhelm, T. Braun, F. Grimminger, et al., Phosphodiesterase-4 promotes proliferation and angiogenesis of lung cancer by crosstalk with HIF, Oncogene 32 (2013) 1121–1134.

[43] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M.A. Ferreira, D. Bender, J. Maller, P. Sklar, P.I. de Bakker, M.J. Daly, P.C. Sham, PLINK: a tool set for whole-genome association and population-based linkage analyses, Am J Hum Genet 81 (2007) 559–575.

[44] N. Reynoird, P.K. Mazur, T. Stellfeld, N.M. Flores, S.M. Lofgren, S.M. Carlson, E. Brambilla, P. Hainaut, E.B. Kaznowska, C.H. Arrowsmith, et al., Coordination of stress signals by the lysine methyltransferase SMYD2 promotes pancreatic cancer, Genes Dev 30 (2016) 772–785.

[45] Y. Sato, N. Yamamoto, H. Kunitoh, Y. Ohe, H. Minami, N.M. Laird, N. Katori, Y. Saito, S. Ohnami, H. Sakamoto, et al., Genome-wide association study on overall survival of advanced non-small cell lung cancer patients treated with carboplatin and paclitaxel, J Thorac Oncol 6 (2011) 132–138.

[46] K. Shiraishi, T. Kohno, C. Tanai, Y. Goto, A. Kuchiba, S. Yamamoto, K. Tsuta, H. Nokihara, N. Yamamoto, I. Sekine, et al., Association of DNA repair gene polymorphisms with response to platinum-based doublet chemotherapy in patients with non-small-cell lung cancer, J Clin Oncol 28 (2010) 4945–4952.

[47] R. Sirera, R.M. Bremnes, A. Cabrera, E. Jantus-Lewintre, E. Sanmartin, A. Blasco, N. Del Pozo, R. Rosell, R. Guijarro, J. Galbis, et al., Circulating DNA is a useful prognostic factor in patients with advanced non-small cell lung cancer, J Thorac Oncol 6 (2011) 286–290.

[48] N. Spellmon, J. Holcomb, L. Trescott, N. Sirinupong, Z. Yang, Structure and function of SET and MYND domain-containing proteins, Int J Mol Sci 16 (2015) 1406–1428.

[49] X.L. Tan, A.M. Moyer, B.L. Fridley, D.J. Schaid, N. Niu, A.J. Batzler, G.D. Jenkins, R.P. Abo, L. Li, J.M. Cunningham, et al., Genetic variation predicting cisplatin cytotoxicity associated with overall survival in lung cancer patients receiving

platinum-based chemotherapy, Clin Cancer Res 17 (2011) 5801–5811.

[50] S. Tang, Y. Pan, Y. Wang, L. Hu, S. Cao, M. Chu, J. Dai, Y. Shu, L. Xu, J. Chen, et al., Genome-wide association study of survival in early-stage non-small cell lung cancer, Ann Surg Oncol 22 (2015) 630–635.

[51] B. Wang, J.M. Cunningham, X.H. Yang, Seq. 2pathway: an R/Bioconductor package for pathway analysis of next-generation sequencing data, Bioinformatics 31 (2015) 3043–3045.

[52] R. Wang, X. Deng, Y. Yoshioka, T. Vougiouklakis, J.H. Park, T. Suzuki, N. Dohmae, K. Ueda, R. Hamamoto, Y. Nakamura, Effects of SMYD2-mediated EML4-ALK methylation on the signaling pathway and growth in non-small cell lung cancer cells, Cancer Sci. (2017).

