REGULOME-SEQ: A NOVEL APPROACH FOR THE IDENTIFICATION OF NON-CODING VARIANTS ASSOCIATED WITH HUMAN DISEASE. ASSESSMENT OF ITS APPLICABILITY IN 89 BRUGADA SYNDROME INDIVIDUALS

Mel·lina Pinsach Abuin

DOCTORAL THESIS

Regulome-seq: a Novel Approach for the Identification of Non-coding Variants Associated with Human Disease. Assessment of its Applicability in 89 Brugada syndrome Individuals.

Mel·lina Pinsach Abuin

2018

Universitat de Girona

UC San Diego
SCHOOL OF MEDICINE

DOCTORAL THESIS

Regulome-seq: a Novel Approach for the Identification of Non-coding Variants Associated with Human Disease. Assessment of its Applicability in 89 Brugada syndrome Individuals.

Mel·lina Pinsach Abuin

2018

Programa de Doctorat en Biologia Molecular, Biomedicina i Salut

Dirigida per la Dra. Sara Pagans Lista

Codirigida pel Dr. Ivan Garcia Bassets

Codirigida pel Dr. Ramon Brugada Terradellas

Tutoritzada per la Dra. Fabiana Scornik Gerzenstein

Memòria presentada per optar al títol de doctor per la Universitat de Girona

4 Annexes

Universitat de Girona

La Dra. Sara Pagans Lista, Directora d'Investigació del Departament de Ciències Mèdiques i professora de la Facultat de Medicina de la Universitat de Girona.

El Dr. Ivan Garcia Basset, "Principal Investigator and Associate Research Scientist" del Departament de Medicina de la Universitat de Califòrnia, San Diego.

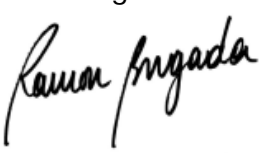
El Dr. Ramon Brugada Terradellas, professor titular de la Facultat de Medicina de la Universitat de Girona, director del Centre de Genètica Cardiovascular de l'Institut d'Investigació Biomèdica de Girona, responsable del servei de Cardiologia de l'Hospital Trueta de Girona i cardiòleg de l'Hospital Josep Trueta de Girona.

DECLAREM: Que el treball titulat **"Regulome-seq: a novel approach for the identification of non-coding variants associated with human disease. Assessment of its applicability in 89 Brugada syndrome individuals",** que presenta la Mel·lina Pinsach Abuin per a l'obtenció del títol de doctor/a, ha estat realitzat sota la nostra direcció i que compleix els requisits per poder optar a Menció Internacional. I, perquè així consti i tingui els efectes oportuns, signo aquest document.

Dra. Sara Pagans Lista

Dr. Ivan Garcia Bassets

Dr. Ramon Brugada Terradellas



Mel·lina Pinsach Abuin

Girona, 6 de Setembre de 2018

Acknowledgements

Bé doncs, ja ha arribat el moment d'escriure els agraïments. No em pensava pas que arribés mai aquest moment! Sembla estrany que després d'haver escrit tota la tesi ara no sàpiga què posar en aquest apartat...

Primer de tot m'agradaria donar les gràcies a la meva directora de tesi, la Sara, i als meus co-directors Ivan i Ramon. Gràcies Sara per haver-me donat la oportunitat de fer aquesta tesi sota la teva supervisió. Treballar amb tu és un plaer, ho fas tot molt fàcil. Ets una persona molt propera i m'has ajudat molt durant aquests anys, tant a nivell professional – per totes les oportunitats que m'has donat –, com personal.

Gràcies Ivan per iniciar (juntament amb la Sara) i "revolucionar" tot el que venia a ser el projecte Regulome-seq. A tu sobretot et vull donar les gràcies per haver-me donat la oportunitat de venir a San Diego i fer de la meva estada una experiència increïble.

Gràcies Ramon per haver-me donat la oportunitat de fer la tesi en el CGC.

En segon lloc, vull donar un seguit d'agraïments als meus companys del CGC Alexandra, Ferran, Eli S, Mireia, Mònica, Anna F, Rebecca, Èric, Adrià (tu et mereixes un extra agraïment per ajudar-me amb els clonatges i les luciferases!), David, Marta Prats, Marta Puigmulé (mira que dir-vos Marta P les dues!) i la Laura, per fer més divertides les hores al laboratori i sobretot els descansos per dinar.

També vull donar les gràcies a en Guillermo i la Fabiana pels seus comentaris i suggerències, l'Anna I per tenir paciència amb mi quan havia de revisar quins pacients triar per l'estudi, l'Òscar, en Marcel i l'Eli C sempre disposada a donar un cop de mà.

Com no, també vull mencionar en els meus agraïments tots els membres del CGC que ja no estan al laboratori però que sempre hi seran presents, l'Anna T (la meva antecessora), l'Irene (gràcies per preocupar-te per mi), la Cris, l'Helena, en Javi, l'Olallo, en Pedro i la Catarina.

Bernat i Txús, vosaltres us mereixeu un agraïment a part. Heu passat a ser membres molt importants del projecte Regulome-seq, sense vosaltres dos no haurien estat possible moltes parts de la tesi però el suport moral que m'heu donat és encara molt més gran. Tot i que vosaltres teníeu les vostres respectives tesis, hem format un petit equip, una petita família que ha treballat colze a colze per tirar endavant aquest projecte. Gràcies pel vostre esforç, pel vostre humor i les vostres bromes. Els darrers anys d'aquesta tesi han sigut molt durs per mi personalment i us vull donar les gràcies per escoltar-me quan ho necessitava i per no fallar-me mai. Ha estat un plaer construir aquest projecte amb vosaltres!

Bernat, a tu no sé què més dir-te que no et digui cada dia...has passat a ser una persona molt important per mi, que “t’apreto ben fort” i espero poder compartir molts més moments al teu costat.

Aquesta tesi ha anat creixent a mesura que ha anat avançant i ha anat recollint persones pel camí, sense la col·laboració de les quals aquest projecte no hauria estat possible. Per això considero que aquestes persones també es mereixen ser mencionades en aquest apartat d’agraïments. Gràcies Daria, Farah, Jing, Evgin, Kenny, Manu i Julia.

Sortint de l’ambient laboral, també vull donar les gràcies a les meves nenes Nuri, Anna, Tina, Mireia, Laura, Laia, Judit, Irene i Alba per haver estat tots aquests anys al meu costat, tan en els bons com en els mals moments, i per haver viscut tantes i tantes experiències juntes – i les que ens queden per venir! –. També us vull donar les gràcies per la paciència que heu tingut amb la meva presència limitada aquests darrers mesos, però no patiu que ja torno a ser aquí.

En aquest grup de persones especials hi falta una persona que ens va deixar massa aviat. Laura, aquesta tesi va dedicada a tu. Encara que no estiguis aquí presencialment, tu sempre estàs amb mi i jo sé que estàs molt orgullosa de que finalment hagi acabat aquesta tesi. Et vull donar les gràcies per tots els moments que em vas regalar però també per haver-me ensenyat tantes coses després d’haver marxat.

Finalment, m’agradaria donar les gràcies a la meva família. Aquesta tesi també és per vosaltres. Encara que no us he donat mai les gràcies per tot el que heu fet per mi, ara us dic que sou un dels pilars fonamentals de la meva vida i que sense el vostre suport no hauria tirat endavant. Gràcies mama i papa – sou els millors pares que hauria pogut tenir mai –, i a tu teta, t’estimo amb bogeria! Gràcies teta per ser la meva “personeta petita”.

Per a la meva família

Abbreviations

AF	Allele Frequency
AP	Action Potential
AVN	Atrioventricular Node
BAM	Binary Alignment Map
BrS	Brugada syndrome
BWA	Burrows-Wheeler Aligner
Ca2+	Calcium ion
CADD	Combined Annotation-Dependent Depletion
CDTS	Context-Dependent Tolerance Score
ChIP-seq	Chromatin Immunoprecipitation followed by massively parallel sequencing
dbSNP	Single Nucleotide Polymorphism database
DHS	DNase I Hypersensitivity Sites
DHS-seq	DNase I Hypersensitivity Sites followed by massively parallel sequencing
DNA	Deoxyribonucleic Acid
EKG	Electrocardiogram
ENCODE	Encyclopedia of DNA Elements
FastQC	Fastq quality control
FN	False Negative
FP	False Positive
GATK	Genome Analysis Toolkit
gnomAD	Genome Aggregation database
GRCh37/hg19	Genome Reference Consortium Human Build 37/Human Genome version 19
gVCF	Genotype Variant Call File
GWAS	Genome-Wide Association Study
HCMs	Human Cardiomyocytes
hESCs	Human Embryonic Stem Cells
hg18	Human Genome version 18
ICD	Implantable Cardioverter Defibrillator
Indel	Insertion and Deletion
iPS cells	Induced Pluripotent Stem cells
K+	Potassium ion
Na+	Sodium ion

NGS	Next Generation Sequencing
NRC	Nextera Rapid Capture
PCR	Polymerase Chain Reaction
PPV	Positive Predictive Value
RNA	Ribonucleic Acid
ROI	Reads of Interest
RVOT	Right ventricular outflow tract
SAM	Sequence aligment map
SAMtools	Sequence Alignment/Map tools
SAN	Sinoatrial Node
SCD	Sudden Cardiac Death
SD	Sudden Death
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant
SV	Structural Variant
t-SNE	T-distributed Stochastic Neighbor Embedding
TAD	Topological Associated Domain
TF	Transcription Factor
TP	True Positive
TSS	Transcription Start Site
TTS	Transcription Termination site
UTR	Untranslated Region
VCF	Variant Call File
VQSR	Variant Quality Score Recalibration
WES	Whole-exome sequencing
WGS	Whole-genome sequencing
X	Coverage

Note: in this thesis genes are cited in uppercase and italics, while the proteins are cited only in uppercase.

Index of figures

Index of tables

General index

Resum

La síndrome de Brugada (SBr) és una malaltia elèctrica cardíaca associada a una elevada susceptibilitat a arrítmies ventriculars i mort sobtada cardíaca. Variants genètiques en regions codificants del gen *SCN5A*, que codifica pel canal de sodi cardíac, expliquen un 11-24% dels casos amb SBr. De manera similar, variants genètiques en regions codificants d'altres gens com els que codifiquen per les subunitats beta reguladores del canal de sodi cardíac, canals de calci o altres proteïnes accessòries, expliquen un 5-10% dels casos amb SBr. En conjunt, les variants codificants que afecten gens de canals iònics i les seves subunitats reguladores expliquen un 25-30% dels casos amb SBr. Així doncs, existeix una elevada proporció de pacients diagnosticats amb SBr en els quals l'etiologia de la malaltia és encara desconeguda (casos amb SBr 'orfes'). Estudis recents d'associació de genoma complet (GWAS) han demostrat que moltes variants associades a malalties es troben en regions no codificants, especialment en regions reguladores en *cis.* En base a aquestes observacions, proposem que variants no codificants en regions reguladores en *cis* dels gens associats a SBr podrien ser una causa inexplorada de la malaltia. En aquesta tesi, s'ha caracteritzat la variació genètica en les regions reguladores en *cis* de sis gens associats a SBr (*SCN5A, SCN2B, SCN3B, CACNA1C, CACNB2* i *CACNA2D*) i es proposen possibles variants candidates per a futurs estudis funcionals.

Per aconseguir el nostre objectiu, s'ha desenvolupat una estratègia dirigida, anomenada Regulome-seq. Utilitzant informació sobre la topologia del genoma humà (TADs), una marca d'histona associada a promotors actius (H3K4me3) i la unió d'un regulador transcripcional (CTCF) al voltant dels gens associats a SBr en cèl·lules cardíaques, es van identificar 1.293 regions no codificants potencialment rellevants per la SBr. Mitjançant l'eina DesignStudio™, s'han dissenyat 5.546 sondes que es van utilitzar per capturar de manera selectiva les 1.293 regions en una cohort de 89 individus *SCN5A* negatius amb SBr. Després de seqüenciar massivament les regions capturades i aplicar una combinació rigorosa de filtres, es van identificar un total de 5,349 variants genètiques en les regions reguladores en *cis* dels 89 individus amb SBr. De totes les variants identificades, 4.837 corresponen a variants en un únic nucleòtid (SNVs) i 512 a insercions i delecions – indels – (212 insercions i 293 delecions). Els resultats obtinguts mostren que aproximadament el 33% de les variants són privades per cada individu amb SBr, mentre que el 67% restant estan compartides per 2 o més individus. També s'ha observat que els SNVs són més prevalents entre les variants privades, mentre que els indels són més prevalents entre les variants més compartides. L'annotació de les variants amb les bases de dades dbSNP150, 1000 Genomes i gnomAD mostra que el 6.86% de les variants

identificades són noves, i que aproximadament el 93% de les variants noves són privades.

Per tal d'identificar quines de les 5.349 variants identificades poden ser funcionalment rellevants per la SBr, s'han utilitzat dues estratègies diferents: una estratègia centrada en les variants afectant la unió de factors de transcripció (FT), i una altra centrada en la comparació de genotips amb una cohort amb envelliment saludable (Wellderly).

En primer lloc, ens vam proposar identificar aquelles variants que solapen llocs d'unió de FT cardíacs. Es van realitzar experiments d'immunoprecipitació de cromatina seguit de seqüenciació massiva (ChIP-seq) dels factors GATA4, GATA6 i NKX2.5 en cardiomiòcits derivats de cèl·lules pluripotents induïdes (iPS). No obstant, per raons tècniques no es van poder obtenir suficients pics i no es va poder assolir l'objectiu plantejat. D'altra banda, també vam sospesar la possibilitat que les variants afectant la unió de CTCF podrien alterar l'aïllament dels TADs dels gens associats a SBr i resultar en una expressió aberrant d'aquests gens. Després d'integrar la informació disponible sobre la unió de CTCF en cardiomiòcits humans amb les 5.349 variants reportades, es van identificar un total de 59 variants en 54 llocs d'unió de CTCF diferents. Mitjançant un algoritme anomenat DeepBind, es va predir que 43 de les variants disminuirien la unió de CTCF, 14 n'augmentarien la unió i 2 no tindrien cap efecte. Es va examinar també l'efecte d'aquestes variants en la unió de CTCF mitjançant assajos luciferasa. Aquests assajos van mostrar que el DeepBind és una estratègia vàlida per prioritzar variants no codificants afectant la unió de FT. La combinació de les prediccions de DeepBind amb els resultats de la luciferasa ha permès la identificació de 21 variants solapant llocs d'unió de CTCF que estarien afectant de manera més robusta la unió d'aquest factor (16 disminuint i 5 incrementant).

En segon lloc, es va comparar la variació genètica present en els 89 individus amb SBr amb 200 individus Wellderly. Aquesta comparativa, juntament amb els anàlisis estadístics aplicats, va permetre la identificació de 537 variants candidates específicament associades a pacients amb SBr (201 exclusives i 336 enriquides). L'anàlisi de les variants segons el seu efecte deleteri (CADD) i tolerància a variació genètica (CDTS) mostra que aquestes 537 variants candidates són més probablement patogèniques que altres variants, especialment quan aquestes es troben al locus del gen *SCN5A*. Aquests resultats suggereixen que, per la patogènesi de la SBr, les variants no codificants afectant l'expressió del gen *SCN5A* poden ser més importants que variants no codificants en altres gens i, a més, remarca la rellevància del gen *SCN5A* en la SBr. L'establiment d'uns llindars de patogenicitat per aquestes 537 variants candidates, essent aquests llindars la presentació d'una puntuació CADD ≥15 i trobar-se en el primer percentil de CDTS, va permetre la identificació de 10 variants candidates que poden estar relacionades amb el fenotip de SBr.

En resum, el present estudi demostra la potencial aplicació de l'estratègia Regulome-seq com a eina diagnòstica de malalties genètiques en un futur. També proposa variants candidates a la SBr per a ser avaluades experimentalment, i demostrar la seva possible associació amb el fenotip de la SBr. Els resultats obtinguts proporcionen un nou enfocament sobre les bases moleculars de les arrítmies cardíaques relacionades a alteracions en les corrents iòniques, les quals podrien explicar alguns casos de SBr 'orfes'.

Resumen

El síndrome de Brugada (SBr) es una enfermedad eléctrica cardíaca asociada a una elevada susceptibilidad a arritmias ventriculares y muerte súbita cardiaca. Variantes genéticas en las regiones codificantes del gen *SCN5A,* que codifica para el canal de sodio cardíaco, explican el 11-24% de los casos de SBr. De forma similar, variantes genéticas en regiones codificantes de otros genes como los que codifican para las subunidades beta reguladoras del canal de sodio cardiaco, los canales de calcio u otras proteínas accesorias, explican el 5-10% de los casos de SBr. En total, las variantes codificantes afectando los genes de canales iónicos así como sus subunidades reguladoras explican un 25-30% de los casos de SBr. Por lo tanto, en una elevada proporción de los pacientes diagnosticados con SBr, la etiología de la enfermedad sigue siendo desconocida (casos de SBr 'huérfanos'). Estudios recientes de asociación de genoma completo (GWAS) han demostrado que muchas variantes asociadas a enfermedades se encuentran en regiones no codificantes, particularmente en regiones reguladoras en *cis.* Partiendo de estas observaciones, proponemos que variantes no codificantes en regiones reguladoras en *cis* de genes asociados al SBr podían ser una causa inexplorada de la enfermedad. En esta tesis, se ha caracterizado la variación genética en regiones reguladoras en *cis* de seis genes asociados a SBr (*SCN5A, SCN2B, SCN3B, CACNA1C, CACNB2* y *CACNA2D*) y se proponen posibles variantes candidatas para futuros estudios funcionales.

Para cumplir con nuestro objetivo, se ha desarrollado una estrategia dirigida, llamada como Regulome-seq. Mediante el uso de información sobre la topología del genoma humano (TADs), una marca de histona asociada a promotores activos (H3K4me3) y la unión de un regulador transcripcional (CTCF) en los alrededores de genes asociados a SBr en células cardíacas, se identificaron 1.293 regiones no codificantes potencialmente relevantes para el SBr. Utilizando la herramienta DesignStudio™, se diseñaron 5.546 sondas que se usaron para capturar de forma selectiva las 1.293 regiones en una cohorte de 89 individuos *SCN5A* negativos con SBr. Después de secuenciar de manera masiva las regiones capturadas y aplicar una combinación rigurosa de filtros, se identificaron un total de 5.539 variantes genéticas en las regiones reguladoras en *cis* de los 89 individuos con SBr. De todas las variantes identificadas, 4.837 corresponden a variantes en un único nucleótido (SNVs) y 512 a inserciones y deleciones – indels – (212 inserciones y 293 deleciones). Los resultados obtenidos muestran que aproximadamente el 33% de las variantes son privadas para cada individuo con SBr, mientras que el 67% restante están compartidas por 2 o más individuos. También se observó que los SNVs son más prevalentes entre las variantes privadas, mientras que los indels son más prevalentes entre las variantes más compartidas. La anotación de las variantes con las bases

de datos dbSNP150, 1000 Genomes y gnomAD muestra que el 6.86% de las variantes identificadas son nuevas, y que aproximadamente el 93% de las variantes nuevas son privadas.

Con el fin de identificar cuáles de las 5.349 variantes identificadas pueden ser funcionalmente relevantes para el SBr, se utilizaron dos estrategias distintas: una estrategia centrada en las variantes afectando la unión de factores de transcripción (FT), y otra centrada en la comparación de genotipos con una cohorte con envejecimiento saludable (Wellderly).

En primer lugar, nos propusimos identificar aquellas variantes que solapan lugares de unión de FT cardiacos. Se realizaron experimentos de inmunoprecipitatión de cromatina seguida de secuenciación masiva (ChIP-seq) de los factores GATA4, GATA6 y NKX2.5 en cardiomiocitos derivados de células pluripotentes inducidas (iPS). Sin embargo, por razones técnicas, no se pudieron obtener suficientes picos y no se pudo alcanzar el objetivo propuesto. Por otro lado, se planteó que las variantes afectando la unión de CTCF podrían afectar el aislamiento de los TADs de los genes asociados a SBr y resultar en una expresión aberrante de dichos genes. Después de integrar la información disponible sobre la unió de CTCF en cardiomiocitos humanos con las 5.349 variantes reportadas, se identificaron un total de 59 variantes en 54 lugares de unión de CTCF distintos. Utilizando un algoritmo llamado DeepBind, se predijo que 43 de las variantes disminuirían la unión de CTCF, 14 aumentarían su unión y 2 no tendrían ningún efecto. Se examinó también el efecto de estas variantes en la unión de CTCF mediante ensayos luciferasa. Estos ensayos mostraron que el DeepBind es una estrategia válida para priorizar variantes no codificantes afectando la unión de FT. La combinación de las predicciones de DeepBind con los resultados de la luciferasa permitieron la identificación de 21 variantes que solapan lugares de unión de CTCF que estarían mas robustamente afectando la unión de este factor (16 disminuyendo y 5 incrementando).

En segundo lugar, comparamos la variación genética presente en los 89 pacientes con SBr con 200 individuos Wellderly. Esta comparación, junto con los análisis estadísticos aplicados, permitió la identificación de 537 variantes candidatas específicamente asociadas a pacientes con SBr (291 exclusivas y 336 enriquecidas). El análisis de las variantes según su efecto deletéreo (CADD) y tolerancia a variación genética (CDTS) muestra que estas 537 variantes candidatas son más probablemente patogénicas que otras variantes, especialmente cuando éstas se encuentran en el locus del gen *SCN5A*. Estos resultados sugieren que, para la patogénesis del SBr, las variantes no codificantes afectando la expresión del gen *SCN5A* pueden ser mas importantes que variantes no codificantes en otros genes y, además, realza la relevancia del gen *SCN5A* en el SBr. El establecimiento de unos umbrales de patogenicidad para estas 537 variantes candidatas, siendo estos umbrales la presentación de una puntuación

CADD ≥15 y encontrarse en el primer percentil de CDTS, permitió la identificación de 10 variantes candidatas que pueden estar relacionadas con el fenotipo de SBr.

En resumen, el presente estudio demuestra la potencial aplicación de la estrategia Regulome-seq como herramienta diagnóstica de enfermedades genéticas en un futuro. También propone variantes candidatas a SBr para ser evaluadas experimentalmente, y demostrar su posible asociación con el fenotipo del SBr. Los resultados obtenidos proporcionan un nuevo enfoque sobre las bases moleculares de las arritmias cardíacas relacionadas con alteraciones en las corrientes iónicas, las cuales podrían explicar algunos casos de SBr 'huérfanos'.

Summary

Brugada Syndrome (BrS) is a cardiac electrical disease with high susceptibility to ventricular arrhythmias and sudden cardiac death. Genetic variants at coding regions of the cardiac sodium channel gene *SCN5A* account for 11-24% of BrS cases. Similarly, genetic variants at coding regions of other genes, such as cardiac sodium channel regulatory beta subunits, calcium channels or accessory proteins, account for 5-10% of BrS cases. Together, coding variants at ion channel genes and their regulatory subunits account for up to 25-30% of BrS cases. Therefore, in a large fraction of BrS diagnosed patients, the etiology of the disease is still unknown (BrS 'orphan' cases). Recent genome-wide association studies (GWAS) have shown that most disease-associated variants lie within non-coding regions, specially at *cis*-regulatory regions. Based on these observations, we propose that non-coding variants located at *cis*-regulatory regions of BrS-associated genes could be a yet unexplored cause of BrS. In this thesis, we profiled genetic variation at *cis*-regulatory elements of six BrS-associated genes (*SCN5A, SCN2B, SCN3B, CACNA1C, CACNB2* and *CACNA2D*) and proposed possible candidate variants for future functional studies.

To accomplish our objective, we developed a targeted strategy, referred to as Regulome-seq. Using information of topological organization in the human genome (TADs), chromatin accessibility (DHS), a histone mark associated to active promoters (H3K4me3), and binding of a transcriptional regulator (CTCF) over BrS-associated genes in cardiac cells, we identified 1,293 non-coding regions potentially relevant to BrS. Using DesignStudio™, we designed 5,546 probes and used them to selectively capture these 1,293 regions in a cohort of 89 *SCN5A*-negative BrS individuals. After massive-parallel sequencing of the captured regions and application of a very stringent combination of filters, we identified a total of 5,349 genetic variants within the *cis*-regulatory regions of the 89 BrS individuals. From all variants identified, 4,837 correspond to Single Nucleotide Variants (SNVs) and 512 correspond to insertions and deletions–indels–(219 insertions and 293 deletions). We observed that approximately 33% of the variants are private to each BrS individual, while the remaining 67% are shared by 2 or more individuals. We also observed that SNVs are more prevalent among private variants, while indels are more prevalent among the most shared variants. Annotation of the variants with dbSNP150, 1000 Genomes and gnomAD showed that 6.84% of the identified variants were novel, and approximately 93.44% of the novel variants are private.

To identify which of the 5,349 variants identified might be functionally relevant to BrS, we took advantage of two different strategies: one strategy focused on variants affecting

transcription factor (TF) binding, and the other focused on genotype comparison with a healthy-aging cohort (Wellderly).

First, we sought to identify those variants that overlap the binding sites of cardiac TFs. We performed chromatin immunoprecipitation followed by massively parallel sequencing (ChIP-seq) experiments for GATA4, GATA6 and NKX2.5 in induced pluripotent stem (iPS)-derived cardiomyocytes but, for technical reasons, we did not obtained enough peaks and could not accomplish this purpose. Instead, we hypothesized that genetic variants affecting the binding of CTCF at boundary elements could disrupt TAD insulation of BrS-associated genes and result in an aberrant expression of these genes. After integrating the available information of CTCF binding from human cardiomyocytes with the reported 5,349 variants, we identified a total of 59 variants within 54 different CTCF binding sites. Using a machine learning-based algorithm named DeepBind, we predicted that 43 variants would decrease CTCF binding, 14 would increase CTCF binding and 2 would have no effects in CTCF binding. We also tested the effect of these variants on CTCF binding in luciferase reporter assays. These assays showed that DeepBind is a valid approach to prioritize non-coding variants affecting TF binding. The combination of DeepBind predictions and luciferase results led to the identification of 21 CTCF-overlapping variants more robustly affecting CTCF binding (16 decreasing and 5 increasing).

Second, we compared genetic variation from the 89 BrS individuals with 200 Wellderly individuals. This comparison, together with the statistical analysis applied, led to the identification of 537 candidate variants specifically associated to BrS patients (201 BrS-specific and 336 BrS-enriched). Scoring of the variants based on their deleteriousness (CADD) and tolerance to genetic variation (CDTS) showed that these 537 candidate variants are more likely to be pathogenic than other variants, especially when they are found in the *SCN5A* locus. These results suggest that non-coding variants affecting *SCN5A* gene expression might be more important for BrS pathogenesis than non-coding variants in other genes, and further underscores the important role of *SCN5A* in BrS. We also established a pathogenicity threshold for these 537 candidate variants, consisting of variants presenting a CADD score ≥15 and being found in the 1st CDTS percentile. Application of this threshold led to the identification of 10 candidate variants that may be related to BrS phenotype.

In summary, this study demonstrates the potential of our Regulome-seq approach to be used as a diagnostic tool for genetic diseases in the future. It also proposes BrS candidate variants to be experimentally evaluated, and demonstrate their putative association to BrS phenotype. The results obtained shed new light into the molecular basis of cardiac arrhythmias related to alterations in cardiac ion currents, which might explain some BrS 'orphan' cases.

I. Introduction

1. The human genome

The human genome, composed by 3 billion DNA base pairs (bp) plus the mitochondrial DNA, is defined as the total genetic content found inside a human cell. The basic unit of the genome is the DNA or deoxyribonucleic acid, that was discovered in 1868 by Friedrich Miescher[1] and, later, in 1953, its structure was described by Francis Crick and James Watson[2] (**Figure 1A**). DNA consists of a double helix with a chemical backbone composed of a five-carbon sugar named 2-deoxyribose. Deoxyriboses are joined together by phosphodiester bonds (phosphate-phosphate bonds) between the third and fifth carbon atoms of adjacent deoxyriboses. These asymmetric bonds result in the two DNA strands oriented in opposite directions, and favors a directionality of the DNA molecule from five prime (5') to three prime (3'). Deoxyriboses, in turn, are bound to one of the four different nitrogenous bases–adenine (A), thymine (T), cytosine (C) and guanine (G) through the first carbon–. In turn, nitrogenous bases are bound to complementary bases from opposite strands (A-T and G-C base pairs) by hydrogen bonds. Together, the deoxyriboses, the nitrogenous bases and the phosphodiester bonds form the nucleotides (also known as building blocks of DNA), and the different combinations of nucleotides found in the DNA are known as DNA sequence[3].

Inside the nucleus of eukaryotic cells, DNA is wrapped around histone octamers called nucleosomes (**Figure 1B**). Nucleosomes, in turn, are packed into a condensed structure called chromatin. During cell division, chromatin fiber is further packed into the 23 chromosome pairs–22 autosomal, plus the sex-determining X and Y chromosomes–[3]. All these levels of DNA compaction are explained in more detail in sections 1.2.3 and 1.2.4 for their role in regulation of gene transcription.

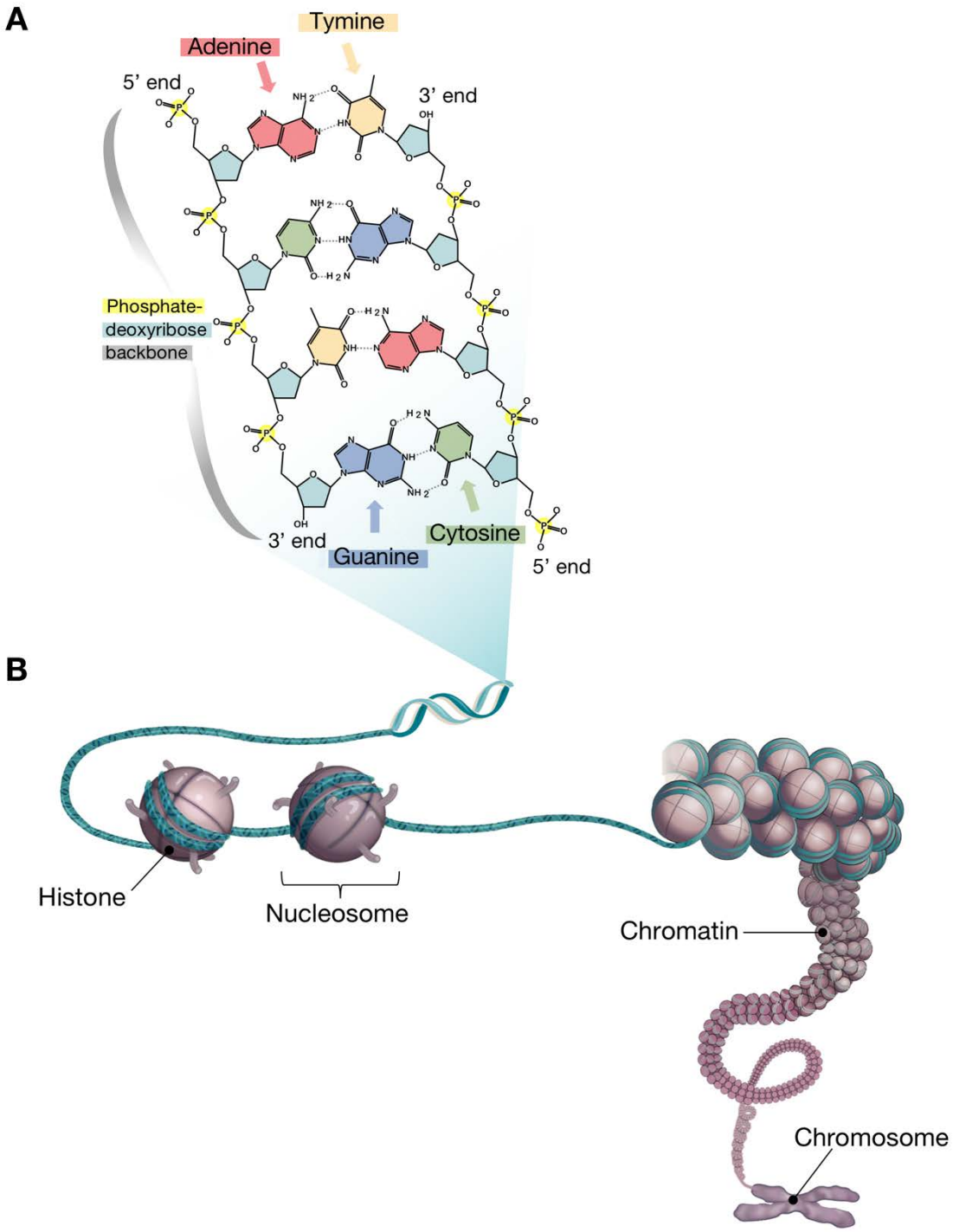


Figure 1. Structure and organization of DNA in the nucleus of eukaryotic cells. (A) Chemical structure of DNA with its deoxyribose backbone, the phosphodiester bonds and the 4 nitrogenous bases. Hydrogen bonds are shown as dotted lines. Figure adapted from Madeleine Price Ball. **(B)** Graphical representation of how the DNA double helix is wrapped around the histone octamer to form the nucleosomes, which are further condensed to form the chromatin and mitotic chromosomes. Figure adapted from Ecker *et al*.[4].

1.1. Composition of the human genome

The human genome can be divided into two main groups of regions (coding and non-coding), which are explained in the sections below.

1.1.1. Coding regions

Coding regions account for 1-2% of the genome. These regions, also known as protein-coding genes, consist of to DNA sequences encoding for proteins. Protein-coding genes are the most widely and best understood component of the human genome. Previous to the publication of the first draft of the human genome, it was estimated that as many as 140,000 protein-coding genes were present in the human genome[5,6]. However, a startling finding of this first draft was that the number of human genes appeared to be significantly lower than previous estimates, and they were revised down to ~20,000 when the full genome sequence was completed in 2003.

Protein-coding genes are composed by exons and introns that, together, are transcribed into a pre-messenger RNA (mRNA) by the RNA polymerase II (RNA pol II). This mRNA is then processed by a mechanism called splicing, in which the introns are removed, resulting in a mature mRNA formed by exons. During translation, the mRNA slides through the ribosome, which reads the mRNA three nucleotides at a time (codons)[3]. During this process, mRNA codons are recognized and bound by complementary transfer RNAs (tRNAs) that carry specific amino acids, which are chained together to form the final protein. The fact that the mRNA is read in triplets of nucleotides establishes a reading frame that is defined by the initial triplet from which translation starts, and finishes when the translation machinery reaches a stop codon. When a stop codon is found, the translation machinery dissociates and the synthesized amino acid chain is further folded and subjected to post-translational modifications to obtain the final functional protein[3].

Gene expression is the process that comprises all the aforementioned steps that lead to a functional gene product from a DNA sequence (in this case protein-coding genes).

The rules that dictate which amino acid corresponds to each specific codon is known as the genetic code. The genetic code is highly similar among all organisms and can be expressed in a table with 64 entries (**Figure 2**). The genetic code is redundant because a particular amino acid can be specified by different codons, but it is not ambiguous because different amino acids are not specified by the same codon. This property is known as degeneracy of the genetic code and it is believed to confer some robustness in front of genetic variants that might change the

codon sequence[7]. For example, the conversion of UCU codon to UCC by a genetic variant, will continue encoding for the amino acid Serine.

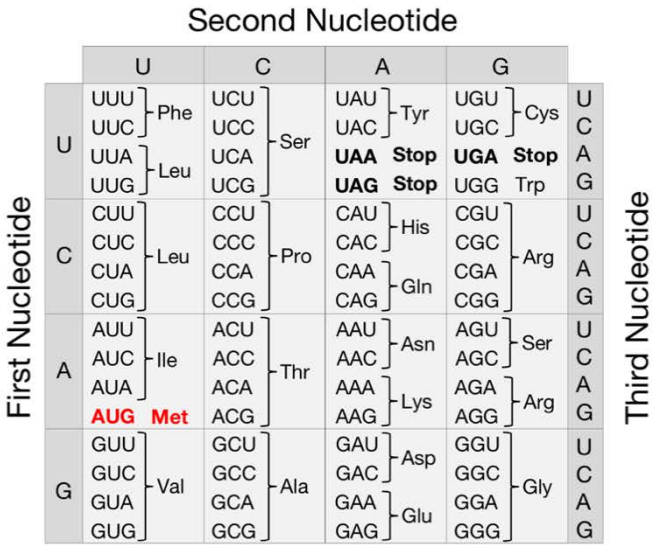


Figure 2. The genetic code. Table showing the genetic code used for translating each nucleotide triplet in the mRNA sequence into an amino acid or a termination signal in a nascent protein. Figure adapted from Alberts *et al.*[3].

1.1.2. Non-coding regions

Non-coding regions account for the remaining 98% of the genome and consist of DNA sequences that do not encode for proteins[3]. Opposite to protein-coding regions, the amount of non-coding DNA is highly variable among species, which has been correlated with organism complexity (**Figure 3**).

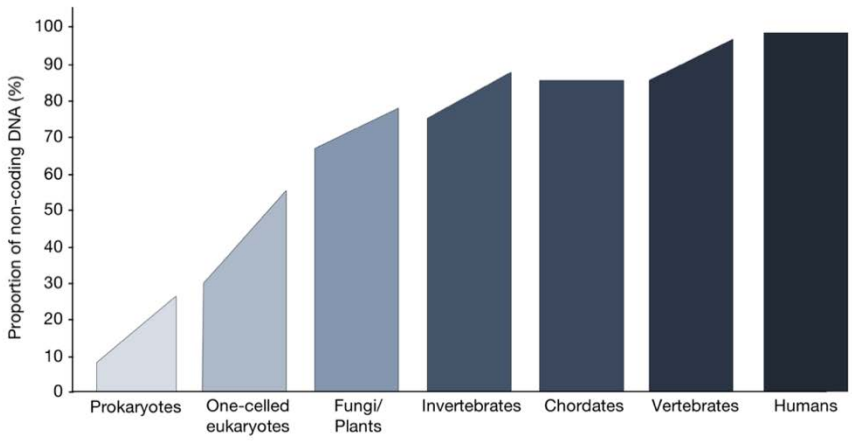


Figure 3. Relationship between non-coding DNA and organism complexity. The graph shows how the proportion of non-coding DNA increases with the complexity of organisms. Figure adapted from The ENCODE Project[8].

In contrast to protein-coding regions, for which the genetic code is understood, the rules that govern translation of non-coding regions into a biological function are still not fully understood. In this regard, several projects such as the Encyclopedia of DNA Elements (ENCODE) project and the NIH Roadmap Epigenomics Mapping Consortium were launched to build a comprehensive list of functional elements in the human genome in several tissues and cell types. An important observation from their results was that, in addition to the known functional role of protein-coding regions, several non-coding regions could be associated with a biochemical function at the organism level[8]. Based on the current knowledge, non-coding regions can be classified into five groups according to their properties:

A) Non-coding RNAs

Non-coding RNAs (ncRNAs) comprise those genes that do not encode for proteins. This group of genes are transcribed (either by RNA pol II or III) to RNA molecules that are functional and are not further translated into proteins.

Until a few years ago, most of the known ncRNAs fulfilled relatively generic functions in cells, such as the ribosomal RNA (rRNA)–that form the ribosomes–, tRNAs involved in mRNA translation, or small nuclear RNAs (snRNAs) involved in splicing[9]. However, as the extent of ncRNA functions are beginning to be explored, these ncRNAs appear to include a hidden layer of internal signals that control various levels of gene expression, including regulation of chromatin architecture (long ncRNAs), transcription (long ncRNAs) and translation (micro RNAs–miRNA–and small interfering RNAs– siRNAs–)[9]. Another class of recently discovered short ncRNAs comprises enhancer RNAs (eRNAs), whose expression has been found to correlate with the activity of its corresponding enhancer in a context-dependent fashion[10].

B) Pseudogenes

Pseudogenes are commonly defined as DNA sequences that resemble known genes but they do not encode for proteins. It is postulated that pseudogenes are once-functional genes that have been mutated over the course of evolution and, even they did loose their ability to encode for proteins, they are transcribed into functional ncRNAs such as miRNAs or siRNA[11]. Therefore, pseudogenes are considered to be involved in the regulation of different biochemical processes in cells[12].

Pseudogenes are randomly distributed throughout the genome and they can be broadly classified into two categories: unprocessed and processed. Unprocessed pseudeogenes can derive either from a gene that has been duplicated and inactivated, or from direct inactivation of genes. This group of pseudogenes keep their intron-exon structure and they are often located

close to the paralogous parent gene. In contrast, processed pseudogenes derive from retrotransposition. Retrotransposition occurs when an RNA molecule is reverse transcribed to DNA and inserted into a new genomic location, which can be a different chromosome than the parental gene. This group of pseudogenes lack introns and promoter sequences, and are flanked by sequence repeats[13].

C) Introns

An intron is any nucleotide sequence within a gene that is removed by RNA splicing during the maturation of the RNA product[7]. Introns are found within genes of most organisms and, even they are mostly located in protein-coding genes, they can also be found in some non-protein coding genes such as those that generate rRNA and tRNA[14].

Introns impose a huge energetic burden to the cell, as they have to be removed from the pre-mRNA by splincing. However, as intron function is being studied, there is growing evidence that introns have important roles in several processes related to gene transcription[15].

Broadly speaking, at least four distinct classes of introns have been identified: (i) introns in protein-coding genes that are removed by spliceosomes (spliceosomal introns); (ii) introns in ncRNAs removed by proteins; (iii) self-splicing group I introns removed by RNA catalysis; and (iv) self-splicing group II introns removed by RNA catalysis[3].

The most common introns are spliceosomal introns. They are characterized by the presence of t GT and AG motifs at the 5' and 3' ends of the intron, respectively. These motifs are recognized by the spliceosome complex that removes the introns from the mRNA and joints exon ends. Splicing is a highly regulated process, since alterations in splicing might lead to changes in protein levels that can result in aberrant cellular metabolism and/or function[16,17]. For example, several genetic variants altering the consensus splicing sequences have been associated to hereditary monogenic disorders[17].

D) Repetitive sequences

Repetitive DNA accounts for up to 50% of eukaryotic genomes[18], and corresponds to short and long DNA sequences that are repeated hundreds or thousands of times. Repetitive DNA can be classified into two major groups: tandem repeats and dispersed repeats. The formers comprise short DNA sequences (5-60 bp) that are repeated and placed adjacent to each other. They can be found as blocks of tandem repeats concentrated at specific chromosomal loci (for example repeats composing centromeric and telomeric regions), or scattered throughout the genome as simple sequence repeats[18].

Dispersed repeats, in contrast, comprise longer DNA sequences that can change their position within a genome. These type of repeats are known as transposable elements, and they are classified into, at least, two classes: Class I and Class II elements. Class I elements or retrotransposons are DNA sequences that are transcribed into RNA, which is then reverse transcribed to DNA and inserted back into the genome at a new position. This class of transposable elements are highly heterogeneous in composition and include several subclasses such as long terminal repeats (LTR), long interspersed nucleotide elements (LINEs) and short interspersed nucleotide elements (SINEs). On the contrary, Class II elements or DNA transposons are DNA sequences that are directly excised from their genomic position and are inserted into a new genomic location[19].

The significance of repetitive DNA in the genome is not completely understood, but some repetitive sequences such as short tandem repeats found in centromeres and telomeres play crucial roles in maintaining chromosome structure. Other repetitive elements have been related to regulatory functions, acting at several stages of gene expression, such as transcription–several enhancers and promoters are enriched in repetitive sequences–, post-transcriptional RNA processing and translation into proteins[20].

E) *Cis*-regulatory elements

Cis-regulatory elements are defined as non-coding DNA sequences that regulate the expression of protein-coding genes. These elements are crucial for the proper regulation of the genome as they orchestrate the genetic programs required for embryonic development, as well as differentiation and function of the different cell types that compose the organism[21–23].

There are several types of *cis*-regulatory elements that participate in the regulation of gene expression, and they do so using different mechanisms: some *cis*-regulatory regions regulate gene expression by interacting with the transcriptional machinery, while others influence in gene expression by participating in the spatial organization of the genome[24].

Due to their importance in the regulation of gene expression, *cis*-regulatory elements will be explained in more detail in the section below.

1.2. Regulation of gene expression at the transcriptional level

Regulation of gene expression includes a wide range of mechanisms aimed to increase or decrease the production of specific gene products (protein or RNA). Gene expression programs are regulated in a highly sophisticated manner to ensure a precise and tight spatiotemporal expression of genes. Thus, any alteration in the mechanisms that regulate gene expression may

have severe consequences that can lead to disease[25]. Gene expression can be regulated at any step: transcription, mRNA processing, and post-translational modifications of a protein (**Figure 4**).

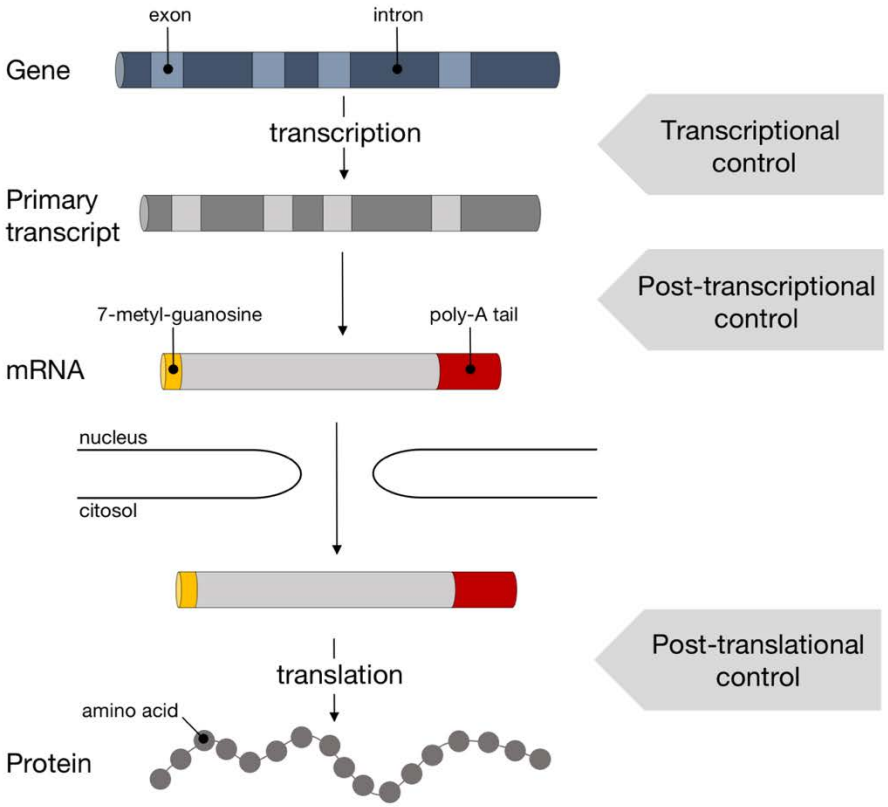


Figure 4. Regulation of gene expression. Schematic representation of the different stages in the DNA-mRNA-protein pathway in which gene expression can be regulated.

Transcriptional regulation represents the first and most important level of regulation. It involves all the mechanisms that control assembly of the RNA pol II complex to start gene transcription. Transcriptional regulation mostly depends on two factors: (i) binding of transcription factors (TFs) at *cis*-regulatory elements[26] (section 1.2.2), and (ii) chromatin modifications that will allow the accessibility of TFs to DNA[27] (see section 1.2.3). Apart from these two factors, a third element that is also related to the control of gene transcription is the tridimesional organization of the genome (section 1.2.4). This spatial organization of the genome facilitates the interaction between *cis*-regulatory elements and their target genes, but also blocks the interaction of *cis*-regulatory elements to off-target regions[28].

Post-transcriptional regulation represents the second level of regulation. It involves all the modifications that occur during mRNA processing before it is translated into a protein. These modifications include: (i) addition of a poly-A tail at the 3' end, (ii) addition of a 7-metil-guanosine at the 5' end and, (iii) splicing of introns[3].

Post-translational regulation represents the third level of regulation. It refers to all chemical modifications of proteins (phosphorylation, glycosylation, acetylation, methylation, etc.) added after translation. These modifications modulate several properties of proteins such as the activation/inactivation of their catalytic function, their localization to different cell compartments, their degradation or their DNA binding affinity.

Post-translational regulation also includes other protein modifications such as their tridimensional folding or addition of prosthetic groups, which are crucial for the activity of several proteins such as enzymes[3].

1.2.1. *Cis*-regulatory elements

Cis-regulatory elements and TFs are explained in this thesis in separate sections. However, their function can bee seen as a single entity given that *cis*-regulatory elements require the binding of TFs to accomplish their regulatory role[29].

In human genomes, we can distinguish four types of *cis*-regulatory elements:

1.2.1.1. Promoters

Promoters are DNA sequences where transcription of protein-coding genes begins. Two different regions can be distinguished inside the promoter sequence: the core promoter and the proximal promoter (**Figure 5**).

The core promoter is the region that serves as the docking site for the basic transcriptional machinery to assemble (RNA polymerase II and general TFs). It defines the position of the transcription start site (TSS), as well as the direction of transcription[30]. The first described core promoter element was the TATA-box, named for its conserved DNA sequence TATAAA. The first hypothesis after the identification of the TATA-box was that core promoters were universal. However, statistical analysis of ~10,000 predicted human promoters revealed that core promoters were not as universal as previously thought[31], and a series of promoters containing other types sequences than the TATA-box were found[31]. This diversity in core promoter composition is believed to have a functional significance, as it has been observed that different core promoters can limit the regulatory inputs to which they will respond[30,32].

The proximal promoter is a region located immediately upstream from the TSS, containing a high density of binding sites for transcriptional activators[31]. Approximately a 60% of human proximal promoters fall near a CpG island, which is a relatively short sequence of DNA (typically 500 bp to 2 kb in length) enriched in GC nucleotides. Many CpG islands scattered throughout the genome are methylated; however, those CpG islands close to proximal promoters are

unmethylated. Indeed, correlations exist between the presence of CpG islands and certain core promoter elements. For example, TATA-boxes are more common in promoters that do not have a CpG island nearby[31,33].

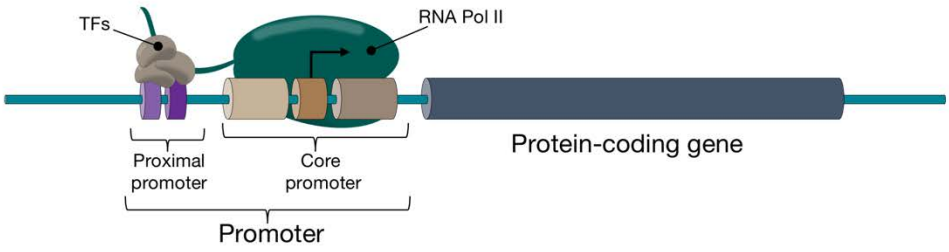


Figure 5. Schematic representation of a promoter region. The proximal promoter is located upstream of the TSS and contains multiple binding sites for TFs. The core promoter contains the TSS and it serves as the docking site for the RNA pol II. Arrow indicates TSS.

1.2.1.2. Enhancers

Enhancers are DNA sequences that increase gene transcription and orchestrate the precise gene expression patterns required for numerous processes, including embryonic development and function of organisms[34]. Unlike most proximal promoters, enhancers are typically long-distance elements that can reside several hundred kilobases upstream of their target gene, and they can regulate gene transcription independently of their orientation, position and distance to promoters[35]. The interaction between enhancers and their target promoters is mediated by a complex of several TFs and co-factors named Mediator[36] (**Figure 6A**). The Mediator complex is involved in the formation of a DNA loop that brings enhancers and promoters into close proximity, allowing the interaction of enhancers with the transcriptional machinery anchored in their target promoters.

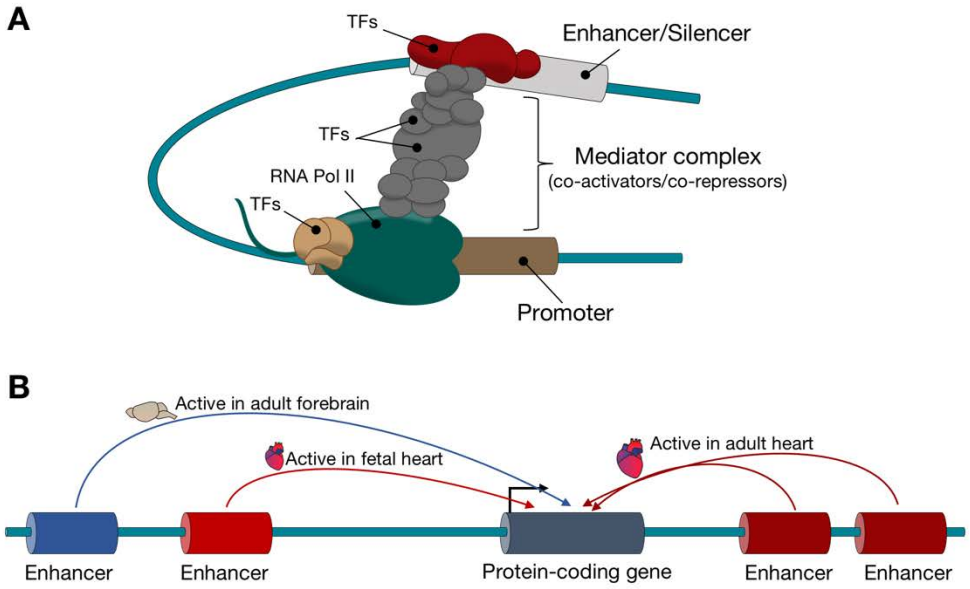


Figure 6. Schematic representation of enhancer function. (A) Interaction between enhancers or silencers with the transcriptional machinery (TFs and RNA pol II) anchored in their target promoters. The Mediator complex involved in the DNA loop formation is also shown. **(B)** Exemplification of how different enhancers might be regulating the same gene at different tissues and distinct spatiotemporal moments. Arrow indicates TSS.

To date, >100,000 enhancers have been identified in the human genome, outnumbering genes by approximately an order of magnitude[8]. This finding, together with the uncertainty of what gene or genes are targeted by a given enhancer, has raised the question of enhancer functionality. It has been suggested that different enhancers might be regulating the same promoter, but it remains uncertain whether different mammalian enhancers regulate the same promoter at distinct spatiotemporal moments[37,38], or if this regulatory complexity results in functional redundancy among enhancers associated with the same gene[39,40] (**Figure 6B**). In this regard, Osterwalder *et al.,*[41] used the mouse developing limb to study enhancer usage during embryonic morphogenesis. In this study, they individually deleted ten conserved enhancers that regulate genes associated with mouse and human congenital limb malformation. Strikingly, they did not observe any significant change and, more important, they did not observe any limb abnormalities. In contrast, when they simultaneously deleted those enhancers located at close proximity that showed similar activity patterns, they observed an alteration of target-gene expression levels and the appearance of a severe limb phenotype[41].

Together, these observations reinforce the hypothesis of enhancer redundancy and could explain why the presence of genetic variants causing enhancer loss-of-function could result in no phenotypic effects[42]. However, we have to take into account that not all genes are regulated

by more than one enhancer and, that there are several cases of genetic variants altering enhancer activity that have been related to disease phenotypes[43,44].

Apart from enhancers, another class of regulatory elements has been recently described. These are referred to as super-enhancers because they are associated to genes that dictate cell identity. Super-enhancers tend to span large genomic regions, with their median size generally an order of magnitude larger than that of conventional enhancers[45]. Active superenhancers are highly occupied by TFs and co-activators, especially the Mediator complex[46].

1.2.1.3. Silencers

Silencers are DNA sequences where TFs bind to repress gene transcription. Similar to enhancers, silencers function independently of the distance and orientation relative to their target promoters. It has been suggested that silencers repress gene transcription using the same mechanism of loop formation as enhancers[33] (**Figure 6A**).

1.2.1.4. Insulators or boundary elements

Insulators are DNA sequences that restrict the functional limits of enhancers and silencers (**Figure 7**). They divide the genome into discrete areas of gene expression regulation and play a crucial role in the tridimensional organization of the human genome[3,24] (section 1.2.4).

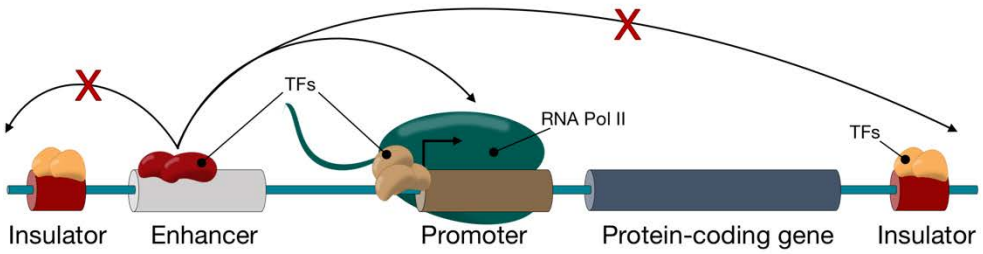


Figure 7. Schematic representation of insulator function. The two insulators shown block the spread of enhancer activity to other genes found outside the boundaries. Arrow indicates TSS.