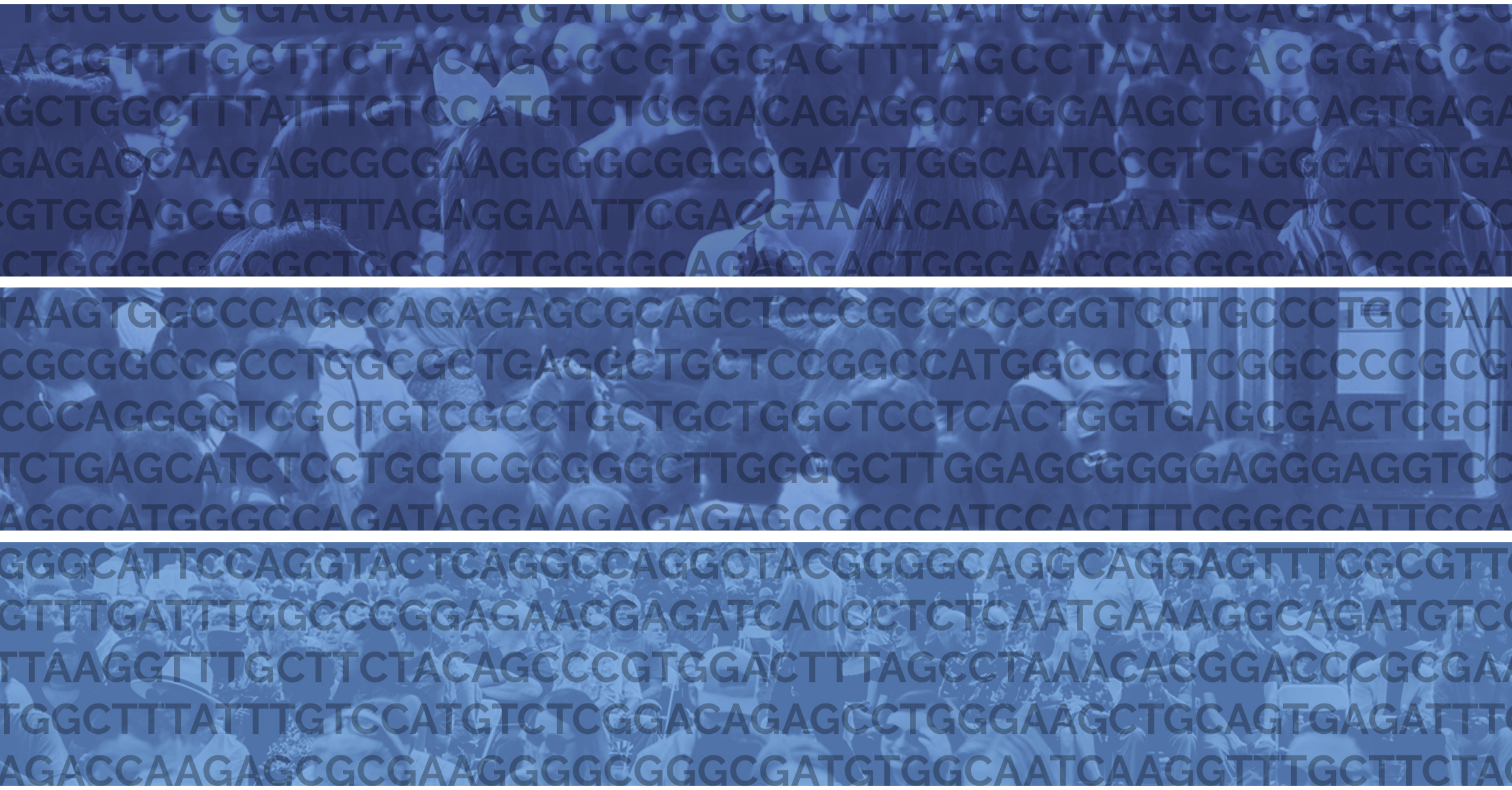
[53] H.J. Westra, M.J. Peters, T. Esko, H. Yaghootkar, C. Schurmann, J. Kettunen, M.W. Christiansen, B.P. Fairfax, K. Schramm, J.E. Powell, et al., Systematic identification of trans eQTLs as putative drivers of known disease associations, Nat Genet 45 (2013) 1238–1243.

[54] X. Wu, Y. Ye, R. Rosell, C.I. Amos, D.J. Stewart, M.A. Hildebrandt, J.A. Roth, J.D. Minna, J. Gu, J. Lin, et al., Genome-wide association study of survival in non-small cell lung cancer patients receiving platinum-based chemotherapy, J Natl Cancer Inst 103 (2011) 817–825.

[55] K.A. Yoon, M.K. Jung, D. Lee, K. Bae, J.N. Joo, G.K. Lee, H.S. Lee, J.S. Lee, Genetic variations associated with postoperative recurrence in stage I non-small cell lung cancer, Clin Cancer Res 20 (2014) 3272–3279.

[56] G.M. Zhang, F.N. Wan, X.J. Qin, D.L. Cao, H.L. Zhang, Y. Zhu, B. Dai, G.H. Shi, D.W. Ye, Prognostic significance of the TREK-1 K2P potassium channels in prostate cancer, Oncotarget 6 (2015) 18460–18468.

[57] M. Zhou, Y. Sun, W. Xu, Z. Zhang, H. Zhao, Z. Zhong, J. Sun, Comprehensive analysis of lncRNA expression profiles reveals a novel lncRNA signature to discriminate nonequivalent outcomes in patients with ovarian cancer, Oncotarget 7 (2016) 32433–32448.



UNIVERSITAT DE
BARCELONA