1.2.2. Transcription factors

TFs are proteins that regulate gene transcription by binding to *cis*-regulatory elements[47]. For this reason, they are sometimes referred to as *trans*-regulatory elements.

1.2.2.1. Function of transcription factors

According to their function, TFs can be broadly classified into two groups: general TFs and specific TFs[33].

A) General transcription factors

General TFs (TFIIA, TFIIB, TFIID, TFIIE and TFIIH) bind to promoters and they are required for gene transcription initiation (**Figure 8**). Briefly, TFIID binds to specific core promoter sequences and recruits TFIIA and TFIIB. Binding of these three factors recruits RNA pol II to the TSS. Then, TFIIE and TFIIH bind to the complex, finishing with the formation of the transcriptional machinery. Once settled this machinery, phosphorylation of the C-terminal domain of the RNA pol II by TFIIH induces the elongation of the mRNA transcript.

The assembly of the transcriptional machinery to the promoter is sufficient to induce basal transcription, but the interaction with specific TFs bound to enhancers or silencers will increase or repress the levels of gene transcription, respectively[48].

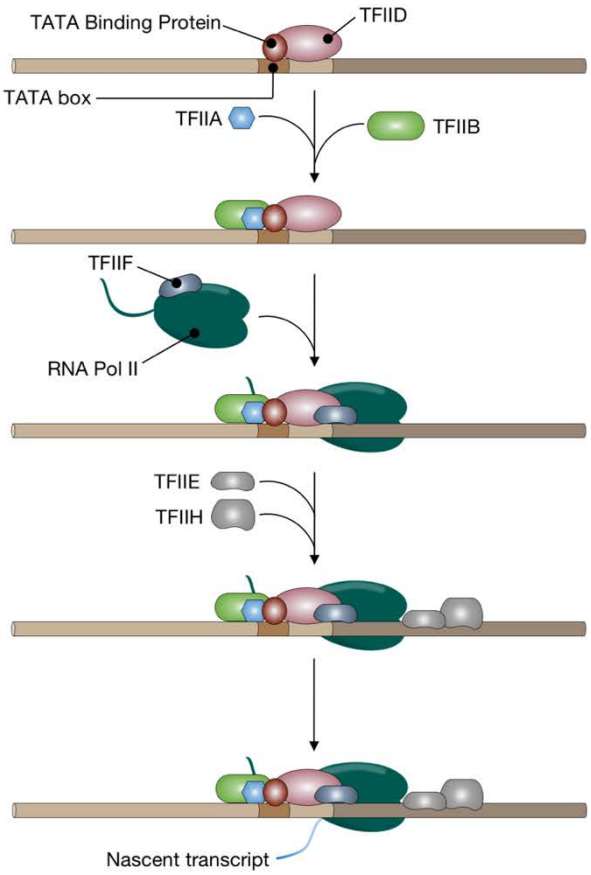


Figure 8. Basal transcription in eukaryotes. Schematic representation of the first steps in the activation of transcription by basal TFs (TFIIA, TFIIB, TFIID, TBP, TFIIE, TFIIH) and RNA Pol II. Figure adapted from Levine[49].

B) Specific transcription factors

Specific TFs are responsible for the differential gene transcription in a temporal- and tissue-specific manner. These TFs can act as activators or repressors depending on the *cis*-regulatory element they bind, and the same TF can act as activator or repressor depending on the context[47].

Activators bind to enhancers and interact with the transcriptional machinery anchored at promoter regions via DNA-loop formation (**Figure 6A**). This interaction is facilitated by the Mediator complex, which is composed of several TFs and co-factors bound by protein-protein interactions[50].

Repressors bind to silencers and have been proposed to repress the activity of the transcriptional machinery via two different mechanisms. They can either compete with activators to interact with the transcriptional machinery via the same DNA-loop formation, or they can directly bind to activators via protein-protein interactions to prevent the interaction of activators with the transcriptional machinery–also known as squelching–[51].

In addition, activators and repressors can influence transcription by changing the chromatin structure in the vicinity of their binding sites. They can recruit chromatin modifiers (section 1.2.3) that, in turn, will increase or decrease the DNA accessibility to other TFs[52].

1.2.2.2. DNA recognition by transcription factors

A basic feature of TFs is that they contain DNA-binding domains that recognize specific sequences within *cis*-regulatory elements[53]. These DNA-binding domains can adopt different structural properties that can be shared by several TFs. Accordingly, TFs can be classified in different families based on their DNA-binding domain[54,55] (**Table 1**).

Table 1. Classification of TFs based on their DNA-binding domains. Adapted from Gonzalez[55] and Wingender[56].

Superclass	Specific properties
1: Basic domains (Helix-loop-helix and Leucine zipper)	They are rich in basic amino acid side chains. TFs with basic domains require dimerization with other factors via specific dimerization areas
2: Zinc-coordinating DNA-binding domains (Zinc fingers)	Their folding is organized by one or two coordinated zinc ions
3: Helix-turn-helix	They contain a DNA-recognition helix, the structure and DNA-binding of which is stabilized by additional helices. This is the second largest and functionally very heterogeneous group of TFs
4: Beta-scaffold factors with minor groove contacts	They bind to DNA through extended strands or beta-sheets, which preferentially bind in the minor groove of the DNA
0: Other TFs	They are composed of an heterogeneous class of TFs with different DNA-binding domains

TFs interact with their binding sites using a combination of electrostatic (ex: hydrogen bonds) and Van der Waals forces. Due to the nature of these chemical interactions, most TFs bind to DNA in a sequence specific manner. However, not all the bases in the TF binding site may actually interact with the TF. In addition, some of these interactions may be weaker than others. Thus, TFs do not bind just one sequence but are capable of binding a subset of closely related sequences, each with different strength of interaction.

DNA sequences bound by TFs are named **motifs** and they are usually represented as sequence logos (**Figure 9**). These logos or consensus motifs are graphical representations of the sequence conservation of the different nucleotides among the TF binding site. They can be obtained by comparing all the different sequences bound by a specific TF and, the more invariant/conserved is a nucleotide at a given position, the bigger it is represented in the sequence logo. The more conserved nucleotides, also referred to as **core nucleotides**, are essential for TF binding and that genetic variants affecting them will have a higher impact in TF binding. In contrast, the presence of less conserved nucleotides has been proposed to confer some tolerance to genetic variation, which might explain why many changes in TF binding sequence do not result in measurable changes in gene transcription levels[57,58]. This tolerance to genetic variation is thought to be important for evolution as it might contribute to species-specific adaptations as well as to within-species variation in complex phenotypes[59,60].

Figure 9. TF binding consensus motif. Representation of GATA4 and the NKX2.5 consensus motifs. The size of the nucleotide in the logo corresponds to conservation, which can be linked to the importance of the nucleotide for TF binding. Motifs extracted from JASPAR[57].

Despite of the importance of the motif sequence for TF binding, the presence of the motif has been observed to be not sufficient for TF binding. Binding motifs are usually short (4 to 30 bp in length), which results in potential TF binding sites occurring by chance in the genome[57]. However, not all the compatible binding sequences are bound by TFs. In fact, there are numerous features intrinsic to TFs or DNA-binding sites that can modulate TF-DNA recognition (**Figure 10**). Many TFs require the formation of dimers to bind DNA (**Figure 10A**). These dimers can be formed by two identical molecules (homodimers) or by two different molecules (heterodimers). Other TFs require protein-protein interactions with co-activators or co-repressors for efficient DNA binding (**Figure 10B**). In addition, TF binding can also be influenced by sequence features such as DNA methylation, the genomic context where the TF binding site is embedded, or the presence of genetic variants at TF binding sites[61] (**Figure 10C-E**).

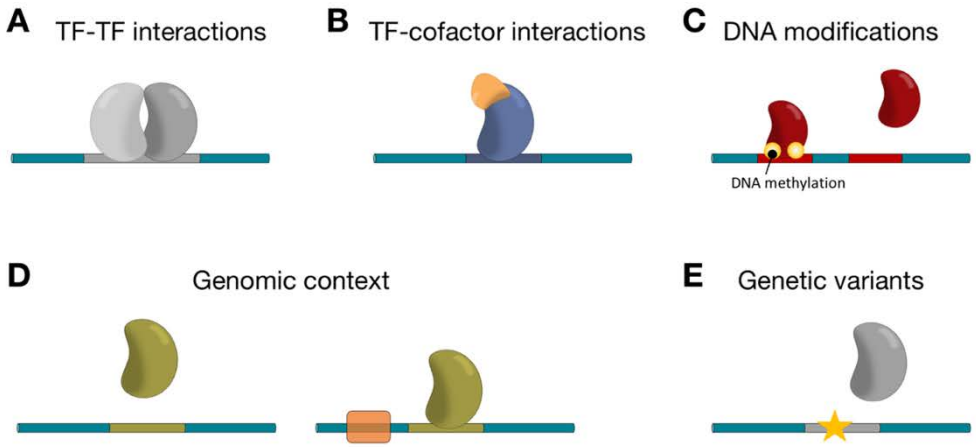


Figure 10. Modulation of TF-DNA recognition by different mechanisms. (A) Interaction with other TFs. **(B)** Interaction with cofactors. **(C)** DNA modifications such as DNA methylation. **(D)** Genome context (depicted by orange box), such as GC content. **(E)** Presence of genetic variants inside the TF binding site. Figure adapted from Inukai *et al.*[61].

1.2.3. Chromatin

1.2.3.1. Chromatin structure

As mentioned earlier, genomic DNA is compacted inside the eukaryotic nucleus in the form of a dynamic polymer called chromatin (**Figure 11A**). The nucleosome is the fundamental unit of chromatin and it is composed of an octamer of the four core histones (H2A, H2B, H3 and H4) around which 147 bp of DNA are wrapped (**Figure 11B**). In addition to the histone octamer, histone H1 participates in the formation of nucleosomes by stabilizing the interaction of the wrapped DNA with the histone octamer[62]. The core histones are predominantly globular except for their N- and C-terminal tails. These are projected outside the nucleosome and are subject to reversible post-translational modifications at multiple residues that modulate chromatin states (**Figure 11C**).

Classically, two chromatin states have been described depending on the levels of compaction. Euchromatin represents a more relaxed chromatin state, facilitating accessibility of TFs and the transcriptional machinery to *cis*-regulatory elements. For this reason, euchromatin comprises regions with transcriptionally active genes. Conversely, heterochromatin is more compacted and less accessible to TFs and the transcriptional machinery, being associated to transcriptionally inactive genes. Heterochromatin can be further classified into two different types: constitutive heterochromatin, whose distribution is shared among all the cells and plays a vital role in chromosome structure (for example the telomeres and centromeres); and facultative heterochromatin, whose compaction is cell-type or cell-state specific[63].

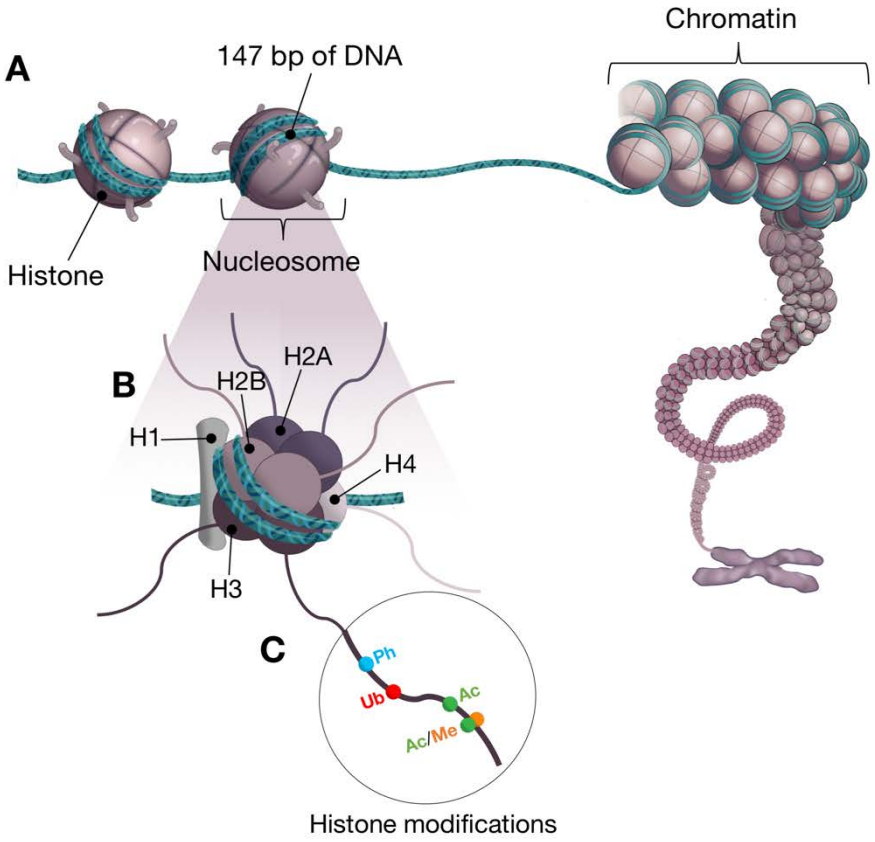


Figure 11. Chromatin structure. (A) Representation of a nucleosome, formed by 147 bp of DNA wrapped around the histone octamer. **(B)** Schematic representation of the histone octamer, showing the H2A, H2B, H3 and H4 histones with corresponding N-terminal tails. The anchor histone H1 is also depicted. **(C)** H3 N-terminal tail with several post-translational modifications that can be added. Figure adapted from Ecker *et al.*[4].

1.2.3.2. Chromatin dynamics

Chromatin is a highly dynamic structure modulated by multiple modifications that influence its compaction, organization and function. Chromatin modifications include: histone modifications, chromatin remodeling by nucleosome positioning, and DNA methylation. Importantly, all these modifications affect chromatin condensation without altering the DNA nucleotide sequence, and they are involved in the regulation of several processes such as transcription, DNA replication or repair[64].

A) Histone modifications

As previously mentioned, the N-terminal tails of histones are subject to multiple post-translational modifications at multiple residues[64]. To date, more than 30 different histone modifications have been described. Depending on the type of modification and the residue modified, histone modifications will be involved in different processes such as regulation of gene

transcription or DNA replication. **Table 2** and **Figure 12** show the most well-known histone modifications involved in transcriptional regulation.

Table 2. Different classes of modifications identified on histones. Adapted from Kouzarides[64].

Chromatin Modifications	Residues Modified	Functions Regulated
Acetylation	**K**-ac	Transcription, Repair, Replication, Condensation
Methylation (lysines)	**K**-me1, **K**-me2, **K**-me3	Transcription, Repair
Methylation (arginines)	**R**-me1, **R**-me2a, **R**-me2s	Transcription
Phosphorylation	**S**-ph, **T**-ph	Transcription, Repair, Condensation
Ubiquitylation	**K**-ub	Transcription, Repair
Sumoylation	**K**-su	Transcription
ADP ribosylation	**E**-ar	Transcription
Deimination	**R** > Cit	Transcription
Proline Isomerization	**P**-cis > **P**-trans	Transcription

Overview of different classes of modification identified on histones. The functions that have been associated with each modification are shown. K (lysines), R (arginines), S (serine), T (threonine), E (glutamic acid), P (proline).

Histone modifications can influence gene transcription by two different mechanisms (**Figure 12**). The first mechanism is related to the interaction of histones and DNA within the nucleosome, which is stabilized by the difference in charges between histones (positively charged) and DNA (negatively charged). Several histone modifications may change the histone charge, thereby altering DNA-histone interactions leading to a more condensed chromatin state. Chromatin condensation, in turn, influences DNA accessibility to TFs and, in consequence, the levels of gene transcription[64] (**Figure 12A**). The second mechanism involves the recognition of specific histone modifications by effector proteins that contain reader domains. Recruitment of these proteins, that might either add other histone modifications or change the nucleosome position through remodeling complexes leads to changes in gene transcription[65,66] (**Figure 12B**).

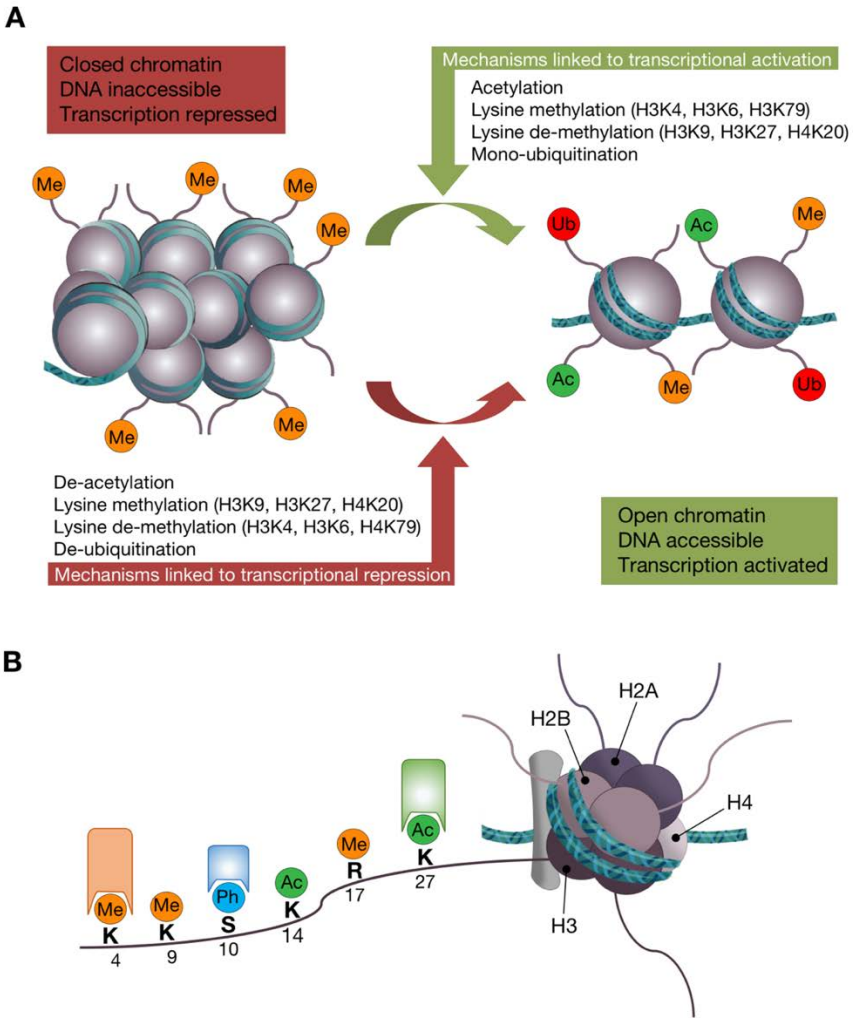


Figure 12. Role of histone modifications in gene expression. **(A)** Left: representation of closed chromatin, in which the DNA is inaccessible to the transcriptional machinery and transcription is therefore repressed. Right: representation of open chromatin, in which the DNA is accessible to the transcriptional machinery ant transcription is therefore activated. Some examples of histone modifications involved in this process as well as the residues modified are also shown. Figure adapted from Basset and Barnett[67]. **(B)** Histone H3 N-terminal tail displaying several post-translational modifications, some of them being recognized by distinct effector proteins. Figure Adapted from Musselman *et al.*[68]. Me (methylation), Ub (Ubiquitination), Ac (Acetylation) and Ph (Phosphorylation).

It has been proposed that several modifications influence the deposition of other modifications and that the different patterns of histone modifications store great information that influences gene transcription (**Figure 13**)[64]. In this sense, the hypothesis of the "histone code" proposed that specific combinations of these modifications dictate the transcriptional activity of a determined promoter[69]. Given the great diversity of histone modifications, it is thought that each modified residue has a biological function, but unfortunately many of these functions are yet to be discovered.

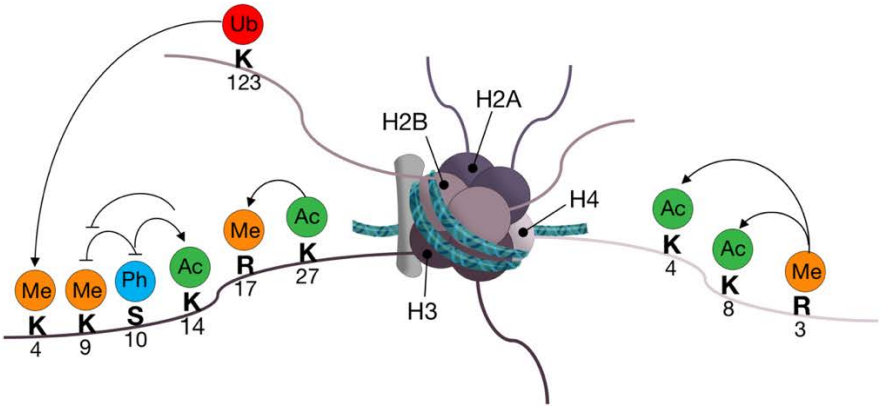


Figure 13. Crosstalk between histone modifications. The positive influence of one modification over another is shown by an arrow and the negative effect by a dash-line. Figure adapted from Kouzarides[64].

B) Nucleosome remodeling by chromatin-remodeling complexes

Chromatin-remodeling complexes or remodelers are ATP-remodeling complexes composed of 10 to 15 subunits, some of them containing reader domains of histone modifications that allow remodelers to bind to specific regions of the chromatin. Once recruited, chromatin-remodelers can induce changes in the position of the nucleosomes that increase the accessibility of TFs, thus facilitating gene transcription. However, nucleosome remodeling can also reduce the accessibility of TFs, blocking gene transcription. The position of nucleosomes can be modified by five different mechanisms: nucleosome displacement, nucleosome elimination, nucleosome assembly, nucleosome spacing, and histone replacement[70] (**Figure 14**).

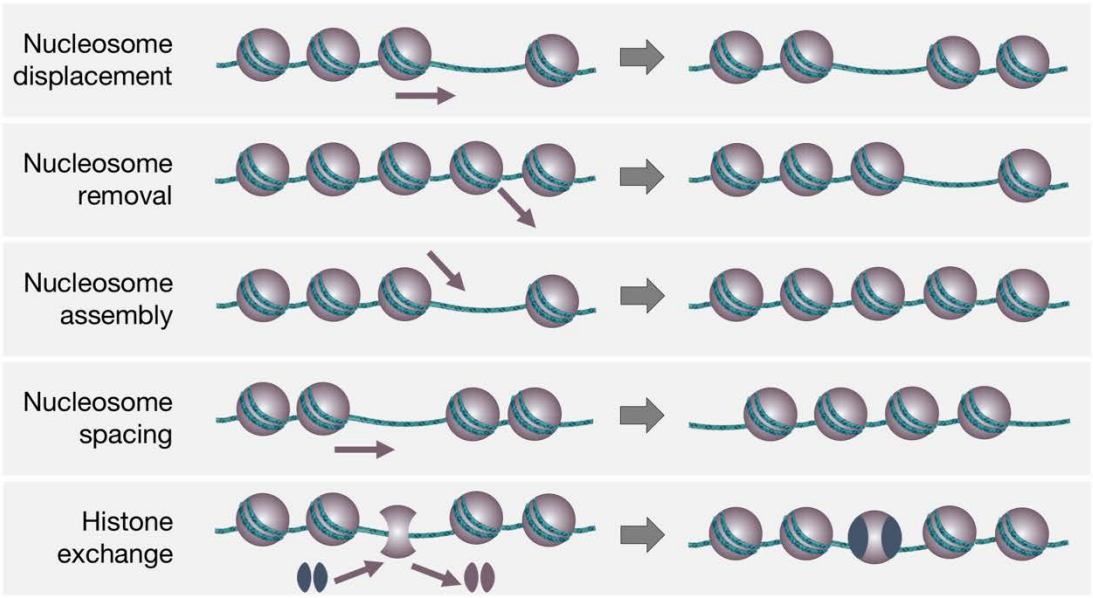


Figure 14. Nucleosome remodeling. Schematic representation of the dynamic functions of chromatin remodeling catalyzed by the chromatin remodeling complexes. Figure adapted from Petty and Pillus[70].

C) DNA methylation

DNA is methylated by adding methyl groups covalently bound to DNA cytosines. DNA methylation is catalyzed by DNA methyltransferases and, even they are not as dynamic as histone modifications, methyl groups can be removed by DNA demethylases[71].

DNA methylation mainly occurs in CpG islands present in the genome, which are short interspersed DNA sequences that are enriched in GC nucleotides[72]. As previously mentioned, about 60% of human promoters contain CpG islands that are related to transcriptional regulation. Methyl groups interfere with DNA accessibility of TFs and, therefore, methylated CpG islands have been associated to transcriptional repression while demethylated CpG islands have been associated to transcriptional activation[73,74].

In summary, the dynamics of the histone modifications, nucleosome remodeling and DNA methylation allows the cell to rapidly respond to several stimuli, modifying accessibility of TFs to DNA and inducing transcriptional changes in the activation/repression balance of genes.

1.2.4. Tridimensional organization of the genome

Regulation of human genome not only depends on the linear genome sequence that embeds millions of *cis*-regulatory elements, but also the tridimensional chromatin architecture that orchestrates the interplay between *cis*-regulatory elements and their target genes[75].

The different levels of chromatin organization in mammalian cells are described below.

1.2.4.1. Insulated neighborhoods

Recent studies of chromatin modification landscapes across a large number of human tissues and cell types have greatly improved the understanding of genome function and regulation[8,76]. As explained before, in a linear DNA, *cis*-regulatory elements may be found several hundreds to thousands of bp away from the genes they regulate. According to the polymeric nature of chromatin fibers, two distant genomic loci would contact each other at very low frequency due to random collision[77]. However, it has been observed that certain loci have a significantly higher contact frequency than expected by chance. These chromatin contacts, known as **insulated neighborhoods**, are cell-type specific and are associated with the DNA-loop formation to allow enhancer-promoter interactions[78,79] (**Figure 15A**).

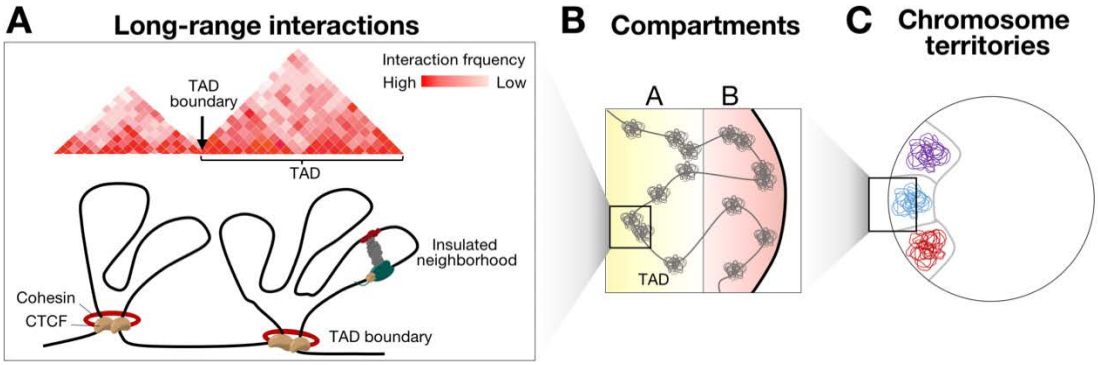


Figure 15. Hierarchical genome organization in mammals. From a fine to a large scale, long-range interactions, TADs, compartment A/B, and chromosome territories are shown (left to right). **(A)** Two types of long-range chromatin interactions are shown: CTCF-CTCF loop mediated by the Cohesin complex establishing TAD boundaries and, promoter-enhancer interactions forming insulated neighborhoods. **(B)** Compartments A and B are indicated in yellow and pink background, respectively. Compartment B is correlated with nuclear lamina. TADs are represented as scrawls symbolizing condensed chromatin. **(C)** Each chromosome territory is denoted by different colors. Only three chromosomes are shown. Figure adapted from Yu and Ren[75].

1.2.4.2. Topological associated domains

Insulated neighborhoods are further organized in to higher order structures named **topological associated domains**[80] (**TADs**; **Figure 15A** and **Figure 16**). TADs represent physically isolated units of genome organization which can be considered as functionally separated of the rest of the genome for two main reasons. First, the contact frequency between genomic loci in the same TAD is several-fold higher than the contact frequency of loci belonging to different TADs. Moreover, one pair of loci from the same TAD is also spatially closer than another pair of similar genomic distance but from different TADs[80], suggesting that TADs are

involved in the co-regulation of genes found within the same TAD (**Figure 16A**). Second, TADs have an insulation property that block the spread of activity of *cis*-regulatory elements or the spread of repressive chromatin to regions located in neighboring TADs[28] (**Figure 16B** and **C**).

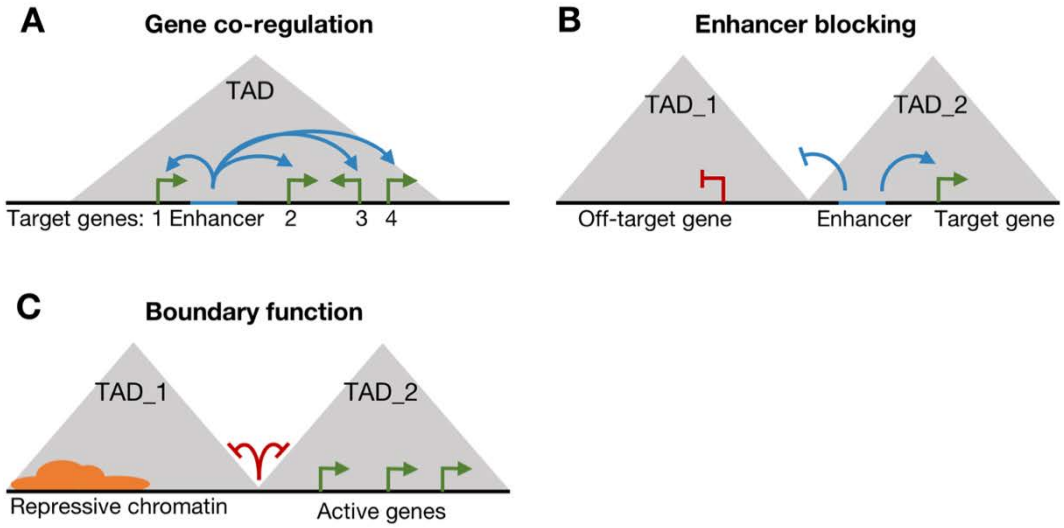


Figure 16. Function of TADs in transcriptional regulation. (A) Co-regulation of multiple genes by a single regulatory element within a TAD. **(B)** Enhancer blocking restricts the interaction of enhancers to target genes within the same TAD. **(C)** Boundary function of TADs restricts the spread of repressive chromatin into active domains and vice versa. Figure adapted from Dixon *et al*.[28].

The limits between adjacent TADs are defined by insulators. These insulators are bound by **CCCTC-binding factor** (**CTCF**), which is a 11 zinc finger TF involved in the formation of higher order organization of the human genome as well as in the formation of chromatin barriers to protect from heterochromatin[81]. Despite the evidence that about 80% of TAD boundaries are bound by CTCF, the exact mechanism by which CTCF insulates chromatin interactions between TADs has been a question of intensive research.

Currently, two different models have been proposed to explain the formation of TAD boundaries: the “handcuff model” and the “extrusion model” (**Figure 17**). The handcuff model posits that the two ends of a TAD are bound by CTCF proteins that interact with each other to recruit the Cohesin complex and stabilize the loop[82] (**Figure 17A**). Cohesin is a chromosome-associated multisubunit protein complex that is highly conserved in eukaryotes. Cohesin plays essential roles in several biological processes including: (i) chromosome segregation in dividing cells, (ii) DNA repair and, (iii) establishment of TAD boundaries[83]. Alternatively, the extrusion model proposes that a chromatin motor complex (such as the Cohesin complex) loads onto DNA and then extrudes a progressively larger loop until it is stalled by CTCF binding at convergent orientation, thus setting up TADs[84] (**Figure 17B**).

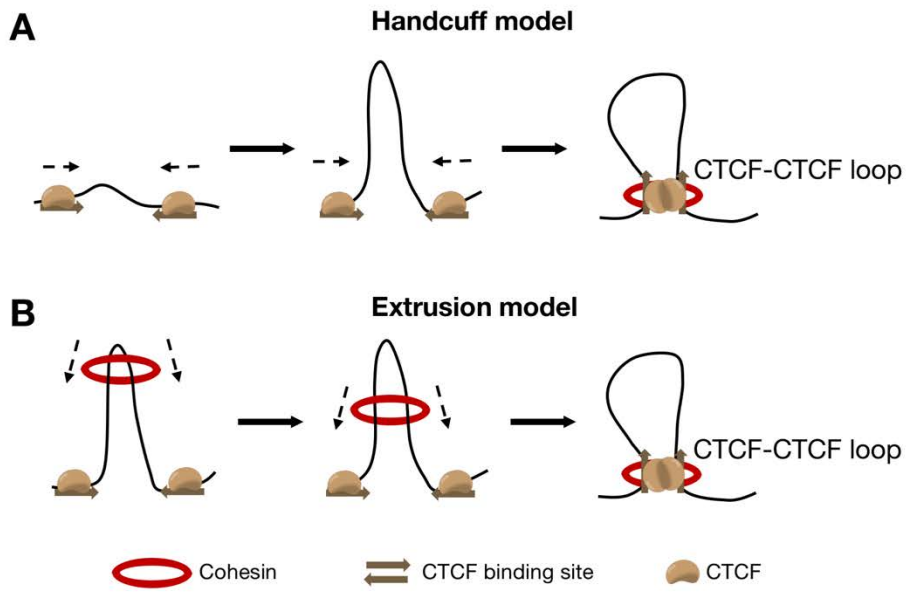


Figure 17. Models proposed for TAD formation. (A) The handcuff model proposes the formation of TADs by CTCF and Cohesin complex connecting the two sequences together. **(B)** The extrusion model involves a pair of tethered CTCF proteins bound to chromatin motors that propel the extrusion of the chromatin fiber while the two CTCF molecules slide the chromatin fiber in opposite directions before pausing at converging CTCF DNA binding motifs. Adapted from Yu and Ren[75] and Dixon *et al.*[28].

Contrary to insulated neighborhoods, TADs are relatively stable across cell types and appear to be independent of tissue-specific gene expression or histone modifications. In this line of evidence, TAD positioning has been shown to be evolutionary conserved (50-70% of TAD boundaries are shared between human and mouse ESCs). This observation has led to the consideration of TADs as the fundamental unit of genome organization in different species[85].

1.2.4.3. Cell compartments

TADs are further organized into two different **compartments** (compartment A and B), separated in space and organized in a spatially polarized manner (**Figure 15B**). Compartment A and B partition is cell-type specific, and switches between these two compartments has been observed for ~60% of the human genome[86]. Compartment A sequences are early-replicating, contain a high density of genes, exhibit strong mRNA expression activities and are enriched for H3K36me3[77]. In contrast, compartment B sequences are late replicating, contain a low density of genes (that are transcriptionally silent or expressed at low levels) and are enriched for typical heterochromatin histone marks H3K9me2 and H3K9me3[87,88].

1.2.4.4. Chromosome territories

The largest level of tridimensional organization of the human genome are **chromosome territories,** which are specific regions of the nucleus occupied by specific chromosomes[89] (**Figure 15C**). Small gene-rich chromosomes are generally located close to the center of the nucleus, whereas larger, gene-poor chromosomes are located near the nuclear periphery. The positioning of chromosome territories also correlates with cell type-specific factors such as replication timing and transcriptional activity: early-replicating loci and active genes tend to localize deeper inside the nucleus, whereas late-replicating loci and repressed genes have a preference for nuclear periphery[90,91].

In summary, the regulation of gene expression at transcriptional level is a dynamic, combinatorial process involving a variety of elements and mechanisms that may only operate in particular cell types, at a given stage in development or in response to environmental factors. Therefore, any alteration in these mechanisms can result in functional consequences involving changes in gene expression, that can be finally translated into disease phenotypes.

1.3. Identification of transcriptional regulatory elements

1.3.1. Analysis of transcription factor binding sites

The binding of TFs to *cis*-regulatory elements requires DNA to be accessible within the chromatin structure. This is achieved by changes in the chromatin condensation and nucleosome displacement. Exposed DNA regions are called open chromatin regions and they have been associated to active transcriptional regulatory regions of the genome[92–94].

DNA exposure facilitates the binding of TFs, but also makes DNA more sensitive to enzymatic cleavage. This property has been long used to identify open chromatin regions by digesting the chromatin with the DNase I nuclease. DNase I can cut through open chromatin regions–although open regions with TFs bound will have a mild protection against digestion– (**Figure 18A**). For this reason, open chromatin regions are also referred to as **DNase I Hypersensitive Sites** (**DHS**).

Currently, DNase I digestion is usually coupled with massively parallel sequencing of the digested material (**DHS-seq**), which allows the identification of all open chromatin regions found in the genome for a specific cell-type or tissue[95,96]. After the subsequent bioinformatic analysis of the sequenced fragments, DHS can be visualized as peaks corresponding to the amount of fragments sequenced in a particular genomic position (**Figure 18B**).

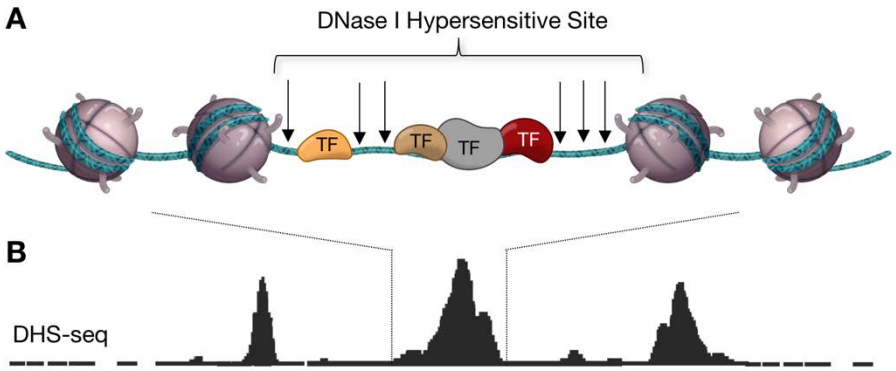


Figure 18. Open chromatin regions. (A) Schematic representation of DNase I hypersensitivity sites with TFs bound. Black arrows indicate regions susceptible to DNase I digestion. **(B)** Overview of the peak tracks obtained after processing DHS-seq data. Figure adapted from Ecker *et al*.[4].

DHS-seq identifies active transcriptional regulatory regions, which is indirectly signaling TF binding regions. However, DHS-seq does not have enough resolution to detect which TF is binding or the exact binding site that is being used. Therefore, the combination of DHS-seq with computational algorithms that perform sequence motif analysis such as HOMER[97] has been proved to be very useful to identify TF binding motifs in the tested sequences (**Figure 19**). This motif analysis strategy can also be applied to the whole genome. However, as mentioned before, not all motifs found in the genome are actually bound by TFs. For example, we can identify a TF motif but it can be located in a condensed region because it might not be required in the cell-type or tissue analyzed (**Figure 19**).

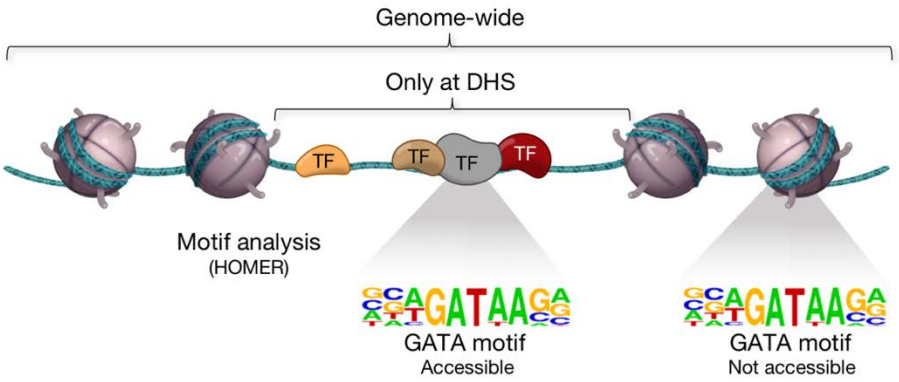


Figure 19. Identification of TF binding sites by motif analysis. Example of a motif analysis that, when performed genome-wide, identifies two GATA motifs but only one of them is accessible to GATA TFs. In contrast, when motif analysis is exclusively applied to DHS, it only identifies the accessible GATA motif. Figure adapted from Ecker *et al.*[4] and motifs extracted from HOMER[97].

If the experimental interest is to profile TF binding in a given cell-type or developmental stage, the most widely used approach is the **chromatin immunoprecipitation followed by massively parallel sequencing** (**ChIP-seq**)[98].

Briefly, ChIP-seq starts with the fixation of the TF-DNA interactions followed by a fragmentation of the chromatin. Fragmented chromatin is incubated with antibodies against the TF of interest to only immunoprecipitate those genomic regions that are bound by the TF of interest. Immunoprecipitated material is then sequenced and, after downstream bioinformatic analysis, TF binding profiles distributed along the genome can be visualized as peaks corresponding to the amount of fragments sequenced in a particular genomic position (**Figure 20**).

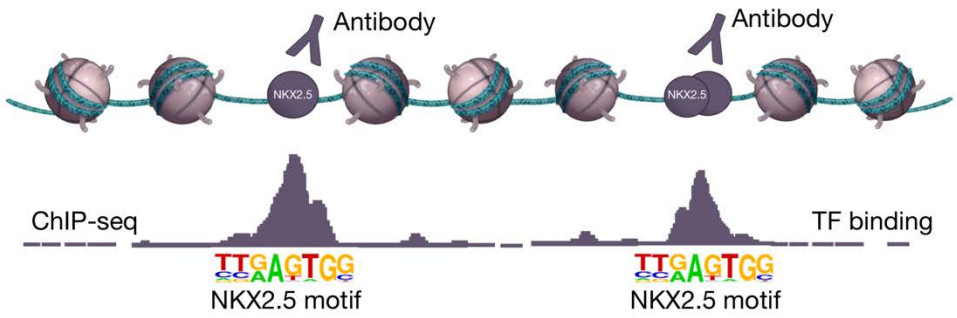


Figure 20. Profiling of TF binding sites by ChIP-seq. Top: schematic representation of TFs bound to chromatin, recognized by specific antibodies. **Bottom:** overview of the peak tracks obtained after processing ChIP-seq data. Figure adapted from Ecker *et al.*[4] and motifs extracted from HOMER[97].

1.3.2. Analysis of histone modifications

As previously mentioned, one of the mechanisms that influence DNA accessibility to TFs is the presence of histone modifications that alter DNA-histone interactions within the nucleosome. These modifications are distributed throughout the genome and it has been observed that particular patterns of histone modifications can be associated to different types of transcriptional regulatory elements. For example, **trimethylation of the lysine 4 of histone 3** (**H3K4me3**) is associated to active promoters[99,100]; and acetylation of lysine 27 of histone 3 (H3K27ac) is associated to active enhancers[100,101]. The patterns of histone modifications can be also detected by ChIP-seq using antibodies against the histone modification of interest, and the results are also represented as peaks corresponding to the amount of regions sequenced in a particular genomic position (**Figure 21**).

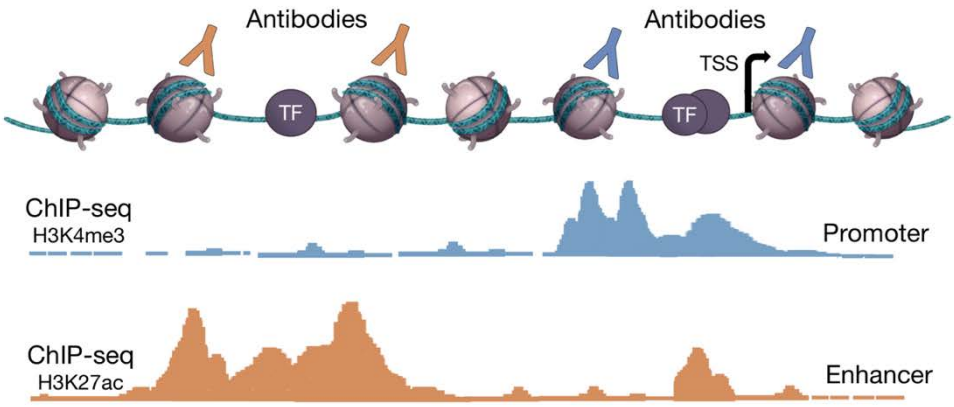


Figure 21. Profiling of histone modifications by ChIP-seq. Top: schematic representation of chromatin with some TFs bound and a TSS. Antibodies recognizing H3K4me3 are used to identify active promoter regions, while antibodies recognizing H3K27ac are used to identify active enhancers. **Bottom:** overview of the peak tracks obtained after processing ChIP-seq data. Figure adapted from Ecker *et al.*[4].

1.3.3. Analysis of the tridimensional organization of the genome

Transcriptional regulatory elements can also be identified using high-throughput chromatin capture techniques, also referred to as C-technologies. These techniques allow the identification of chromatin interactions corresponding to DNA loops between *cis*-regulatory elements, but they are also useful to uncover general features of genome organization such as cell compartments and TADs[102]. Therefore, depending on the experimental purpose, there are several C-technologies available. For example, the 4C technique is applied to interrogate all the regions of the genome that interact with a particular region of interest[103]. In contrast, the **Hi-C** is applied to obtain the interaction profiles of all regions, genome-wide, and their interactions with all other genomic regions[77]. When the assessment of chromatin contacts is performed genome-wide such in Hi-C experiments, C-data is represented as chromatin contact maps where the color intensity is related to the interaction frequency (**Figure 22A**).

Chromatin contact maps can be further complemented with ChIP-seq information for CTCF, to identify TAD boundaries; or DHS or histone modifications, to identify active and inactive regions inside TADs (**Figure 22B**).

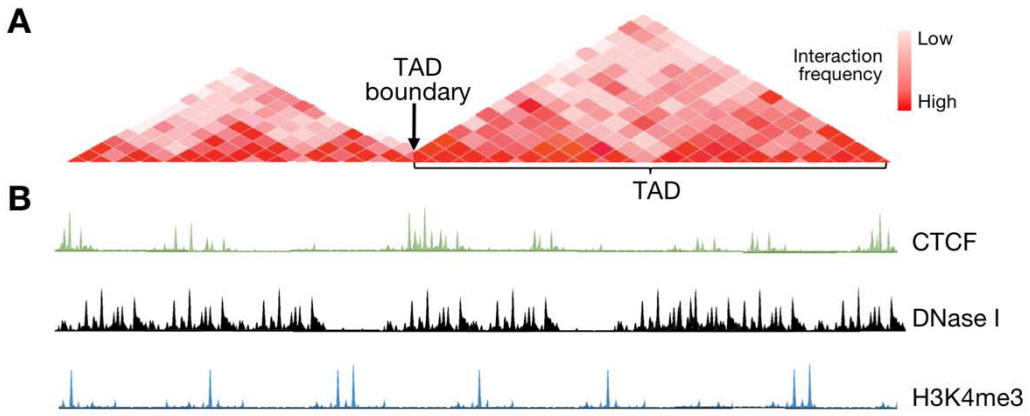


Figure 22. Analysis of genome organization. (A) Schematic representation of Hi-C contact maps indicating the frequencies of interactions between genomic regions. **(B)** Peak tracks showing CTCF-binding sites, open chromatin regions (DHS) and promoters (H3K4me3).

In summary, any of the methods aforementioned can be used to identify transcriptional regulatory elements and their interactions. Their utilization and combination with each other will depend on the experimental purposes. However, it is important to know that, in addition to DHS-seq, ChIP-seq and Hi-C, other methods have been designed for the same purposes but more emphasis has been given to the aforementioned methods because of their importance in this thesis.

2. Genetic variation

Genetic variation refers to the differences in the DNA sequence found in the genome of organisms. These genetic variants are randomly distributed throughout the genome and, in the case of diploid genomes, they can be found as heterozygotes (if they are only affecting one of the two alleles), or as homozygotes (if they are affecting both alleles).

2.1. Types of genetic variation

Generally speaking, genetic variants are classified into three main groups depending on the extent of the nucleotides that are affected. These include: (i) variants affecting a single nucleotide (SNVs)[104], (ii) small insertions or deletions of less than 50 bp in length (indels)[105], and (iii) larger forms of structural variation (SVs) of more than 50 bp in length[106].

The analyses performed in this thesis are only focused on the detection of SNVs and indels thus, all the information included in the sections below will refer to these two types of genetic variants.

2.1.1. Single nucleotide variants

SNVs are the most common type of genetic variant observed in the genome, and are referred to as single nucleotide polymorphisms (SNPs) when they are found at a high frequency among the population (allele frequency; AF > 0.01).

As it will be explained in section 2.4, the less frequent SNVs (AF < 0.01) are more difficult to be detected because they require deep sequencing of multiple genomes. Therefore, it has only been possible to estimate the frequency of SNPs in the human, which accounts for 1 SNP every 800-1000 bp[107].

SNVs occur when a single nucleotide is substituted for another at a given position in a DNA sequence (**Figure 23**). If the nucleotide substitution involves a purine to purine (A ↔ G) or pyrimidine to pyrimidine (T ↔ C) exchange, SNVs are referred to as transitions. On the other hand, if the substitution involves purines and pyrimidines or vice versa, SNVs are referred to as transversions[104].

Figure 23. Single Nucleotide Variants. Two different examples of SNVs representing a transition (purine-purine or pyrimidine-pyrimidine) and a transversion (purine-pyrimidine).

2.1.2. Small indels

Indels consist of an insertion or deletion of one or more DNA nucleotides into the genome (**Figure 24**). They represent the second most common type of genetic variation, although the estimated frequency in the human genome varies from one study to an other[108].

The identification and annotation of indels can be difficult, especially for complex indels that include both inserted and deleted nucleotides at the same time. Moreover, indels often involve areas with repetitive sequences that further complicate the determination of their exact extent[109]. All these complexities make more difficult the determination of the exact genotype and localization of indels in the human genome.

Figure 24. Indels. Two different examples of indels representing an insertion of 2 nucleotides and a deletion of 2 nucleotides.

2.1.3. Structural variants

The term SV is typically used to describe genetic variation that occurs over a large DNA sequence, from 50 to thousands of bp. They are less frequent than SNVs or indels but they affect a higher number of nucleotides[106]. SVs can be found as changes in the copy number (deletions, insertions and duplications), orientation (inversions) or chromosomal location (translocations) between individuals[110] (**Figure 25**).

Independently of their extent (from intermediate size to large changes that involve entire chromosomes), SVs can be classified as: (i) balanced, if there is no loss or gain of genetic material (inversions and translocations) and; (ii) unbalanced, if apart of the genome is lost or duplicated (deletions, insertions and duplications)[110].

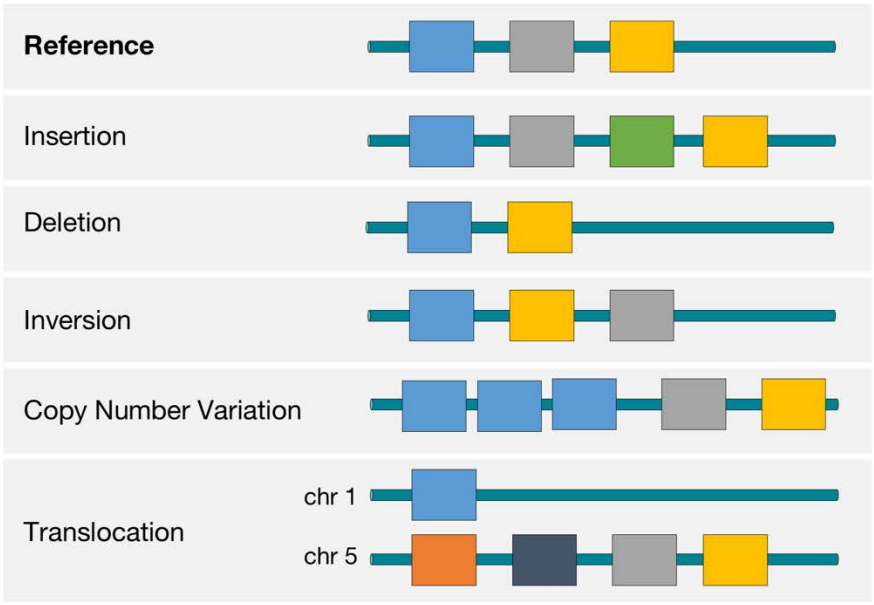


Figure 25. Structural Variants. Depiction of different types of structural variants compared to the reference genome. Each different box represents a different genomic element.

2.2. Sources of genetic variation

As previously defined, genetic variation refers to differences in the DNA sequence found in the genome of organisms. These differences can be originated by: (i) exogenous sources such as environmental factors or, (ii) endogenous sources such as cell metabolism products, errors in DNA replication and failure of DNA repair mechanisms[104,109]. Additionally, DNA recombination during meiosis and transposable elements also constitute two different sources of endogenous variation that may lead to indels[109].

Endogenous DNA damage is more frequent than exogenous damage, although the types of alterations produced are very similar[111].

2.2.1. Environmental factors

Broadly speaking, there are two main groups of environmental factors that can generate genetic variants. These include chemical products and radiation.

In the case of chemical products, there is almost an infinite amount of chemicals that can alter the DNA sequence via several mechanisms. For example, base analogs, which are similar to DNA nitrogenous bases, can be inserted into DNA (**Figure 26A**). These analogs have different pairing properties than those of the normal bases, and may lead to incorrect insertion of nucleotides in the opposite strand during DNA replication[112].

Radiation-induced DNA damage can be classified into two general categories: damage caused by ultraviolet radiation (UV light) and damage caused by ionizing radiation. UV sunlight induces covalent bonds between adjacent thymine residues in the same strand of DNA, producing bulky intra-strand TT pyrimidine dimers (**Figure 26B**). If unrepaired, these lesions promote SNVs through errors during DNA replication[113–115]. On the other hand, ionizing radiation such as X-Rays can induce a broad spectrum of DNA damages including base lesions, crosslinks, and double-stranded breaks (**Figure 26C**)[113–115].

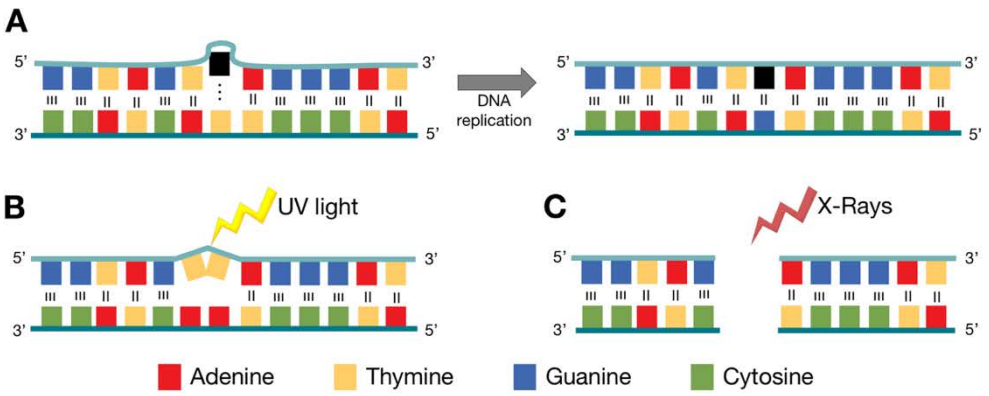


Figure 26. DNA damage caused by environmental factors. (A) Insertion of a base analog that ends with the addition of a mispaired guanine instead of the original thymine. **(B)** Formation of a TT-pyrimidine dimer caused by the incidence of UV light. **(C)** Double-stranded break produced by the incidence of X-Rays. Figure adapted from Khan Academy courses[116].

2.2.2. Cell metabolism

Genetic variants can also be generated by hydrolysis, exposure to reactive oxygen substances (ROS) and other reactive metabolites that are produced in the routine metabolism of the cell[104]. These damaging mechanisms by which genetic variants will be introduced are different depending on the type of metabolite generated. For example, the generation of 8-oxodeoxyguanosine as an oxidation product represents a major cause of genetic variation during replication. 8-oxodeoxyguanosine is the equivalent of guanine but it can pair with adenine almost as efficiently as cytosine, causing G-C to T-A transversions[104].

2.2.3. DNA replication errors

Cell cycle is defined as the period that takes place between successive divisions of a cell. It is divided into four main phases (M, G1, S and G2). During the S phase, and prior to cell division, DNA is replicated by a group of enzymes called DNA polymerases.

In human cells, DNA polymerases involved in replication are highly accurate, with an error rate of 10^{-6} to 10^{-7} errors per bp. This error rate is further diminished by DNA repair mechanisms that lower the net rate to approximately 10^{-10} errors per bp[104]. However, even with this high level of DNA replication fidelity, some errors in DNA replication are incorporated in the final DNA sequence, thus originating genetic variants.

In the case of SNVs, replication errors are produced by the addition of a non-complementary base pair by DNA polymerase (**Figure 27A**). In contrast, indels can be originated by two different replication errors: (i) polymerase slippage and (ii) secondary structure formation during replication. Polymerase slippage occurs when the DNA polymerase and the newly synthesized DNA strand temporarily dissociate from the template DNA. Especially in areas with repetitive sequences, the polymerase may re-associate with the template strand in a position ahead or behind of where it left off– introducing a deletion or an insertion, respectively– (**Figure 27B**). Secondary structures can be formed during DNA replication of regions with inverted repeats. These secondary structures can induce DNA polymerase slippage to another region after the secondary structure, resulting in the introduction of a deletion[109] (**Figure 27C**).

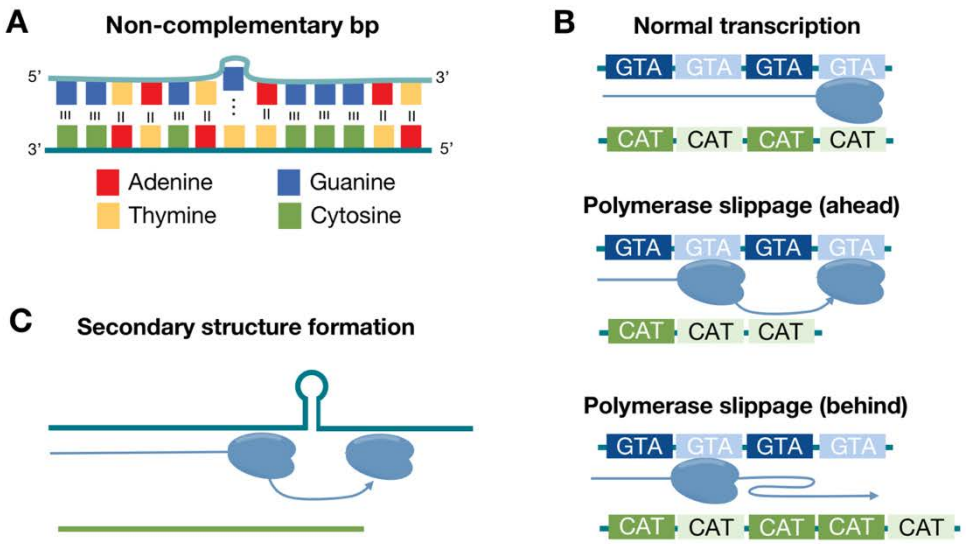


Figure 27. DNA replication errors resulting in SNVs and indels. **(A)** SNV originated by a non-complementary base incorporated by DNA polymerases. **(B)** Indel formation produced by DNA polymerase slippage during replication of repetitive nucleotides. The top sequence (blue) corresponds to the template DNA while the bottom sequence (green) corresponds to the newly synthesized sequence. **(C)** Indels originated by secondary structures that induce a polymerase slippage for a position ahead of the secondary structure. Figure adapted from Khan Academy courses[116].

2.2.4. DNA repair mechanisms

In eukaryotic cells, DNA repair mechanisms are activated to correct alterations in the DNA sequence and maintain its integrity. DNA repair mechanisms are included as a source of genetic variation because, even they seek to restore the DNA sequence when an alteration is introduced, their activity can also result in the insertion of genetic variants.

Simplistically, DNA repair mechanisms can be classified into two main groups. The first group includes all those mechanisms intended to correct DNA replication errors (proofreading and mismatch repair). The second group includes all those mechanisms intended to correct DNA damage induced by environmental factors or cell metabolites (base excision repair, nucleotide excision repair and double-stranded break repair)[116].

A) Proofreading

During DNA replication, several DNA polymerases have the ability to detect incorrectly paired nucleotides. If the polymerase detects that a mispaired nucleotide has been introduced, it will replace the nucleotide before proceeding with the synthesis of DNA (**Figure 28A**)[117]. Proofreading is highly accurate and fixes about 99% of DNA replication errors, but sometimes mispaired nucleotides are unnoticed and therefore are introduced in the DNA sequence as genetic variants[116].

B) Mismatch repair

Immediately after DNA replication, mismatch repair mechanisms are activated to correct either errors caused by mispaired nucleotides that skipped proofreading, or small indels that were originated due to polymerase slippage or secondary structure formation (**Figure 28B**). Incorrect nucleotides are recognized and bound by a protein complex that cuts the DNA near the mismatch, removing the mispaired nucleotide as well as its neighbors. A DNA polymerase replaces the missing DNA sequence with the correct nucleotides and then, an enzyme called ligase seals the gap[118].

The mismatch repair reduces the final DNA replication error rate. However, some sequence alterations might remain unnoticed or might be incorrectly repaired. All those sequence alterations that still remain following mismatch repair become permanent genetic variants after the next cell division because, once such errors are established, the cell no longer recognizes them as errors[119].

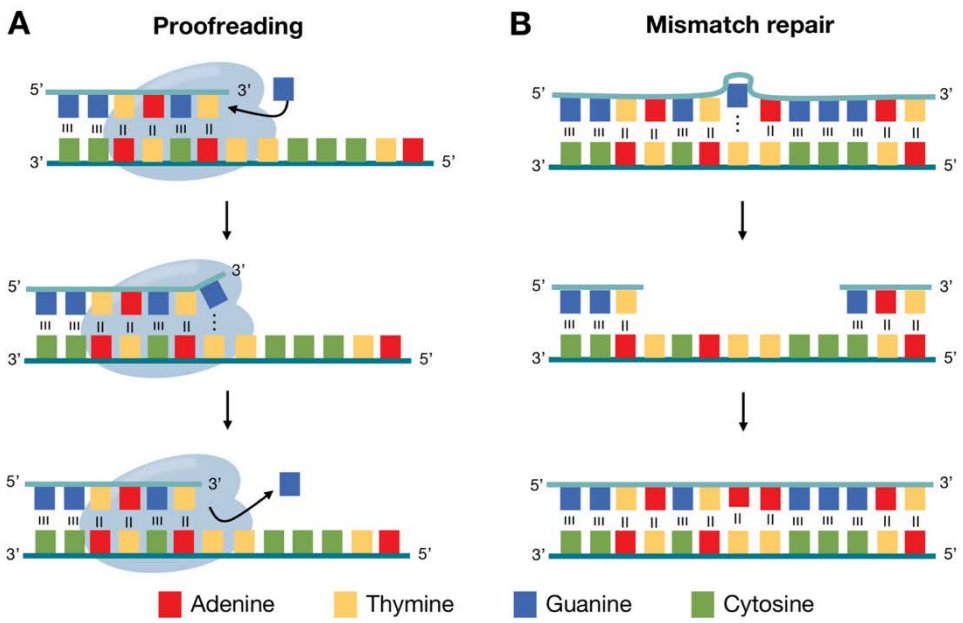


Figure 28. DNA repair mechanisms to correct replication errors. (A) The DNA polymerase introduces an incorrect nucleotide to the new DNA strand. Due to its proofreading property, the polymerase detects that the inserted nucleotide is mispaired and removes it. **(B)** If a mismatch is detected in the newly synthesized DNA, the new strand is cut and the mispaired nucleotide as well as its neighbors are removed. The missing sequence is replaced with correct nucleotides by DNA polymerase and the ligase seals the gap. Figure adapted from Khan Academy courses[116].

C) Base excision repair

Base excision repair is a mechanism used to detect and remove certain types of damaged nucleotides. For example, a chemical reaction called deamination can convert a cytosine base into uracil, a base typically found only in RNA. To prevent such DNA alterations, a specific enzyme detects and removes deaminated cytosines leaving a nucleotide gap. This gap will be further filled by DNA polymerases and then, a ligase will seal the backbone[119] (**Figure 29A**). However, if the deaminated cytosines are not detected or the DNA polymerase introduces a mispaired base, the result will be the introduction of a genetic variant that will become permanent in the genome.

D) Nucleotide excision repair

Nucleotide excision repair is a mechanism used to remove and replace damaged nucleotides that distort the DNA double helix such as intra-strand TT pyrimidine dimers induced by UVB radiation. Once the TT pyrimidine dimers are detected, a group of enzymes cut the dimer and the surrounding DNA, leaving a gap. Similarly to previously mentioned mechanisms, the gap will be filled by a DNA polymerase and the backbone sealed by a ligase[104] (**Figure 29B**). In this case, as in the base excision repair mechanism, if the TT pyrimidine dimer is not detected or the DNA

polymerase introduces a mispaired nucleotide, the introduced genetic variants will become permanent in the genome.

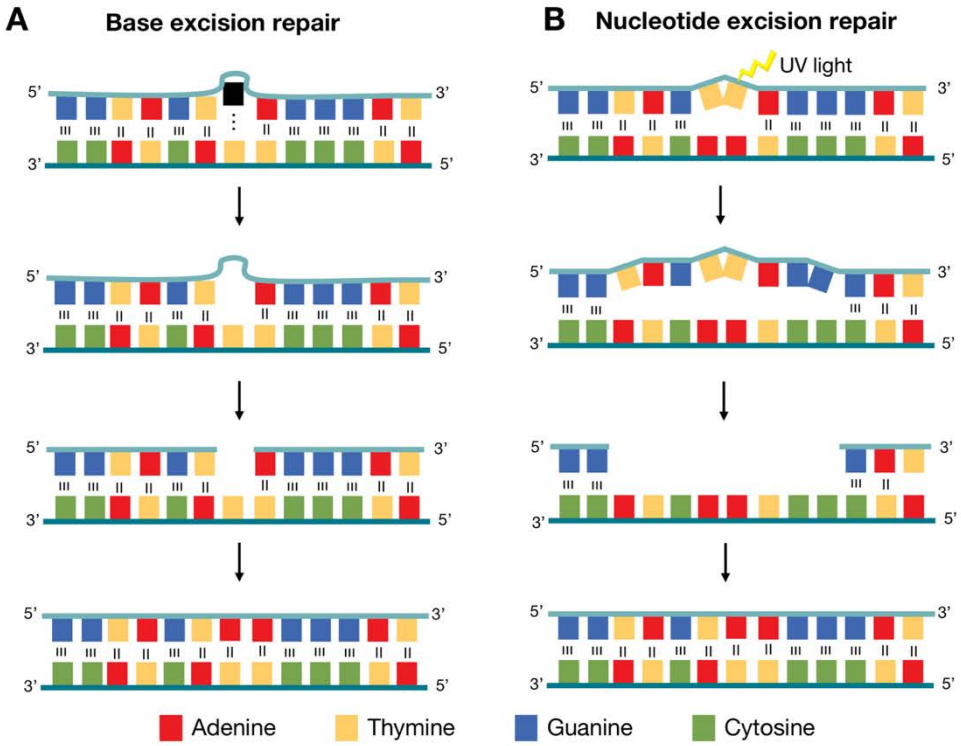


Figure 29. DNA repair mechanisms to correct damaged nucleotides. (A) Example of base excision repair mechanism that removes deaminated cytosines, leaving a gap that is filled by DNA polymerases. **(B)** Example of nucleotide excision repair mechanism that removes intra-strand TT pyrimidine dimers induced by UVB radiation and surrounding nucleotides, leaving a gap that is filled by DNA polymerases. Figure adapted from Khan Academy courses[116].

E) Double-stranded break repair

Some types of environmental factors such as high-energy radiation can cause double-stranded breaks in the DNA. Double-stranded breaks are dangerous because large segments of chromosomes, and the genes they contain, may be lost if the break is not repaired. There are two different pathways involved in double-stranded DNA breaks repair: the non-homologous end joining and the homologous recombination (**Figure 30**).

In non-homologous end joining, the two broken ends of the break are directly reunited. This mechanism avoids loosing genetic material but it typically involves the loss or addition of a few nucleotides at the breakpoint, giving rise to small indels[119]. In contrast, in homologous recombination, the missing DNA is copied from the homologous sequence, resulting in a sequence that is identical to the original[119].

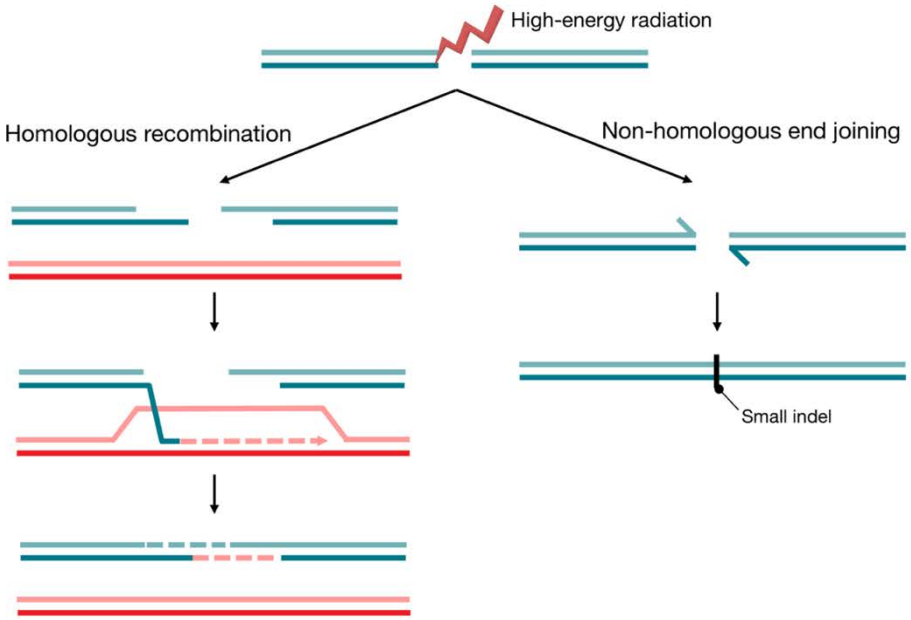


Figure 30. DNA repair mechanisms to correct double-stranded breaks. Representation of the two possible mechanisms that can be activated to repair double-stranded breaks produced by ionizing radiation: the homologous recombination (left) and non-homologous end joining (right). Figure adapted from Khan Academy courses[116].

2.2.5. Mobile DNA elements

Mobile elements dispersed over the genome (transposons and retrotransposons) are considered an important source of SVs, but they can also contribute to the formation of small indels. These small indels are formed because, when mobile elements are excised from their genomic position, they leave a signature similar to that of double-stranded breaks. Therefore, the double-stranded break repair mechanism will be activated and, if the break is repaired by non-homologous end joining, small indels might be introduced[18].

2.2.6. Meiotic recombination

In diploid organisms such as humans, in which the cells have two copies of each chromosome, sexual cells undergo a process named meiosis. Meiosis is an especialized cell division that reduces the chromosome number by half, originating four haploid cells. This chromosome number reduction is important because in sexual reproduction, when the maternal egg is fertilized with the paternal spermatozoid, the resulting zygote will have the diploid composition restored[3].

Before meiosis begins, during S phase of the cell cycle, the DNA of each chromosome is replicated. Immediately following DNA replication, meiotic cells enter a stage known as meiotic

prophase I. During this period, homologous chromosomes pair with each other and undergo genetic recombination–a process named crossover–, in which there is DNA information exchange between the two homologous chromosomes (**Figure 31**)[3].

In general, meiotic recombination does not involve the introduction of SNVs or indels; however, when meiotic recombination involves misaligned homologous chromosomes, small indels can appear[109].

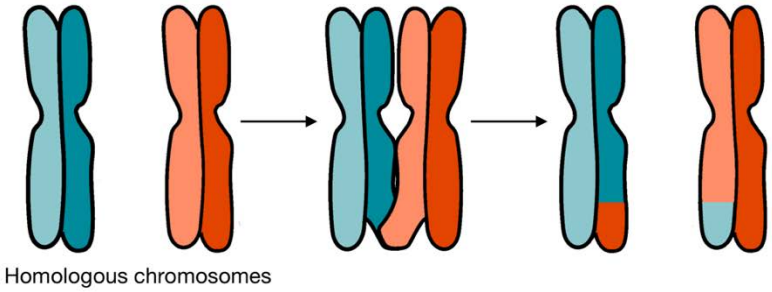


Figure 31. Crossover. Schematic representation showing the pairing of homologous chromosomes during meiotic prophase I, resulting in an exchange of genetic material between non-sister chromatids. Figure adapted from Khan Academy courses[116].

2.3. The role of genetic variation in human disease

Human diseases can be caused by genetic variants that encompass the full range of variant types, from SNVs to SVs. These variants span a broad frequency spectrum, from the very rare to the common[120] (**Figure 32**). Common variants refer to those variants found in a relatively high frequency in a population (AF > 0.01). They are thought to be originated from old variants that arose in a common ancestor and that have been selected until they reached a significant frequency among the population[121]. In contrast, rare variants refer to those variants found in a relatively low frequency in a population (AF < 0.01). These variants have been generally associated either to recent variants where the selection did not have time to act, or variants that are being selected against due to their deleterious nature[121].

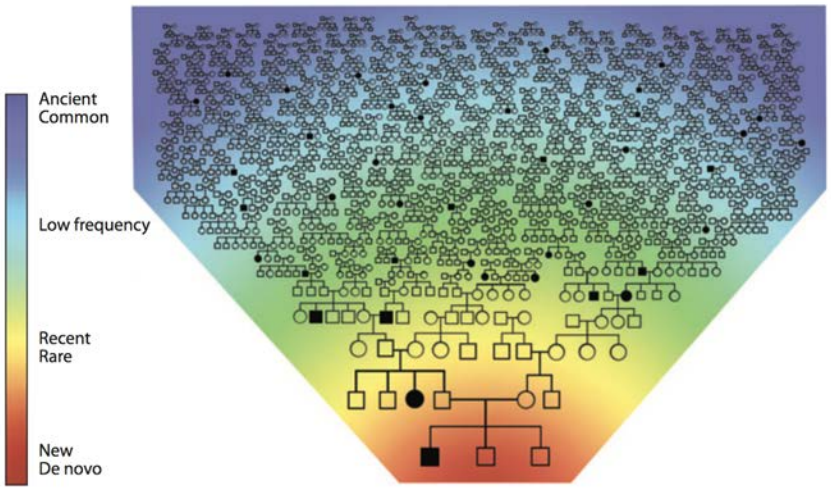


Figure 32. Frequency spectrum of human genetic variation. Example of a family tree that expands backward in time to distant ancestors. The heat map indicates that the frequency of each variant is related to its age, increasing from bottom to top. Figure extracted from Lupski *et al.*[120].

When studying human genetic diseases, genome-wide association studies (GWAS) have provided valuable information. GWAS aim to unravel those variants associated to a specific disease or trait by comparing common variants (SNPs) in diseased individuals versus genetic variants found in healthy individuals. These studies have let to the identification of >1,200 genomic regions harboring genetic variants associated with >165 human diseases and traits. However, even this impressive information gleaned from GWAS, the results obtained only explain a few percent of the apparent genetic variation contributing to human diseases[122]. These observations could be related to the fact that GWAS are focused on the detection of common variants, suggesting that part of these so-called missing heritability could be elucidated, to some extent, by rare variants[122].

However, detection of rare variants, has been particularly challenging because, in order to properly determine the frequency and population distribution of these genetic variants, highly accurate sequencing data needs to be generated from many samples[120]. For example, the HapMap project[123–125] that provided an early survey of single-base variation across major human populations, only cataloged a fraction of the genetic variation above an AF of 0.05. Even the 1000 Genomes Project pilot studies comprehensively captured only variation at greater than an AF of 0.01[126].

As these and other projects have been growing and more genomes have been sequenced, it is becoming apparent that genetic variation between individuals is greater than was previously expected[127–134]. For example, the sequencing of >2,500 individuals from 5 populations (African, American, East Asian, European and South Asian) by the 1000 Genomes Project revealed that, when compared to the reference genome, each individual genome contains on average 3.5

million SNPs and >2,100 SVs. Moreover, many of these variants appear to be rare in the population from which the individual was sampled, and the identification rate of variants that have not been previously described continues growing with every new individual sequenced[135,136].

From these studies, it was also observed that the total number of variants identified greatly differs among populations (**Figure 33**). Although most common variants are shared across the world, rare variants are typically restricted to closely related populations. Individuals from African ancestry populations harbor the greatest number of variant sites, as predicted by the out-of-Africa model of human origins. In contrast, individuals from recently admixed populations show great variability in the number of variants, roughly proportional to the degree of recent African ancestry in their genomes[107]. For this reason, when studying human disease, it is very important to take into account the population of origin of the individuals studied. The same variant can be common in one population but rare in another population.

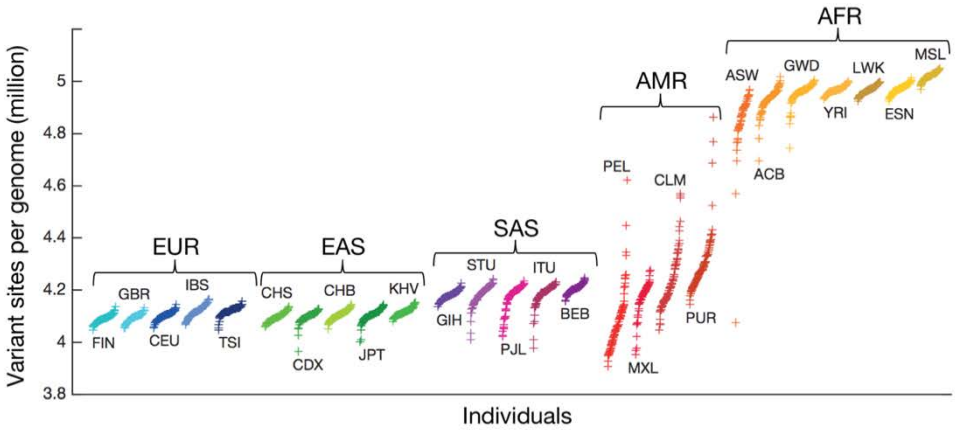


Figure 33. Distribution of variants among human populations. Representation of the number of variants sites per genome classified according to the population of origin. Each color palette corresponds to one of the 5 super-populations (EUR, Europeans; EAS, East Asians; SAS, South Asians; AMR, Americans; AFR, Africans). Figure adapted from The 1000 Genomes Project Consortium[107].

Together, the high proportion of variants present in the genome and the fact that some of the variants are private to each individual genome, suggests that the most important aspect when studying human genetic disease is not to focus disproportionately on specific variants, but rather to integrate across all classes of risk-associated variants. In some individuals, genetic disease might be caused by an unusual combination of common variants, whereas in others it might be caused by a smaller number of rare variants, and each of them can have a range of small to major effects[120] (**Figure 34**).

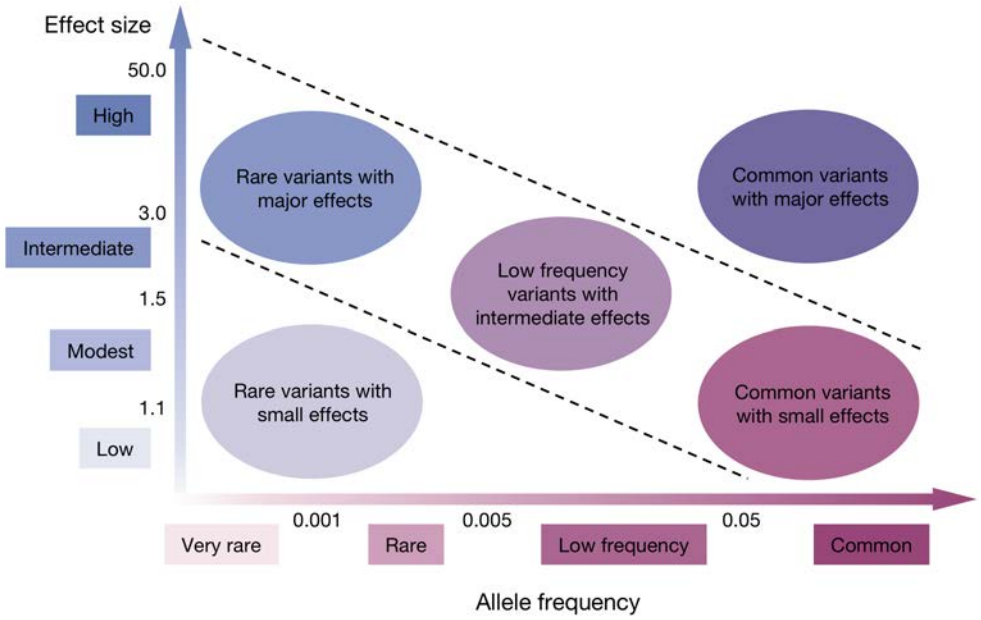


Figure 34. Disease-risk allele frequencies and effect size. Relation between risk allele frequency and its effects in human disease (odds ratio). Figure adapted from Manolio *et. al.*[137].

2.4. Examples of non-coding variants associated to human disease

The functional impact of non-coding variants still remains a challenge, but the recent progress in the study of non-coding variants has allowed the identification of a considerable number of non-coding variants associated to several diseases.

For example, when studying colorectal cancer, Katainen *et al.,*[138] found that microsatellite-stable colorectal cancer tumors presented a higher accumulation of genetic variants at CTCF-Cohesin binding sites than it was expected. Transcriptional enhancers usually interact with their target genes through the formation of DNA loops, which are typically constrained within larger CTCF-Cohesin-mediated loops. Therefore, this accumulation of genetic variants at CTCF-Cohesin binding sites observed in Colorectal Cancer may disrupt the CTCF-Cohesin-mediated loops, causing aberrant gene expression that may drive tumorigenesis.

Other examples of non-coding variants associated to disease were identified by Hniz *et al.,*[139] when studying T-cell acute lymphoblastic leukemia (T-ALL). In their case, they found that several recurrent deletions in T-ALL patients resulted in an aberrant gene expression of two different proto-oncogenes, the *TAL1* and *LMO1.* These two proto-oncogenes are silenced by the same mechanism of insulation in a CTCF-Cohesin-mediated loop. However, the deletions found in T-ALL patients were removing two different CTCF-Cohesin binding sites involved in the insulation of *TAL1* and *LMO1.* The loss of insulation resulted in *TAL1* and *LMO1* overexpression, which is highly associated to T-ALL phenotypes.

The two aforementioned examples represent non-coding variants with pathogenic effects. However, not all non-coding variants always have a negative effect and, sometimes, they can be involved in the protection against disease. In this regard, a clear example was observed by Wang *et al.,*[140] when analyzing the presence of disease-associated GWAS at antioxidant response elements bound by NRF2/sMAF TFs. In their analysis, the authors found a GWAS variant (rs242561) that was affecting the binding of NRF2/sMAF at an ARE active in brain tissue. Interestingly, the presence of the T allele was increasing the binding of NRF2/sMAF compared to the C allele, which results in an increased expression of MAPT Tau protein. The MAPT Tau protein has seven major isoforms formed by alternative splicing of exons 2, 3 and 10, and the T allele has been associated to higher levels of inclusion of exon 3. The inclusion of exon 3 results in a type of Tau protein with less propensity for aggregation and amyloid-beta toxicity. Therefore, individuals displaying the T alle will have a reduced risk of developing diseases caused by aggregation of Tau protein such as Parkinsonian Disorders.

In summary, human genomes contain a considerable number of genetic variants which, in most cases, have neutral effects. The identification of functionally relevant variants among the pool of neutral variants is complex and several computational approaches have been designed to help in this purpose, especially for the identification of functionally relevant non-coding variants. Due to scientific and technological advances in the field of human genetics, the number of non-coding variants that have been linked to diseased phenotypes–either increasing or decreasing the risk of disease–continue to grow, which encourages to continue with the study of non-coding regions and their function.

2.5. Identification of genetic variants using Illumina sequencing

High-throughput detection of genetic variants is currently possible due to the newly developed sequencing platforms that allow the processing of multiple DNA sequences at the same time (known as massively parallel sequencing). The most used technology worldwide is the Illumina sequencing by synthesis, which has also been used in this thesis. Basically, the Illumina sequencing workflow consists of four basic steps: (i) DNA library preparation, (ii) cluster generation, (iii) sequencing and, (iv) data analysis.

2.5.1. DNA library preparation

Illumina sequencing can be used for several applications such as sequencing of entire genomes (whole-genome sequencing; WGS), sequencing of exons (whole-exome sequencing; WES) or sequencing of a limited number of specific regions (targeted sequencing). Depending on the final application, library preparation workflow will lightly differ, but the basics are shared by all of them.

Library preparation includes a fragmentation of genomic DNA and indexing of the fragments obtained. Indexes are used to tag all the DNA fragments from the same sample, allowing the multiplexing of samples from different individuals during the same sequencing run. In addition to indexes, adaptor sequences at the 5' and 3' ends of the DNA fragments are also added during library preparation. These adaptors are required to attach the DNA fragments to the flow cell, which is a glass slide with lanes where the sequencing takes place. The attachment of the DNA fragments to the flow cell is mediated by two different types of oligonucleotides fixed to the flow cell surface (**Figure 35**).

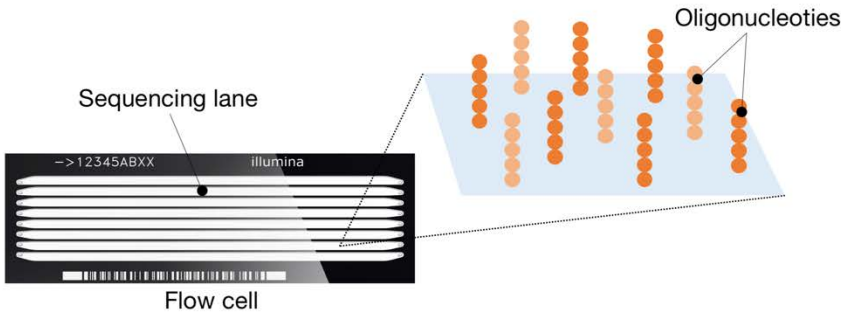


Figure 35. Schematic representation of an Illumina flow cell. Each channel of the flow cell is referred to as lane and it is where the sequencing takes place. Each lane is coated with pairs of oligonucleotides that will be used for DNA fragments to attach to the flow cell.

2.5.2. Cluster generation

After the DNA libraries are loaded on the Illumina sequencing platform, cluster generation takes place. First, DNA library fragments are denatured, and only hybridize with one of the two types of oligonucleotides attached to the flow cell, ensuring that all fragments have the same orientation. Attached fragments are then amplified by a DNA polymerase, and the original DNA fragments are washed away[143].

The adaptor of the free end of the newly synthesized DNA fragment hybridizes with the other type of oligonucleotide attached to the flow cell, forming a bridge (**Figure 36**). Then, polymerases generate the complementary strand, forming a double-stranded bridge. The bridge

is denatured, leaving two single-stranded copies of the molecule that are tethered to the flow cell. The process is repeated ~35 cycles and occur simultaneously for millions of clusters, resulting in a clonal amplification of all the fragments. After bridge amplification, reverse strands are cleaved and washed-off, leaving only the forward strands attached to the cell[143].

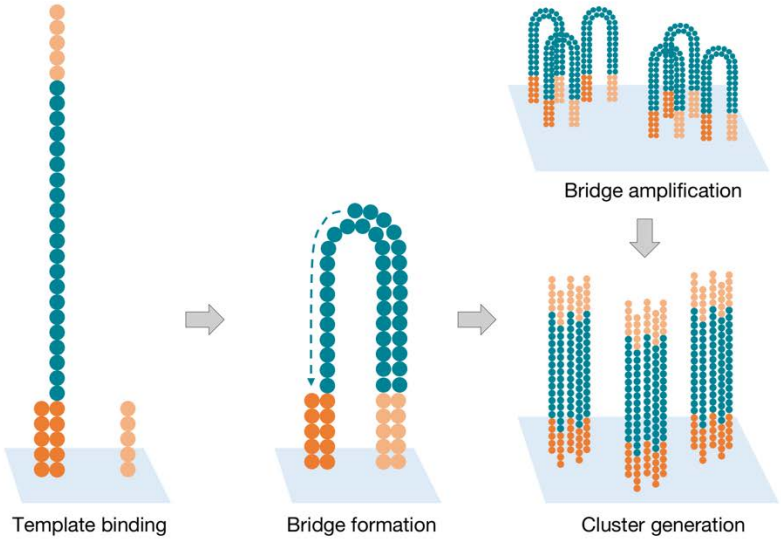


Figure 36. Cluster generation. Schematic representation of the typical bridge PCR-amplification from Illumina platforms that result in the formation of millions of clusters. Figure adapted from Illumina® tutorials (www.illumina.com).

2.5.3. Sequencing

Illumina sequencing is based on a sequencing by synthesis with fluorescently tagged nucleotides. At each cycle, fluorescent nucleotides are introduced, but only the complementary nucleotide will be incorporated to the growing chain. After the addition of each nucleotide, clusters are excited by a light source and a characteristic fluorescent signal is emitted (**Figure 37**). The emission wavelength, together with the signal intensity, determine the nucleotide that has been added (process named base call). It is important to note that, for a giving cluster, all identical stands are read simultaneously and that hundreds of millions of clusters are sequenced in a massively parallel process, generating millions of sequences at the same time[143].

The number of sequencing cycles applied is directly related to the final length of the reads, with a maximum of 150 cycles for short-read sequencing–the same used in this thesis–.

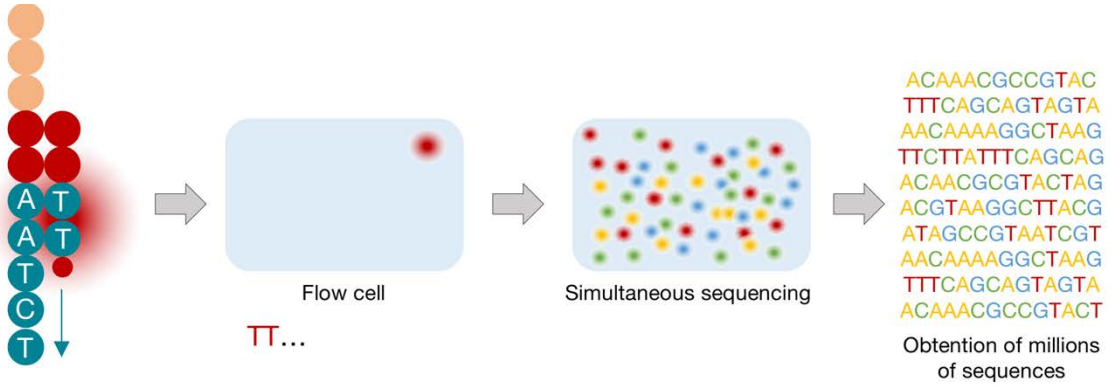


Figure 37. Illumina sequencing by synthesis. At each sequencing cycle, the four fluorescently labelled oligonucleotides are added. When the complementary nucleotide is introduced, the clusters are excited and the characteristic fluorescent signal emitted by the nucleotide incorporated is used to determine which nucleotide has been added. The process is repeated multiple times and simultaneously to all clusters, which allows the obtention of millions of sequences in the same sequencing run. Figure adapted from Illumina® tutorials (www.illumina.com).

In the case of paired-end sequencing, the sequencing continues by washing away the forward read and allowing the DNA template to fold over and bind to the second oligonucleotide on the flow cell. Polymerases extend the bridge forming a double-stranded bridge. This double-stranded DNA is linearized and the original forward strand is washed away, leaving the reverse strand. The remaining DNA molecule is sequenced similarly to the first read, originating the reverse read[143].

All reads obtained from multiplexed sample libraries are separated based on the unique index sequences introduced during library preparation and stored in Fastq files, which contain all the raw sequences called, as well as a quality score for each nucleotide of the sequence. In the case of paired-end sequencing, two Fastq files corresponding to the forward and reverse reads will be obtained.

2.5.4. Data analysis

The standard pipeline for the identification of genetic variants from massively parallel sequencing is composed of two main steps: (i) alignment/mapping of the reads to a reference genome (using computational tools called aligners/mappers) and, (ii) variant discovery (using computational tools called variant callers).

Nowadays, there are several aligners and variant callers available that use different strategies, all resulting in different performances and accuracies. The algorithmic basis of these tools will not be discussed in this thesis. This section is intended to give a basic overview of the purpose for which they were designed.

A) Read alignment

Reads from sequencing platforms represent DNA sequences of the individual from whom the library was prepared. However, these reads can derive from different genomic regions with different sequences. In order to detect genetic variants in these reads, it is very important to first determine the exact genomic position from which the reads were generated (**Figure 38A**). For humans and other species, read alignment is standardized by the building of reference genomes generated by a combination of several individuals. For example, the human reference genome GRCh37/hg19 is derived from thirteen anonymous volunteers[141].

B) Variant discovery

Once the reads are aligned, variant callers compare the sequence of each read to the sequence of the reference genome for that particular location (**Figure 38B**).

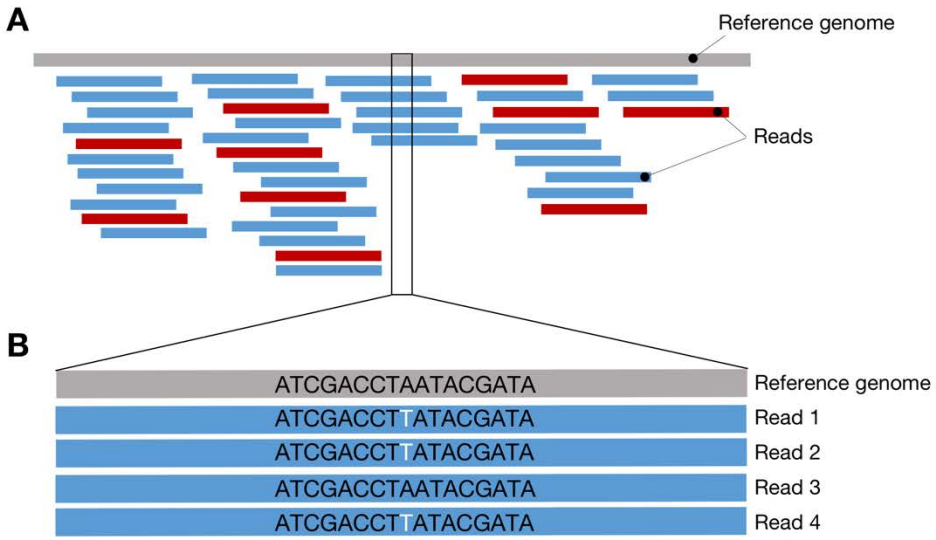


Figure 38. Schematic representation of sequencing data analysis. (A) Alignment of sequencing reads to the human reference genome. The number of reads supporting each interrogated position are reflected in the coverage tracks. Forward (blue) and reverse (red) reads are indicated. **(B)** Variant discovery example. The T variant in question (white) is supported by 3 of the reads aligning at this region.

All genomic positions that show differences in the sequence between the read and the reference genome are reported by the variant callers in a file called VCF (variant call file). The VCF contains several fields of information such as the genomic position of the genetic variant, the reference and alternative alleles and the genotype of each sample analyzed (**Table 3**). VCFs can also be further complemented with information regarding the quality of the variant, the number of reads standing for the reference and alternative alleles or the provability that the variant is a sequencing artifact.

Table 3. Example of a variant call file showing the position of the variant, the reference and alternative alleles and the genotypes for 3 different individuals.

Chr	Position	Ref	Alt	Ind. 1	Ind. 2	Ind. 3
chr3	37428076	G	A	0/0*	0/1*	0/0
chr3	37428323	G	T	0/1*	0/1*	0/0*
chr3	37443935	T	A, C	0/0*	0/1*	0/2*

Chr (chromosome), Ref (reference allele), Alt (alternative allele), Ind (individual).
*0 (reference allele), 1 (first alternative allele), 2 (second alternative allele).

2.6. Prioritization of non-coding variants

The great progress achieved in massively parallel sequencing technologies as well as in algorithms designed for variant identification enable the efficient acquisition of genetic information on a genome-wide scale[142–144]. This progress has been complemented by numerous efforts to functionally annotate both coding and non-coding genomic elements and genetic variants in the human genome.

Nowadays, there is a good understanding of the functional impact of protein-coding variants owing to historical studies of Mendelian disorders, the predictable consequences of amino acid changes (provided by the genetic code) and the considerable amount of recently available exome data[145]. However, about 80% of disease-associated variants identified by GWAS are significantly enriched in non-coding regions[146]. These non-coding variants are increasingly being recognized as acting through changes in regulatory elements that control gene activity[147–149] but very little is known about their functional consequence. To address this issue, several computational approaches have been designed to facilitate the identification of functionally relevant non-coding variants from the whole set of variants found in the genome.

The three computational approaches used in this thesis will be described in this section.

2.6.1. Combined Annotation-Dependent Depletion

Combined Annotation-Dependent Depletion (CADD)[150] is a machine-learning tool designed for scoring the deleteriousness of SNVs as well as small indels in the human genome. Deleterious variants are defined as all genetic variants that increase an individual's susceptibility or predisposition to a certain disease. Briefly, CADD was trained with two datasets of variants: one labelled as benign and the other labelled as deleterious. The benign dataset was obtained from the comparison of the human genome with the inferred genome of the most recent shared human-chimpanzee ancestor, and corresponds to those variants that are not found in the common ancestor but are fixed in the human population. On the other hand, the deleterious

dataset was obtained from the comparison of the benign dataset with an equivalent number of simulated variants generated on the basis of models of mutation rates across the genome.

The two datasets of variants were then annotated using a wide range of data types including conservation metrics such as GERP[151], phastCons[152] and phyloP[153]; regulatory information from the ENCODE Project[8] such as DHS[95] and TF binding[154]; transcript information such as distance to exon-intron boundaries or expression levels in commonly studied cell lines[8]; and protein-level scores like Grantham[155], SIFT[156], and PolyPhen[157]. After the annotation, different combinations of features were tested and only those features that more efficiently predicted if a variant was benign or deleterious were included in the final CADD model.

The main advantage of CADD is that it integrates diverse genome annotations into a single score of deleteriousness, which highly improves its performance compared to other methods based on individual annotations.

CADD is implemented as a web-based tool in which users can upload the variants to score SNVs and small indels. Otherwise, pre-computed CADD scores for all 8.6 billion possible human SNVs can be downloaded from:
http://krishna.gs.washington.edu/download/CADD/v1.3/whole_genome_SNVs.tsv.gz

2.6.2. Context-Dependent Tolerance Score

Context-Dependent Tolerance Score (CDTS)[158] measures the tolerance of genomic regions to genetic variation in the context of surrounding sequences. To obtain the CDTS scores, di Iulio *et al*., based their strategy in the generation of metaprofiles, which consist of multiple alignments that integrate and score sequence variation and frequency across elements of the same nature in the genome.

In their case, the metaprofiles were built based on all the possible heptameric sequences found in the genome. They scored the probabilities of variation of the middle nucleotide for each of the heptamers to obtain the expected variation of each nucleotide genome-wide. This expected variation was then compared to the observed variation in 11,257 whole-genomes, and the absolute difference between the observed and expected variation was defined as the CDTS.

Once obtained the CDTS, the authors ranked every region in the genome from the least tolerant to variation (1st percentile) to the most tolerant to variation (100th percentile). Interestingly, they found that most of the pathogenic variants affecting non-coding regions of the genome are embedded in genomic regions ranked at their 1st CDTS percentile.

The main advantage of CDTS is that it does not require any prior annotation of the variants and thus captures a specific set of pathogenic variants that are not detected by other metrics based on functional annotations.

Pre-computed CDTS scores for all the regions of the genome can be downloaded from: http://www.hli-opendata.com/noncoding/Pipeline/CDTS_diff_perc_coordsorted_gnomAD_N15496_hg19.bed.gz

2.6.3. DeepBind

DeepBind[159] is a machine-learning algorithm that was designed to predict sequence specificities of TFs, which can be used to identify functionally relevant variants based on their effects in TF binding. The rationale of DeepBind is that the control of gene transcription requires the binding of TFs to their target regulatory elements and that any genetic variant affecting this binding might be of functional relevance.

In order to compute the binding effects of genetic variants, a binding model for the TF of interest was first created. Briefly, DeepBind was trained using a set of sequences–TF binding sites–that had been experimentally obtained from ChIP-seq experiments. In the first training step, DeepBind identified local sequence patterns (i.e. TF binding motifs) that summarized the frequency of each nucleotide at each position of the sequence. These local features were then synthesized into higher-level structures that were considered in many combinations and orientations and that were summarized into a final binding model for the TF of interest.

To test the validity of DeepBind models in identifying functionally relevant variants, Alipanahi *et al.,* scored several disease-associated variants that had been already shown to affect gene expression. With this test, the authors corroborated that DeepBind is able to predict TF binding alterations produced by the presence of a genetic variant.

The main advantage of DeepBind versus previous prediction models is that DeepBind is trained from *in vivo* data, which highly increases its accuracy when scoring TF binding events.

Currently, DeepBind models for 137 TFs that can be directly downloaded from http://tools.genes.toronto.edu/deepbind/.

3. The heart

The heart is a muscular organ that plumps blood through the blood vessels of the circulatory system to provide the body with oxygen and nutrients, as well as with assistance in the removal of waste products of cell metabolism[160].

In humans, the heart is divided into four chambers: the upper left and right atria, which receive the blood; and the lower left and right ventricles, which eject the blood. In turn, these four chambers are separated by a muscular wall named septum that avoids blood exchange between left and right parts of the heart[160] (**Figure 39**).

The heart wall is composed of three principal layers named from the most external to most internal: epicardium, myocardium and endocardium (**Figure 39**). The epicardium is a serous membrane that protects the external area of the heart. The myocardium is the heart muscular tissue itself and contains the cardiac cells or cardiomyocytes. Finally, the endocardium protects the internal part of the heart chambers and valves[161].

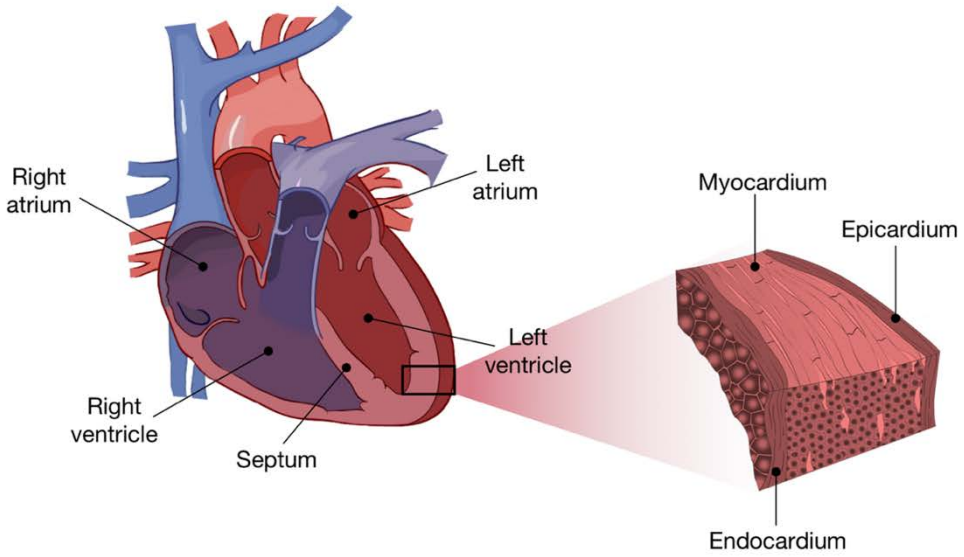


Figure 39. Heart anatomy. Schematic representation of the four heart chambers and the septum (left) and the three principal layers that form the cardiac wall (right). Blue corresponds to deoxygenated blood while red corresponds to oxygenated blood. Figures adapted from Science Learning Hub[162] and Studyblue[163].

3.1. Heart development

In humans, the heart is the first functional organ to develop. Its function starts early during embryogenesis and is crucial to supply the embryo with nutrients and oxygen. From the beginning of its formation, the heart itself generates and propagates the electrical impulse that is required to initiate coordinated contractions to efficiently pump blood throughout the body[164].

The adult heart derives from four distinct pools of progenitors: the first heart field, the second heart field, the proepicardial organ and the cardiac neural chest. These progenitors migrate, proliferate, and differentiate to a myriad of different cell types that comprise the adult heart: cardiomyocytes, endothelial cells, vascular smooth muscle cells, fibroblasts, and the conduction system[165] (**Figure 40A** and **B**) .

Briefly, the progenitors comprising the heart fields coalesce to form a parallel pair of vessels which fuse to form the primitive heart tube. This tube elongates and undergoes rightward looping, followed by a series of septation and fusion events to give rise to the four chambered heart (**Figure 40C** and **D**). Shortly after birth, cardiomyocytes undergo terminal differentiation, loosing their proliferative capacity[166].

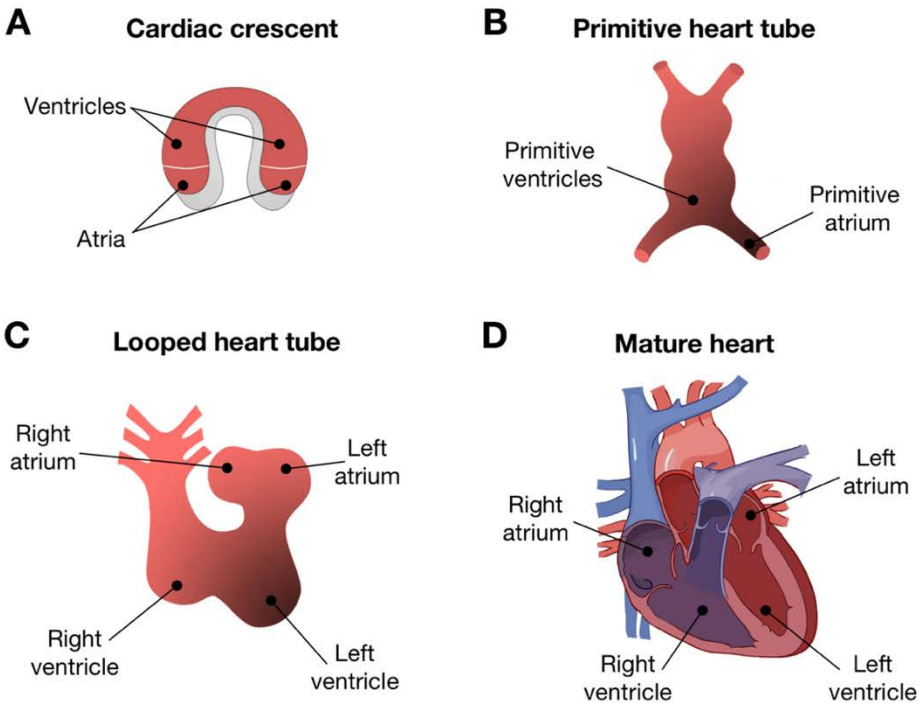


Figure 40. Schematic representation of heart development during human embryogenesis. (A) Cardiac crescent at day 15. The first heart field (red) forms particular segments of the linear heart tube. The second heart field (grey) is located medial and caudal of the first heart field and will later contribute to the arterial and venous pole formation. **(B)** Primitive heart tube with its primitive ventricle and atrium at day 21. **(C)** By day 28, the linear heart tube loops to the right (D-loop) to establish the future position of the cardiac regions (atria, ventricles, outflow tract). **(D)** By day 50, in the mature heart, the chambers and outflow tract of the heart are divided by the atrial septum, the interventricular septum, two atrioventricular valves (tricuspid valve, mitral valve) and two semilunar valves (aortic valve, pulmonary valve). Adapted from Kloesel *et al.,*[165] and Lindsey *et al.*[167].

Each of these events relies on specific orchestrated functions of conserved cardiogenic TFs, many of which appear to have chamber-specific expression patterns. Among them, the most critical transcriptional modulators of heart development are **NKX2.5**, **MEF2C**, **HAND1/2**, **TBX5** and **GATA4**[166]. These TFs have been shown to interact and form multiproteic complexes to regulate their target genes and, in fact, the co-occupancy by these cardiac TFs has been proposed as a mechanism to identify active transcriptional enhancers in the heart[168].

A) NKX2.5

NKX2.5 is a member of the homeodomain family of TFs that are critical regulators of organ development[169]. NKX2.5 has a key role in heart formation and development. Accordingly, genetic alterations in NKX2.5 are related to cardiac development defects. NKX2.5-null mice show a heart without proper looping and decreased expression of ventricular markers[170,171]. In humans, genetic variants affecting the DNA binding ability of NKX2.5 have been related to atrial septal defects with alterations in the atrioventricular electrical conduction[172].

B) MEF2C

MEF2C is a member of the MADS box transcription enhancer factor 2 (MEF2) family of proteins, which play a role in myogenesis. MEF2C is a transcription activator which binds specifically to the MEF2 element and controls cardiac morphogenesis and myogenesis. MEF2C null mice embryos fail to properly develop a right ventricle, do not exhibit cardiac looping, and fail to express a subset of cardiomyocyte-specific genes[173,174]. At the presumed onset of cardiac looping, MEF2C-null mice also show down-regulation of HAND2 expression and altered HAND1 expression, which supports the role of MEF2C in ventricular development[173].

C) HAND1/2

HAND1/2 are members of the basic helix-loop-helix family of TFs. They are asymmetrically expressed in the developing ventricular chambers and play an essential role in cardiac morphogenesis. During the linear tube stage of murine cardiogenesis, HAND2 expression is higher in the right ventricle, while HAND1 expression is higher in the left ventricle[175,176]. HAND1-null mice show embryonic lethality by an arrest of cardiac development while targeted deletion of HAND2 in mice causes embryonic lethality and hypoplasia of the right ventricle[176].

D) TBX5

TBX5 is a member of the T-box TF family and is primarily known for its role in cardiac and forelimb development. It appears to be particularly important for the correct cardiac looping and

formation of the septum that separates the right and left sides of the heart[166]. TBX5-null mice display arrested cardiac development with failure of cardiac looping, and hypoplasia of the left ventricle, resulting in embryonic lethality[177].

E) GATA4

GATA4 is a member of the GATA family of double zinc-finger TFs, which have been identified as crucial mediators of vertebrate cardiogenesis, and show potential functional redundancy with one another[178–180]. GATA4 is widely expressed in the developing heart and in adult cardiomyocytes, and plays a critical role in cardiac differentiation and morphogenesis[181–184]. It has been shown to directly regulate the transcriptional activity of numerous cardiac-restricted genes that control cardiac progenitor cell differentiation, including MEF2C, HAND2 and GATA6. GATA4 interacts with TBX5 and NKX2.5 to form multiprotein transcriptional complexes that regulate the expression of a set of downstream genes involved in cardiac development[177,185,186]. GATA4-null mice show gross cardiac defects that result in embryonic lethality[183,184]. In humans, genetic variants affecting GATA4 result in familial septal defects and are involved in a range of cardiac defects including right ventricular hypoplasia and cardiomyopathy[187].

Apart from its role during embryonic development, GATA4 is also important for the electrical activity of the heart. In a recent study conducted by Tarradas *et al.,*[188] it was shown that GATA4 regulates the expression of *SCN5A* in adult human hearts. As it will be explained in the following sections, *SCN5A* encodes for the alpha subunit of the cardiac sodium channel, which plays a crucial role in the generation of the cardiac action potential.

3.2. The electrical activity of the heart

Every heartbeat is characterized by a perfectly coordinated relaxation (diastole) and contraction (systole) of the atria and ventricles. During systole, the ventricles contract to pump blood to the pulmonary artery and aorta, while the atria are relaxed and collecting blood. During diastole, the ventricles are relaxed and the atria contract to pump blood to the ventricles[189].

The coordination between systole and diastole is mediated by the electrical conduction system of the heart, which transmits the electrical impulse from the atria to the ventricles (**Figure 41**). The cardiac electrical activity is originated at the autonomous cells of the sinoatrial node (SAN), located in the wall of the right atrium. The SAN is also referred to as the heart's natural pacemaker as it continuously and autonomously initiates the electrical impulse, which is propagated to the atrial cardiomyocytes, resulting in excitation of the atria. The excitation wave originated in the SAN propagates to the atrioventricular node (AVN), which electrically connects

the atria and the ventricles. The AVN delays the cardiac pulse to ensure that the atria have contracted and the blood has been ejected into the ventricles before ventricle contraction[190]. Finally, the excitation wave propagates from AVN to the right and left ventricles by the left and right bundle branches and Purkinje fibers[160].

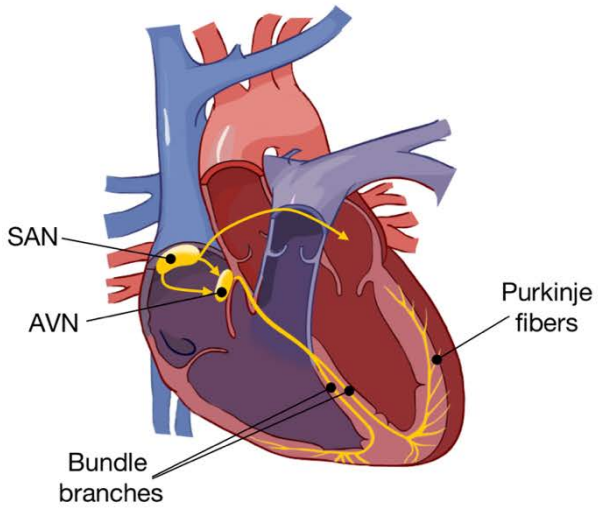


Figure 41. Electrical conduction system of the heart. Schematic representation of the components that constitute the electrical conduction system. The electrical impulse is initiated at the sinoatrial node (SAN) and it is propagated to the atrioventricular node (AVN), from where it propagates to the right and left ventricles by the left and right bundle braches and Purkinje fibers. Figure adapted from Munshi[191].

3.3. Voltage-gated ion channels

Voltage-gated ion channels are integral membrane proteins that enable the passage of selected ions across cell membranes. They open and close in response to changes in transmembrane voltage, and play a key role in electrical signaling of excitable cells such as neurons or cardiomyocytes[192].

Voltage-gated ion channels contain a voltage sensor, which is a region of the protein bearing charged amino acids that relocate upon changes in the membrane electric field. The movement of the sensor initiates a conformation change in the gate of the conducting pathway thus controlling the flow of ions (**Figure 42A**). Voltage-gated ion channels can exist in three functionally distinct states: closed (resting), open (active), and inactivated. Both closed and inactivated states are non-conducting, but channels that have been inactivated are refractory unless the cell is returned to the initial membrane potential to allow them to return to the closed state[193] (**Figure 42B**).

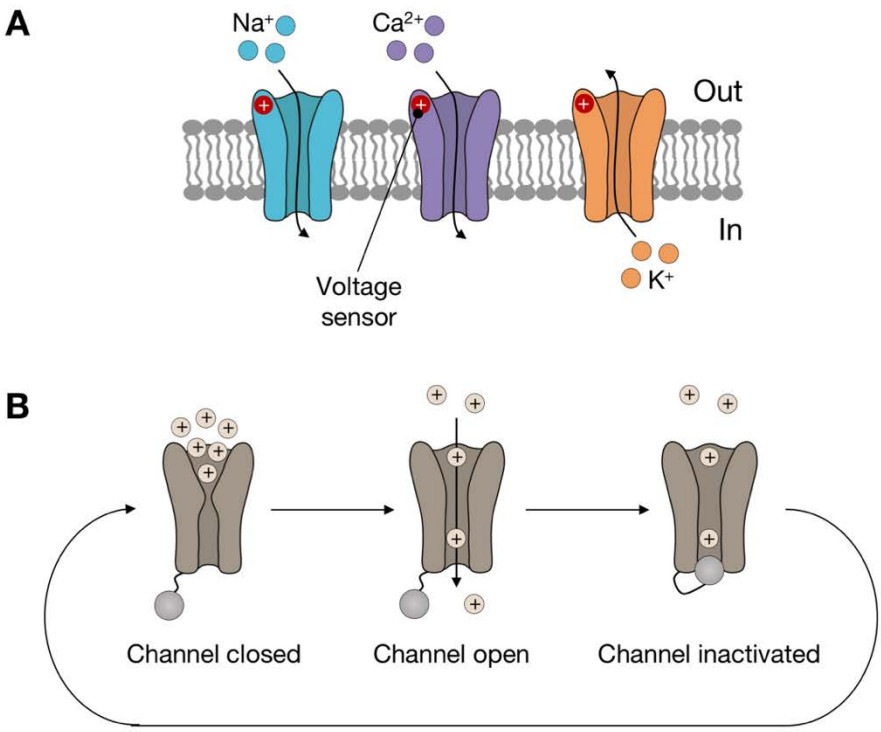


Figure 42. Voltage-gated ion channels. (A) Schematics of the three main cardiac voltage-gated ion channels (Na^+; sodium, Ca^{2+}; calcium and K^+; potassium). Na^+ and Ca^{2+} participate in the inward currents while K^+ participate in the outward currents. Adapted from Amin et al.,[194]. **(B)** Conformational states of voltage-gated ion channels. Figure adapted from Theile and Cummins[195].

3.3.1. Sodium channels

Voltage-gated sodium channels (Na_V channels) are dynamic transmembrane proteins that govern action potential initiation and propagation in excitable membranes[196]. In mammals, *SCN1A-SCN11A* are the genes encoding for a family of nine functionally expressed voltage-gated sodium channels ($Na_V1.1$- $Na_V1.9$), which are more than 50% homologous in amino acid sequence[197].

All voltage-gated sodium channel isoforms show the same overall structure, consisting of a large pore forming α-subunit and one or more of four auxiliary β-subunits (β1, β2, β3 and β4), encoded by *SCN1B-SCN4B* genes, respectively[198,199]. An extra β1 subunit (β1b) is obtained by alternative splicing of the *SCN1B* gene[200]. The main functional properties of voltage-gated sodium channels reside in the α-subunit, although its activity can be modulated by β-subunits that can regulate the amount of channels in the plasma membrane or α-subunit kinetics[199].

Distinct α-subunit isoforms are expressed in tissue-specific patterns and exhibit differences in gating properties that tailor them for distinct physiological roles. The α-subunit expressed in cardiomyocytes is the $Na_V1.5$, encoded by *SCN5A* gene[201].

$Na_V1.5$ is a 227 kilo Daltons (kDa) tetramer of four homologous domains (DI-DIV) linked by intracellular loops (**Figure 43**). Each domain consists of six transmembrane segments (S1-S6). S4 is positively charged and is involved in voltage-dependent activation of the channel, while the loops between S5 and S6 of each domain curve back into the membrane to form the sodium-conducting channel pore[202]. The four homologous domains DI-DIV fold and position the S5 and S6 segments of each domain at the internal part of the tetramer, which forms the final pore with selectivity for Na^+ ions[203,204].

The ***SCN5A***, ***SCN2B*** and ***SCN3B*** genes (encoding the sodium channel α-subunit $Na_V1.5$ and two accessory β-subunits, respectively) are important for this thesis.

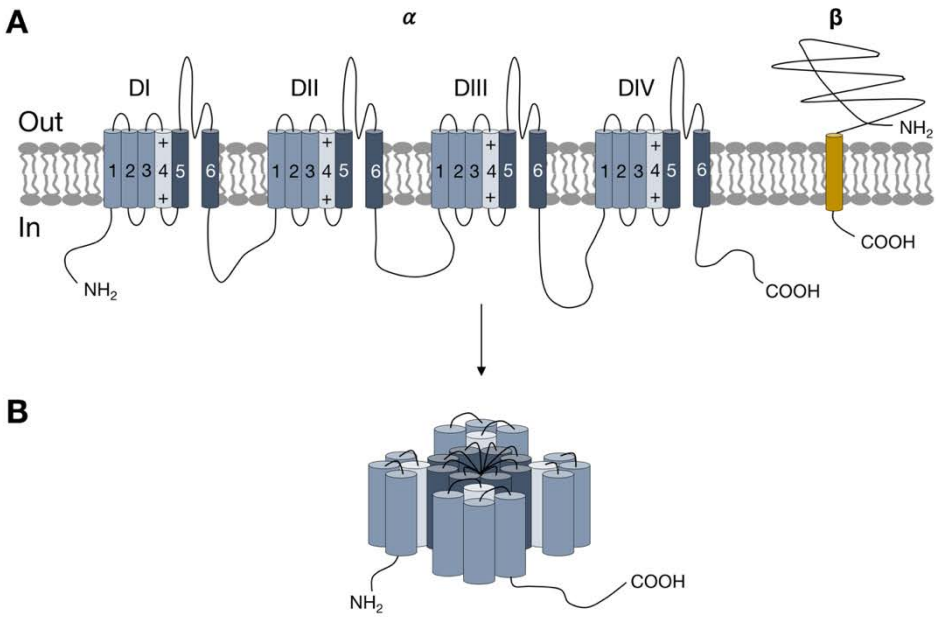


Figure 43. Cardiac voltage-gated sodium channel. (A) $Na_V1.5$ α-subunit with the 4 domains (DI-DIV), each composed by six transmembrane segments. An accessory β-subunit is also represented. Figure adapted from Vacher *et al.*[205]. **(B)** Folding of the four $Na_V1.5$ domains around an ion-conducting pore, which is lined by the loops between the S5 and S6 segments. Figure adapted from Amin and Asghari-Roodsari[206].

3.3.2. Calcium channels

The family of voltage-gated calcium channels serve as the key transducers of cell surface membrane potential changes into local intracellular calcium transients that initiate many different physiological events[207].

All voltage-gated calcium channels are composed by a combination of up to four distinct subunits (α1, α2δ, β and γ). The α1-subunit of 190-250 kDa is the largest subunit and includes the conduction pore, the voltage sensor and the gating apparatus. To date, ten members of the α1-subunit ($Ca_V1.1$-$Ca_V1.4$, $Ca_V2.1$-$Ca_V2.3$, and $Ca_V3.1$- $Ca_V3.3$), encoded by *CACNA1A*-

CACNA1I and *CACNA1S* genes have been characterized in mammals. Similar to the α-subunits of the sodium channel, the α1-subunit of voltage-gated calcium channels is organized in four homologous domains (DI-DIV), each comprising six transmembrane segments (S1-S6; **Figure 44**). S4 is positively charged and is involved in voltage-dependent activation of the channel, while S5 and S6 form the channel pore that determines ion conductance and selectivity. Calcium channels are also tetramers with the S5 and S6 segments of each domain at the internal part of the tetramer, forming the pore with selectivity for Ca^{2+} ions[207].

The α2δ-, β- and γ-subunits are accessory subunits that modulate the trafficking and kinetics of the α1-subunit[208]. The α2δ-subunits are encoded by *CACNA2D1-CACNA2D4* genes, the β-subunits are encoded by *CACNB1-CACNAB4* genes, and the γ-subunits are encoded by *CACNG1-CACNG8* genes.

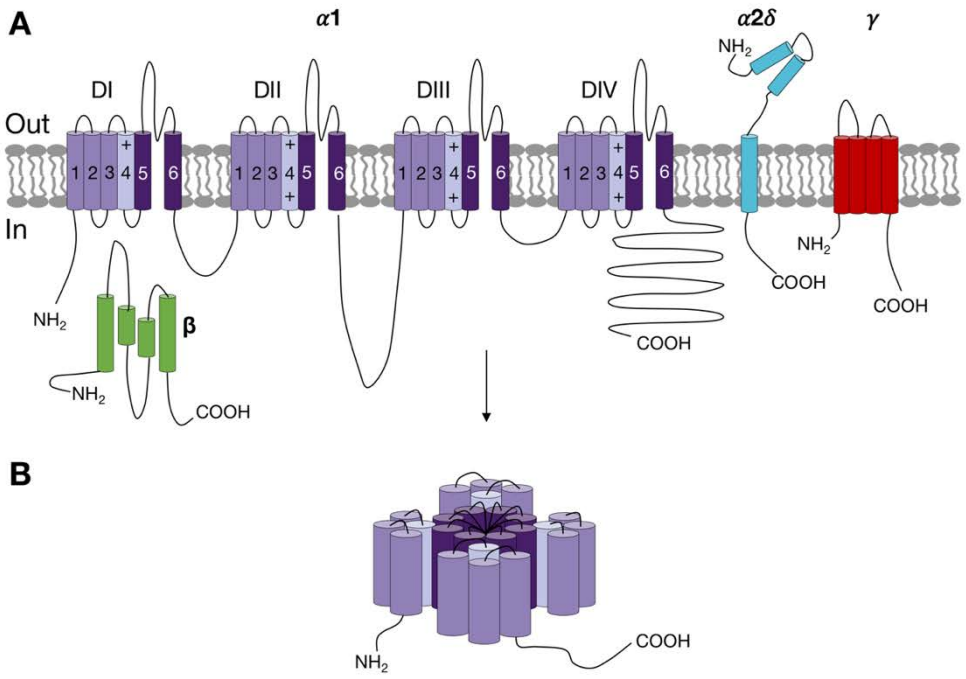


Figure 44. Cardiac voltage-gated calcium channel. (A) $Ca_v1.2$ α-subunit with the 4 domains (DI-DIV), each composed by six transmembrane segments. The accessory α2δ-, β- and γ-subunits are also represented. **(B)** Folding of the four $Ca_v1.2$ domains around an ion-conducting pore, which is lined by the loops between the S5 and S6 segments. Figure adapted from Vacher *et al.*[205].

Voltage-gated calcium channels have been classified into five groups (L-, N-, P/Q-, R- and T-types) according to the type of current they show[207]. These five types of channels show tissue-specific expression patterns, and are involved in different physiological roles. The L- and T-type channels are found in cardiac muscle. More specifically, the L-type channels are found in all cardiac cell types, while the T-type channels are mainly found in pacemaker and Purkinje cells[208].

The ***CACNA1C***, ***CACNB2*** and ***CACNA2D1*** genes (encoding the calcium channel α1-subunit $Ca_V1.2$, the auxiliary β-subunit and the auxiliary $\alpha 2\delta$-subunit, respectively) are important for this thesis.

3.3.3. Potassium channels

Voltage-gated potassium channels display broad distributions in the nervous system and other tissues such as the cardiac muscle[209]. They play an important role in regulating cardiac muscle excitability by returning the depolarized cell to a resting state, and by controlling action potential duration and frequency[210]. In the case of voltage-gated potassium channels, variations in the level of expression of these channels are responsible for regional differences of the action potential configuration in the atria, ventricles and across the myocardial wall.

Each voltage-gated potassium channel is a tetramer of four different or identical pore-forming α-subunits, and it may also contain auxiliary β-subunits that can modulate the channel function and/or localization[205,211] (**Figure 45**). Each pore-forming α-subunit contains six transmembrane segments (S1-S6), with the first four transmembrane segments (S1-S4) forming the voltage sensor, and the last two transmembrane segments (S5 and S6) forming the pore domain[211]. In humans, α-subunits are grouped in 12 subfamilies (K_V1- K_V12) encoded by 40 different genes, while β-subunits are composed by 12 members encoded by 12 different genes[211].

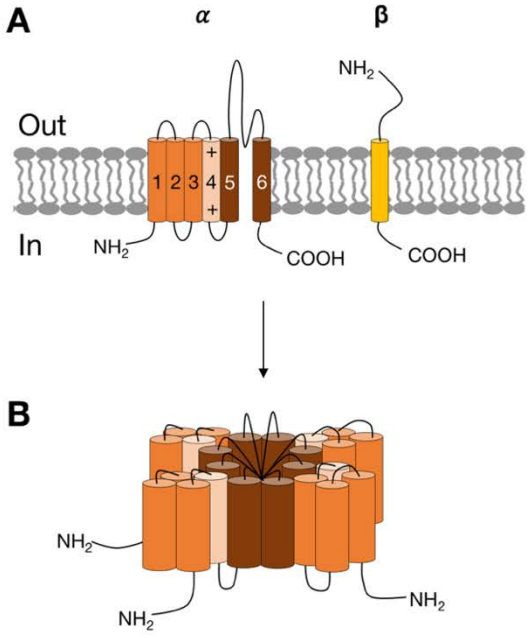


Figure 45. Cardiac voltage-gated potassium channel. (A) $K_V7.1$ ***α***-subunit with the six transmembrane segments. The accessory β-subunit is also represented. **(B)** Folding of the four different or identical $K_V7.1$ α-subunits around an ion-conducting pore, which is lined by the loops between the S5 and S6 segments. Figure adapted from Vacher *et al.*[205].

3.4. Cardiac action potential

Excitable cells such as cardiomyocytes are characterized by a plasma membrane with an asymmetric distribution of charges (ions) between the inside and outside of the cells. The plasma membrane is impermeable to these ions, which creates a difference in potential between both sides of the membrane. The opening and closing (gating) of the ion channels, discussed in the previous section, enable transmembrane ion currents and, as a result, the formation of action potentials (APs)[194]. These APs are responsible for the conduction of the electrical activity throughout the heart that ends with heart contraction.

In general, the resting potential of atrial and ventricular cardiomyocytes (**phase 4**) is stable and negative (approximately -85 mV) due to an accumulation of positive charges at the outer plasma membrane. Upon excitation by electrical impulses from adjacent cells, voltage-gated sodium channels open and allow an inward sodium current (I_{Na}), which gives rise to **phase 0** depolarization (initial upstroke). Phase 0 is followed by **phase 1** (early repolarization), accomplished by the transient outward potassium current (I_{to}) and inactivation of sodium channels. **Phase 2** (plateau) represents a balance between the depolarizing L-type inward calcium current ($I_{Ca,L}$) and the repolarizing potassium currents (I_{Kr} and I_{Ks}). **Phase 3** (repolarization) reflects the predominance of potassium currents (I_{k1}) after inactivation of the L-type calcium channels[194] (**Figure 46**).

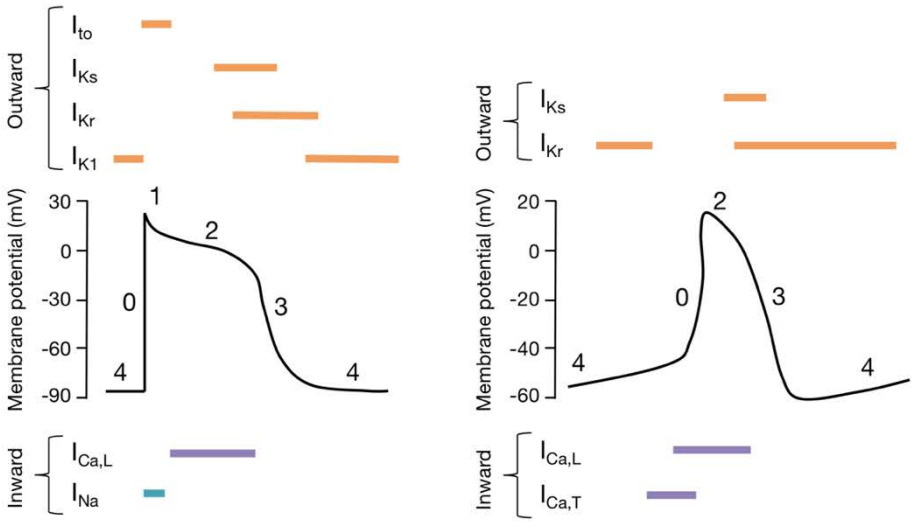


Figure 46. Phases of the cardiac AP. Schematic representation of inward and outward currents that contribute to action potential formation in ventricular cardiomyocytes (left) and sinoatrial node cardiomyocytes (right). Figure adapted from Amin *et al.*[194].

In contrast to atrial and ventricular cardiomyocytes, SAN and AVN cardiomyocytes show slow depolarization of the resting potential during phase 4 (**Figure 46**). This is mostly due to the

absence of I_{K1}, which allows inward currents to gradually depolarize the membrane potential. This slow depolarization during phase 4 inactivates most Na^+ channels and decreases their availability for phase 0. As a consequence, in SAN and AVN cardiomyocytes, AP depolarization is mainly achieved by $I_{Ca,L}$ and the T-type Ca^{2+} current ($I_{Ca,T}$). This phenomenon is what explains the auto-rhythmicity properties of the SAN and AVN[212].

3.5. Electrocardiogram

The propagation of the electrical impulse from the atria to ventricles can be registered in the electrocardiogram (EKG), which is the result of all APs that occur in different regions of the heart[206].

Depolarization of the right and left atria originates the first wave of the EKG called “P wave”. The P wave is followed by a flat line corresponding to the propagation of the electrical impulse from the atria to the ventricles. Depolarization of the ventricles generates the next wave called “QRS complex”. Finally, repolarization of the ventricles results in the last wave called “T wave”. Atrial repolarization is not detected in the EKG because it is masked by the QRS segment– originated during ventricular depolarization– that occurs at the same time[194,213] (**Figure 47**).

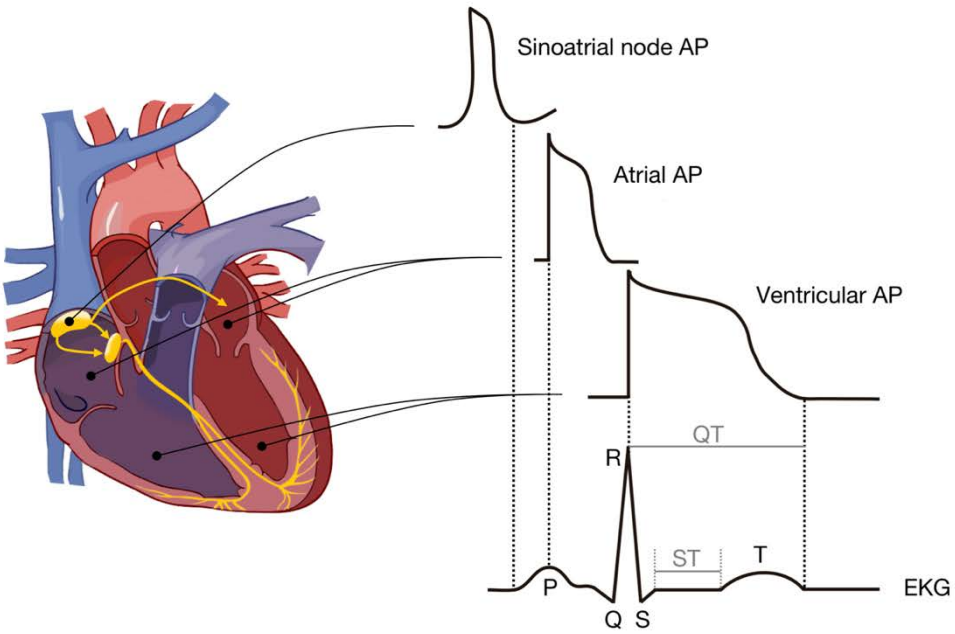


Figure 47. EKG representation. Relationship between EKG and APs of cardiomyocytes from different heart regions. Figure adapted from Amin *et al.*[194].

4. Sudden cardiac death

Sudden death (SD) is defined as a natural and unexpected death that occurs within a short period of time (generally less than 1 h from the onset of acute symptoms) in an apparently healthy person. There are several origins of SD and, even that post-mortem examination often fails to determine the exact cause of death, approximately 85% of all SDs are of cardiac origin, which are known as sudden cardiac death (SCD)[214]. The estimated annual incidence of SCD in United States, Europe and China ranges from 50 to 100 affected per 100,000 individuals[215,216].

Apart from the sudden infant death syndrome, affecting children between birth and 6 months, SCD typically occurs between 45 and 75 years of age[217] and its incidence is 3 to 4 times higher in men than in women[218].

The etiologies of SCD are very diverse, although certain diseases are known to play significant roles in its pathogenesis. The most common cause of SCD is coronary heart disease, accounting for up to 75-80% of SCDs. The remaining 20-25% of SCDs are inherited. From these, 10-15% are caused by cardiomyopathies, primarily related to cardiac structural abnormalities[219]. This type of diseases are associated with genetic alterations in structural proteins, including those of sarcomeres, desmosomes, and the cytoskeleton[219–221]. The most common SCD-related cardiomyopathies include hypertrophic cardiomyopathy (HCM), dilated cardiomyopathy (DCM), arrhythmogenic right ventricular cardiomyopathy (ARVC), and left ventricular noncompaction (LVNC). The remaining 5-10% of inherited SCDs are caused by **channelopathies**, related to an alteration of the electrical activity in structural normal hearts. This type of diseases are associated with genetic alterations in membrane ion channels or their regulatory proteins that compromise their function[222,223]. The major channelopathies associated to SCD include long-QT syndrome (LQTS), short-QT syndrome (SQTS), **Brugada syndrome** (**BrS**), and catecholaminergic polymorphic ventricular tachycardia (CPVT)[220,221,223].

4.1. Electrical diseases related to sudden cardiac death

As mentioned earlier, every heartbeat requires an orchestrated regulation of the inward and outward currents through the voltage-gated ion channels found at the cell membrane of cardiomyocytes. Any change in gating properties or expression levels of cardiac ion channels can affect the propagation of the electrical impulse throughout the heart and may result in cardiac arrhythmias such as those displayed by BrS, LQTS or SQTS channelopaties (**Figure 48**)[224].

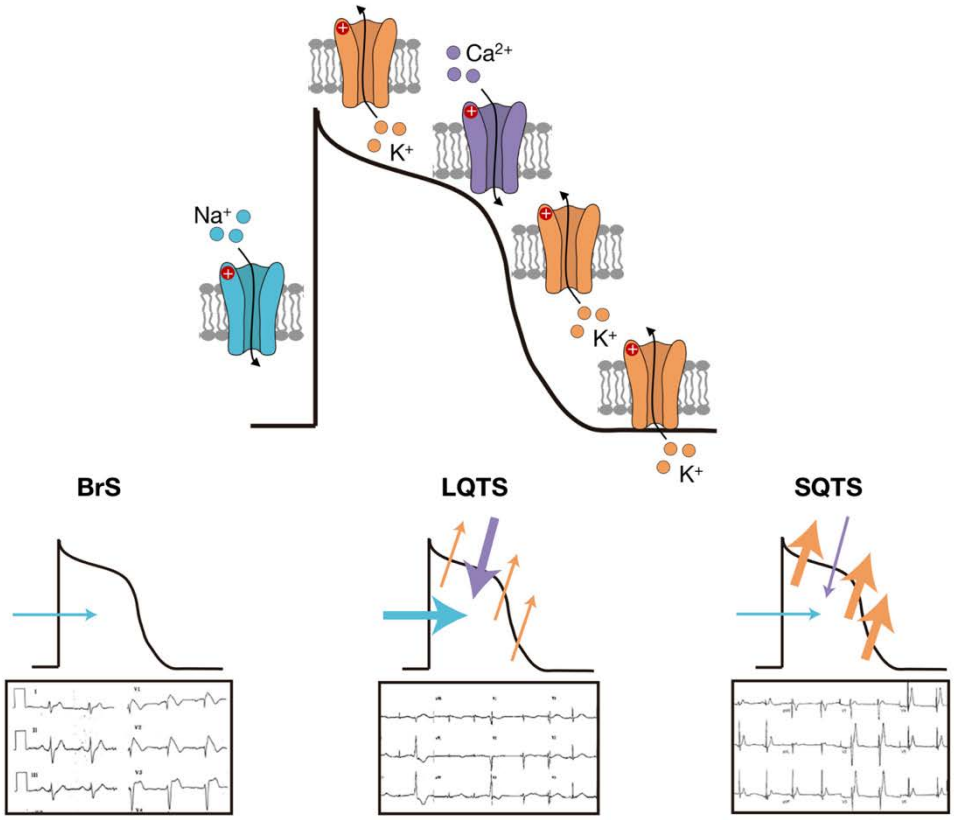


Figure 48. Cardiac channelopaties. Top: diagram showing the principal voltage-gated ion channels participating at each phase of the cardiac AP. **Bottom:** representation of three examples of electrical diseases associated to loss-of-function of the voltage-gated sodium channel (BrS), gain-of-function of the voltage-gated sodium and calcium channels (LQTS) and gain-of-function of voltage-gated potassium channels (SQTS). Figure provided by Helena Riuró.

One of the main causes of cardiac channelopaties is the presence of pathogenic variants at the coding regions of genes encoding cardiac ion channels. These genetic variants can affect both the α-subunits and the auxiliary subunits. However, in the last few years, it has been observed that genetic variants affecting the activity of *cis*-regulatory elements regulating the expression of cardiac ion channels can also result in cardiac conduction defects, increasing the susceptibility to cardiac arrhythmias[44,225].

4.1.1. Brugada syndrome

BrS, first described in 1992, is a familial cardiac disease characterized by a typical EKG pattern with an ST-segment elevation in the right precordial leads V1-V3[226,227] (**Figure 49**). This ST-segment elevation is diagnostic of BrS, and EKGs with this pattern are classified as BrS Type 1. In contrast, EKGs displaying Brs Type 2 and Type 3 patterns are not diagnostic and are considered as susceptibility patterns.

The EKG manifestations of Brugada syndrome are often dynamic and some patients may present a normal EKG until external factors unmask the ST-segment elevation. These external factors include febrile states, antidepressants, cocaine toxicity, or sodium channel-blockers (i.e. ajmaline and flecainide)[228]. Indeed, ajmaline and flecainide are used in the clinical setting to unravel the EKG pattern in those patients with evidence of the disease[229].

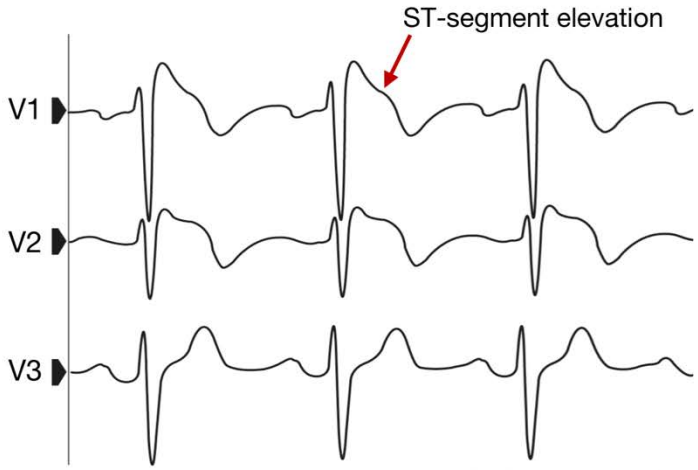


Figure 49. BrS EKG. Typical BrS EKG displaying the ST-segment elevation in the right precordial leads V1-V3 (highlighted by the red arrow). Figure provided by Anna Tarradas.

The prevalence of BrS varies depending on the geographical region, being of 1-5 individuals per 10,000 inhabitants in Europe and 12 individuals per 10,000 inhabitants in Southeast Asia. The pathology has been accounted to be responsible for at least 4% of all sudden deaths[230] and it is also associated to supraventricular arrhythmias (20% of cases) and atrial fibrillation (10-20% of cases) in structural normal hearts[230]. Most of the arrhythmogenic symptoms appear between 40-45 years old and it is more prevalent in males than females (80% versus 20%).

The typical ST-segment elevation observed in BrS patients is caused by a reduction of the inward Na^+ currents during depolarization. However, the exact electrophysiological mechanism leading to this ST-segment elevation is still unresolved. Indeed, two different hypotheses have been proposed: the repolarization disorder hypothesis and the depolarization disorder hypothesis (**Figure 50**)[231,232].

A) The repolarization disorder hypothesis

This hypothesis states that differential shortening in the AP across the myocardial wall (epicardium and endocardium) is primarily responsible for the BrS phenotype[233]. In normal conditions, during the phase 1 of the cardiac AP, the outward K^+ currents counterad the inward Na^+ currents. Given that the K^+ currents are higher in the epicardium than the endocardium, the

difference between the Na^+ and K^+ currents is higher in the epicardium. In BrS patients, characterized by reduced Na^+ currents, the difference between the inner Na^+ and outward K^+ currents is higher, especially in the epicardium, which is reflected as the ST-segment elevation in the EKG. This decompensation in the epicardium currents also results in a longer phase 2 of the AP, reflected as a negative T wave (**Figure 50A**).

B) The depolarization disorder hypothesis

This hypothesis states that the reduction of Na^+ currents during phase 0 of the AP results in a reduction in the conduction velocity of the AP at the right ventricular outflow tract (RVOT) compared to the right ventricle. As a consequence, the ST-segment elevation and the negative T wave are registered in the EKG (**Figure 50B**).

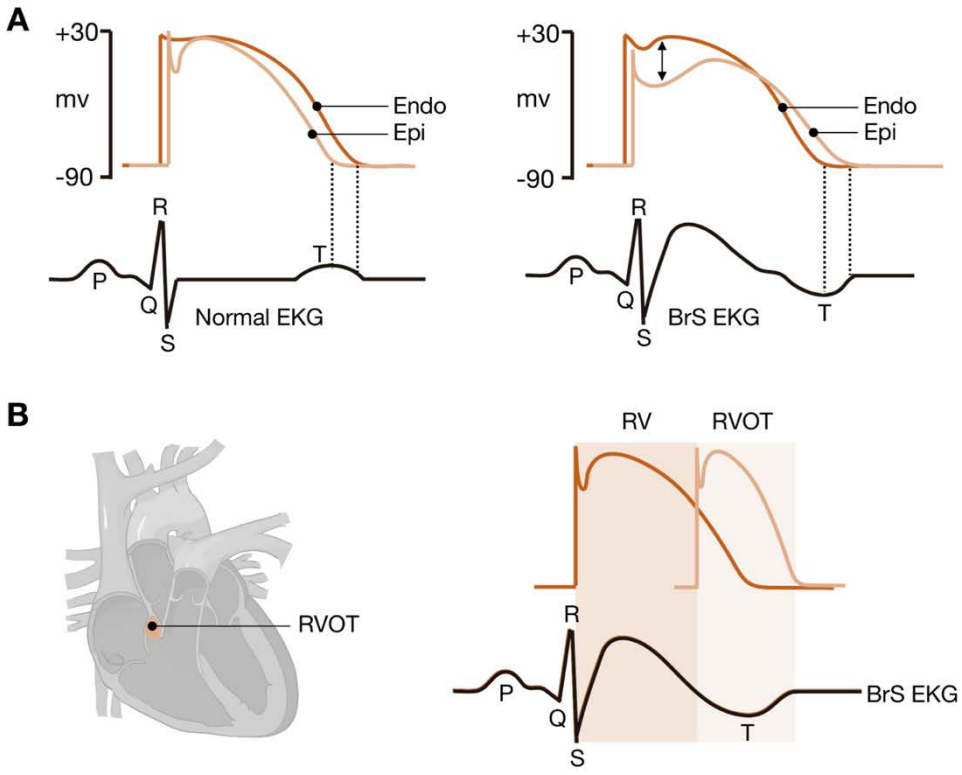


Figure 50. Hypothesis to explain the ST-segment elevation in BrS. (A) Repolarization disorder hypothesis. **(B)** Depolarization disorder hypothesis. Figure adapted from Meregalli *et al*.[231].

The most common cause of BrS is the presence of genetic variants at the protein-coding region of the *SCN5A* gene, which encodes for the α-subunit of the cardiac sodium channel $Na_V1.5$. To date, more than 350 genetic variants at the protein-coding region of *SCN5A* have been described, together accounting for 25% of BrS cases[234]. These variants are related to a loss-of-function of the channel either affecting its function or reducing the number of channels

found in the plasma membrane. Other genes have also been associated to BrS, although with a lower incidence (**Table 4**).

Although more than 20 genes have been associated to BrS pathogenesis, in this thesis we analyzed the genes encoding the cardiac sodium channel (***SCN5A***) and two of their regulatory β-subunits (***SCN2B***[235] and ***SCN3B***[236]), together with the gene encoding the L-type calcium channel (***CACNA1C***[237,238]) and its regulatory β- and $\alpha 2\delta$-subunits (***CACNB2*** and ***CACNA2D1***[237,238], respectively).

Table 4. Genes associated to BrS phenotype. The genes analyzed in this thesis are highlighted in bold.

Gene name	Gene product	Functional effect	Incidence (%)
SCN5A	**$Na_V1.5$**	**Loss-of-function**	**20-25**
SCN1B	$Na_V\beta1$	Loss-of-function	1-2
SCN2B	**$Na_V\beta2$**	**Loss-of-function**	**Rare**
SCN3B	**$Na_V\beta3$**	**Loss-of-function**	**Rare**
SCN10A	$Na_V1.8$	Loss-of-function	2.5-16
RANGRF	MOG1	Loss-of-function	Rare
GPD1-L	G3PD1L	Loss-of-function	Rare
SLMAP	SLMAP	Loss-of-function	Rare
PKP2	Plakophilin-2	Plakophilin-2 cause	2.5
TRPM4	NSCCa	Loss-of-function	8
CACNA1C	**$Ca_V1.2$**	**Loss-of-function**	**6-7**
CACNB2	**$Ca_V\beta2$**	**Loss-of-function**	**4 – 5**
CACNA2D1	**$Ca_V\alpha2\delta1$**	**Loss-of-function**	**Rare**
ABCC9	SUR2A	Gain-of-function	4-5
KCND3	$K_V4.3$	Gain-of-function	Rare
KCNE3	MiRP2	Gain-of-function	<1
KCNJ8	Kir6.1	Gain-of-function	Rare
KCNH2	$K_V11.1$	Gain-of-function	1-2

Although genetic variants at protein-coding regions of all these genes have been related to BrS, in 65-70% of cases the etiology of the disease is still unknown[239] ('orphan' cases).

During the last few years, it has been proposed that the presence of genetic variants at *cis*-regulatory regions involved in the regulation of the expression levels of cardiac ion channels and their accessory subunits could be another mechanism that could explain the BrS phenotype. In this line of evidence, Bezzina *et al.,*[240] identified an haplotype of 6 SNPs in the *SCN5A* promoter in near complete linkage disequilibrium that occurred at an AF of 0.22 in Asian subjects. Luciferase reporter assays performed in rat cardiomyocytes showed that the activity of the

SCN5A promoter containing the haplotype variant is 62% reduced compared to the reference haplotype. Moreover, when comparing individuals with the reference and variant haplotypes, they observed that individuals carrying the haplotype variant show longer PR and QRS intervals.

Similarly, an SNP in the *SCN10A* gene identified through GWAS was associated with alterations in cardiac conduction patterns and susceptibility to arrhythmias. This SNP is located in an enhancer region–within the *SCN10A* gene–that modulates the expression of *SCN5A* gene. Van den Boogard *et al.,*[44,225] provided evidence that the SNP disrupts TBX5 binding in the enhancer region and results in reduced *SCN5A* expression levels.

Together, both studies further support the hypothesis that genetic variants affecting *cis*-regulatory regions might be an important contributor to BrS susceptibility and explain some of the remaining BrS 'orphan' cases.

II. Rationale and Objectives

Rationale

BrS is a familial cardiac disease with high susceptibility to ventricular arrhythmias and sudden cardiac death. The ***SCN5A*** gene, which encodes the alpha subunit of cardiac sodium channel ($Na_V1.5$), was the first gene associated with BrS and still remains as the major gene linked to BrS pathogenesis Genetic variants in protein-coding regions of *SCN5A* gene have been directly linked to 11-24% of BrS cases. In addition, genetic variants in protein-coding regions of other ion channel genes, such as sodium channel regulatory beta subunits ***SCN2B*** and ***SCN3B***, as well as calcium channels ***CACNA1C, CACNB2*** and ***CACNA2D1*** account for 5-10% of BrS cases. Together, protein-coding variants in ion channel genes and their regulatory subunits account for up to 25-30% of BrS cases. Therefore, in a large fraction of BrS diagnosed patients the etiology of the disease is still unknown ('orphan' BrS cases). At present, there is a good understanding of the functional impact of protein-coding variants thanks to classical studies of Mendielian disorders, the predictable consequences of amino acid changes and the recent availability of exome sequencing data. However, protein coding regions represent less than 2% of the total genome, and little is known about the functional consequences of variation in the remaining 98% of the genome. In support of the role of non-coding variants in human disease and phenotypic traits, it has been observed that most disease-associated GWAS variants lie within non-coding regions of the genome, particularly within transcriptional regulatory regions (promoters, enhancers and boundary elements). Hence, we propose that non-coding variants located at *cis*-regulatory regions of BrS-associated genes could be a yet unexplored cause of BrS.

Objectives

In the present work, we aim to profile, for the first time, the genetic variation found at *cis*-regulatory elements of BrS-associated genes and propose possible candidate variants for future functional studies. To achieve this goal, we propose to develop the following objectives:

1. To design a targeted high-throughput sequencing approach (Regulome-seq) based on the pre-selection of putative *cis*-regulatory regions of BrS-associated genes.
2. To screen genetic variants at selected *cis*-regulatory regions (Regulome-seq regions) in a cohort of 89 BrS *SCN5A*-negative individuals.
3. To characterize genetic variation identified in the Regulome-seq regions.
4. To propose possible candidate non-coding variants related to BrS pathogenesis.

III. Materials and Methods

1. Materials

All reagents used in this thesis were purchased at Sigma-Aldrich® unless stated otherwise.

1.1. Samples

1.1.1. Brugada syndrome cohort

Our study cohort consists of 89 unrelated *SCN5A*-negative individuals that were diagnosed with BrS on the basis of a positive BrS EKG, either spontaneous or induced by flecainide or ajmaline administration (**Table 5**). The genetic analysis for the *SCN5A*-negative classification was previous to this thesis and was performed at the Cardiovascular Genetics Center (CGC; Girona). Blood samples from the 89 BrS individuals were received at the CGC in 4 mL EDTA Anti-Coagulant BD Vacutainer tubes. Genomic DNA from blood samples was extracted using Chemagic MSM I instrument (PerkinElmer®) following manufacturer's recommendations and sequenced by conventional Sanger sequencing or massively parallel sequencing. Sanger sequencing was carried out on an 3130XL Genetic Analyzer (Applied Biosystems™) using 31 different pairs of primers covering the *SCN5A* coding regions. Massively parallel sequencing was carried out on a MiSeq (Illumina®) using custom panels developed by Gendiag.exe SL and commercialized by Ferrer InCode. These custom panels include several genes associated to cardiac arrhythmias.

Genomic DNA from BrS individuals is stored at -80ºC for long-term storage.

This investigation conforms the ethical guidelines of the Declaration of Helsinki 2008. All patients signed an informed written consent to participate in the study (**Annex 1**) and all procedures were approved by the ethical committee of University Hospital Dr. Josep Trueta of Girona (Spain).

Table 5. BrS demographic data.

	Male	Female	Total
Number of individuals	70 (78.65%)	19 (21.35%)	89
Age at diagnosis*	46 ±11.98	52 ± 11.60	48 ± 11.87
Symptoms (syncope)	14 (20%)	5 (26.32%)	19 (21.35%)
Familial Sudden Death	21 (30%)	5 (26.32%)	26 (29.21%)
Type 1 EKG	28 (40%)	8 (42.1%)	40 (44.94%)
Type 2 EKG	18 (25.71%)	2 (10.53%)	20 (22.47%)
Type 3 EKG	22 (31.43%)	7 (36.84%)	29 (32.58%)
ICD implanted	28 (40%)	5 (26.32%)	33 (37.08%)
Quinidine treatment	4 (5.71%)	2 (10.53%)	6 (6.74%)

*The results are represented as average ± SD.
EKG (electrocardiogram), ICD (Implantable Cardioverter Defibrillator).

1.1.2. Coriell sample

To determine the quality of the sequencing and to curate the variants identified in the cohort of 89 BrS individuals, we included a Coriell sample (NA12249) in our study (**Table 6**). The genomic DNA from this sample was directly obtained from the Coriell Institute for Biomedical Research Biobank (New Jersey, Camden, USA) and was processed in parallel with the DNA samples from our study cohort of 89 BrS individuals.

The selection of this particular sample was based on the availability of its genotype information at the 1000 Genomes repository (generated from whole genome and exome sequencing and validated using SNP arrays).

Table 6. Coriell NA12249 sample information.

	Information
Original repository	NIGMS Human Genetic Cell Repository
Tissue	Blood
Cell type	B-lymphocyte
Sample type	Genomic DNA
Gender	Female
Age	Unknown
Diagnosis	Unknown

NIGMS (National Institute of General Medical Sciences).

1.1.3. Healthy-aging (Wellderly) cohort

To facilitate the selection of candidate variants associated to BrS phenotype, we compared the genetic variants identified in the BrS cohort with the genetic variants obtained from 200 Wellderly individuals (**Table 7**). These subjects, were part of a bigger cohort of 1,354 individuals that were recruited based on their healthy-aging phenotype by the group of Dr. Eric Topol (Scripps Research Institute, La Jolla, CA, USA)[241]. The Wellderly phenotype was defined as individuals who, at the time of recruitment, were >80 years old with no chronic diseases and were not taking chronic medications.

From the total 1,354 Wellderly subjects, we only used the genomic data of those 200 that were sequenced with the same Illumina platform used in this thesis, thus facilitating the comparison of sequencing data.

The genetic information from the 200 Wellderly individuals was obtained after signing a Material Transfer Agreement with the group of Dr. Eric Topol.

Table 7. Wellderly demographic data (provided by Dr. Manuel Rueda; Scripps Research Institute, CA, USA).

	Male	Female	Total
Number of individuals	69 (34.50%)	131 (65.50%)	200
Age*	84 ± 4.32	85 ± 5.26	85 ± 4.96
Height (cm)*	175 ± 6.09	160 ± 7.75	165 ± 10.46
Weight (kg)*	76 ± 11.80	60 ± 8.85	65 ± 12.60
Never smoked*	36 (52.17%)	75 (57.25%)	111 (55.50%)

*The results are represented as average ± SD.

1.2. Experimental cell models

1.2.1. H9c2 embryonic rat ventricle cells

To evaluate the binding effects of the 59 CTCF-overlapping variants identified in the 89 BrS individuals, we performed luciferase reporter assays in H9c2 cells. This cell line is a sub-clone derived from the original clone BDXI established by Kimes and Brandt in 1976[242], which was originated from embryonic rat ventricle cells.

1.2.2. iPS-derived cardiomyocytes

To obtain the binding profiles of several cardiac TFs (GATA4, GATA6 and NKX2.5), we performed ChIP-seq experiments in iPS-derived cardiomyocytes. These, were kindly provided by Dr. Farah Sheikh (University of California, San Diego; UCSD, CA, USA). To generate iPS-derived cardiomyocytes, skin fibroblasts obtained from a healthy donor were reprogrammed to iPS cells at the University of California, Los Angeles (UCLA, CA, USA). Then, iPS cells were transferred to Dr. Farah's laboratory and differentiated into cardiomyocytes following a protocol adapted from Zanella *et. al.,*[243] (**Figure 51**). Briefly, the differentiation protocol started at day 0 with a mesodermal induction of iPS cells when they reached 80-90% of confluence. The medium used for mesodermal induction was CD3i medium (**Table 8**) supplemented with 6 μM of CHiR99021, an inhibitor of the GSK3 pathway. At day 2, differentiation of iPS to cardiomyocytes was activated by changing to CD3 medium (**Table 8**) supplemented with 5 μM of C59, an inhibitor of the Wnt pathway. Differentiation continued by maintaining the cells in CD3 medium from day 4 to day 20 and changing the medium every two days. Finally, after day 20, maturation of cardiomyocytes was carried out in Maturation Medium (**Table 8**) and cardiomyocytes were maintained in this medium changing it every 2-3 days.

We decided to use these iPS-derived cardiomyocytes because Dr. Farah's protocol allowed the high-throughput differentiation of iPS cells (in 6-well plates), providing more than enough cardiomyocytes for our ChIP-seq experiments.

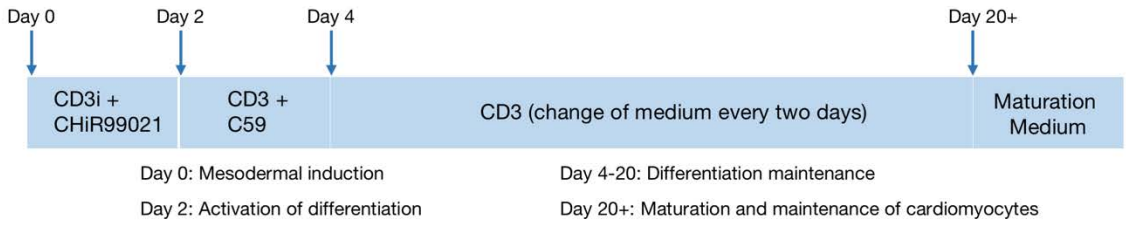


Figure 51. Protocol for iPS differentiation into cardiomyocytes. Scheme of the protocol used in Dr. Farah Sheik's laboratory for differentiation of human iPS cells to cardiomyocytes**.**

Table 8. Media required to differentiate iPS cells to cardiomyocytes.

Medium	Composition
CD3	- 500 mL of RPMI 1640 (with L-glutamine; Life Technologies™) - 500 µg/mL of Human Serum Albumin - 256 µg/mL of L-Ascorbic acid 2-phosphate - 1X of penicillin/streptomycin (Corning)
CD3i	- 50 mL of CD3 medium with 10 µg/mL of insulin (Life Technologies™)
Maturation Medium	- 360 mL of DMEM with L-glutamine and without sodium pyruvate (Corning) - 125 mL of M199 with L-glutamine (Corning) - 500 µg/mL of Human Serum Albumin - 256 µg/mL of L-Ascorbic acid 2-phospate - 10 µg/mL of insulin (Life Technologies™) - 1X of penicillin/streptomycin (Corning)

1.3. Luciferase reporter assay constructs

To experimentally evaluate the binding effects of the 59 CTCF-overlapping variants identified in the 89 BrS individuals, we used the following vectors:

A) pMIR-E-hCTCF-VP64 vector

This vector was generated by fusing the pMIR plasmid expressing the human CTCF (pMIR-E-hCTCF) with the VP64 trans-activator from the lenti dCAS-VP64 (#61425, Addgene) as explained in Methods section 2.8.1. The pMIR-E-hCTCF vector was kindly provided by Dr. Evgin Destici (UCSD, CA, USA).

CTCF is a dynamic TF displaying a context-dependent activity, and can act as an activator, repressor or insulator. By fusing CTCF to VP64, we generate a CTCF-VP64 protein that, when binding to the CTCF binding site, will always act as an activator. The pMIR-E-hCTCF-VP64 vector also confers ampicillin resistance.

B) pGL4.23_variant vectors

These vectors were obtained by cloning 36 bp sequences–containing each of the 59 CTCF-overlapping variants identified in the Regulome-seq regions of 89 BrS individuals–into a pGL4.23 vector (#E8411, Promega) as explained in Methods section 2.8.2. For each variant, the 36 bp sequences (harboring either the reference or alternative allele) were cloned upstream of the pGL4.23 minimal promoter, regulating the expression of the firefly luciferase reporter gene. The pGL4.23_variant vector confers ampicillin resistance.

C) EF-1α-renilla vector

This vector encodes the renilla luciferase reporter gene under the control of the human elongation factor 1 alpha. It also confers ampicillin resistance[244]. The EF-1α-renilla was used to normalize luciferase activity using a Dual-Luciferase Reporter Assay System (Promega).

1.4. Primers and DNA sequences

1.4.1. Primers to generate luciferase reporter assay constructs

All the primers required to obtain the vectors for the experimental evaluation of the 59 CTCF-overlapping variants are described in this section.

1.4.1.1. Primers for site-directed mutagenesis

To generate the pMIR-E-hCTCF-VP64 vector, we first mutated the CTCF stop codon TGA from the pMIR-E-hCTCF vector to GGA (Glycine). We decided to mutate TGA to GGA because Glycine is the smallest amino acid and, therefore, its incorporation into the resultant protein was expected to have a minor effect on the tridimensional folding of the expressed CTCF.

The primers used for site-directed mutagenesis (**Table 9**) were designed using the QuikChange® Primer Design program from Agilent (http://www.genomics.agilent.com/primer designProgram.jsp) and were synthesized by Integrated DNA Technologies®.

Table 9. Primers designed to mutate the CTCF stop codon.

Name	Sequence (5' – 3')
t4250g-Fw	TGATGGACCGGGGAGCGGCCGCC
t4250g-Rv	ACTACCTGGCCCCTCGCCGGCGG

The underlined sequence corresponds to the GGA codon.

1.4.1.2. Primers for PCR-amplification of VP64

To obtain the pMIR-E-hCTCF-VP64, we also PCR-amplified the VP64 trans-activator from the lenti dCAS-VP64 using the primers described in **Table 10**. Amplification primers also introduced the NotI restriction site (GC/GGCCGC), that was further used to ligate the PCR-amplified VP64 with the mutated pMIR-E-hCTCF. The primers were synthesized by Integrated DNA Technologies®.

Table 10. Primers used to PCR-amplify VP64.

Name	Sequence (5' – 3')
VP64-Fw	tcaagtcaGCGGCCGCAGGATCCGGACGGGCTGACGCATTGGACGAT
VP64-Rv	cagtgctgcCGCCGGCGTCAGTTAATCAGCATGTCCAGGTC

The underlined sequence corresponds to NotI restriction site.

1.4.1.3. Oligonucleotides containing CTCF-overlapping variants

To obtain the pGL4.23_variant vectors containing the 59 CTCF-overlapping variants, we designed oligonucleotides with the same 36 bp sequences analyzed by DeepBind (**Figure 52** and **Annex 2**). Oligonucleotides were ordered as single-stranded DNA sequences designed to be ready for cloning after annealing. Reverse oligonucleotides contained four extra nucleotides at both ends (overhangs) that did not anneal with the forward strand. These overhangs were complementary to the pGL4.23 vector digested with KpnI (GGTAC/C) and NheI (G/CTACC). In addition, an extra nucleotide was introduced at both ends of all forward and reverse oligonucleotides to originate a completely different restriction site after ligation with the vector.

For each of the 59 CTCF-overlapping variants, we designed 2 pairs of oligonucleotides (forward and reverse): one containing the reference allele and the other containing the alternative allele, giving a total number of 118 pairs of oligonucleotides to be cloned into the pGL4.23 vector (59 variants x 2 alleles = 118 sequences). All the 118 oligonucleotides were synthesized by Conda® laboratories.

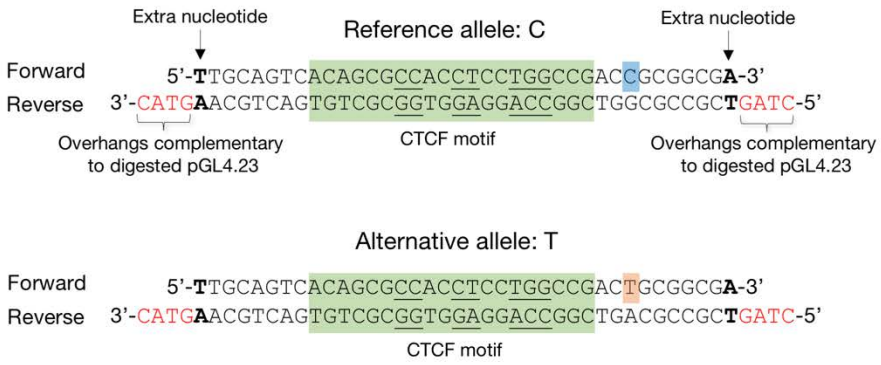


Figure 52. Example of two pairs of oligonucleotides (reference and alternative) used to generate pGL4.23_variant vectors. Top: oligonucleotide pair containing the reference allele (blue). **Bottom:** oligonucleotide pair containing the alternative (orange). Only the reverse oligonucleotides contain 5' and 3' overhangs required for ligation with the pGL4.23 vector (shown in red). Both forward and reverse oligonucleotides include an extra nucleotide introduced to originate a different restriction site after ligation with the pGL4.23 vector (shown in bold). The CTCF motif is highlighted in green and the core nucleotides are underlined.

Using the same strategy aforementioned, we also designed two pairs of oligonucleotides containing either the CTCF consensus motif or a scramble sequence (**Table 11**). The oligonucleotides were synthesized by Conda® laboratories.

Table 11. Oligonucleotides designed to verify the luciferase assay strategy to quantify CTCF binding.

Name	Sequence (5' – 3')
CTCF_consensus-Fw	**T**GCCCCCTGGTGG**A**
CTCF_consensus-Rv	CTAG**T**CCACCAGGGGGC**A**GTAC
CTCF_scramble-Fw	**T**AGTGCGTATGGCAG**A**
CTCF_scramble-Rv	CTAG**T**CTGCCATACGCACT**A**GTAC

Only the reverse oligonucleotides contain 5' and 3' overhangs required for ligation with the pGL4.23 vector (shown in red). Both forward and reverse oligonucleotides include an extra nucleotide introduced to originate a different restriction site after ligation with the pGL4.23 vector (shown in bold). The CTCF core nucleotides are underlined in the CTCF_consensus oligonucleotides.

1.4.2. Primers for Sanger sequencing validation

All the primers used to validate different steps followed for the generation of the luciferase reporter assay constructs are detailed in the table below. These primers were synthesized by Integrated DNA Technologies®.

Table 12. Primers designed for sanger sequencing validation.

Experiment	Sequence (5' – 3')	Hybridization site
pMIR-E-hCTCF mutagenesis	AAACAGAACCAGCCAACAGC	Downstream of CTCF
pMIR-E-hCTCF-VP64 ligation	TTCCTATGCCTACTGCCTCG	Upstream of the VP64 ligation site
pGL4.23_variant ligation	CTCGAAGTACTCGGCGTAGG	Downstream of the ligation site

1.4.3. Indexing adapters for Nextera Rapid Capture

We used the Nextera Rapid Capture (NRC) Custom Enrichment Kit (Illumina®) to prepare DNA libraries from the 89 BrS individuals and the Coriell NA12249 (Methods section 2.2).

To allow the sequencing of multiple samples at the same time (multiplexing), we introduced specific indexing adapters (index 1 and index 2) to each sample by PCR (**Tables 13** and **14**). These adapters also contain a consensus sequence complementary to the flow cell adapters required for cluster generation and sequencing. Index 1 and index 2 adapters are provided as part of the NRC Custom Enrichment Kit (Illumina®).

Table 13. NRC index 1 adapters added to 5' ends of fragmented DNA.

Index	Sequence (5' – 3')
E517	AATGATACGGCGACCACCGAGATCTACACGCGTAAGATCGTCGGCAGCGTC
E502	AATGATACGGCGACCACCGAGATCTACACCTCTCTATTCGTCGGCAGCGTC
E503	AATGATACGGCGACCACCGAGATCTACACTATCCTCTTCGTCGGCAGCGTC
E504	AATGATACGGCGACCACCGAGATCTACACAGAGTAGATCGTCGGCAGCGTC
E505	AATGATACGGCGACCACCGAGATCTACACGTAAGGAGTCGTCGGCAGCGTC
E506	AATGATACGGCGACCACCGAGATCTACACACTGCATATCGTCGGCAGCGTC
E507	AATGATACGGCGACCACCGAGATCTACACAAGGAGTATCGTCGGCAGCGTC
E508	AATGATACGGCGACCACCGAGATCTACACCTAAGCCTTCGTCGGCAGCGTC

NRC index 1 adapters are composed by a sequence complementary to the transposase adaptor sequences (blue), the index 1 (purple) and a sequence complementary to the flow cell adapter (P5; orange). These adapters are provided as part of the NRC Custom Enrichment kit (Illumina®).

Table 14. NRC index 2 adapters added to 3' ends of fragmented DNA.

Index	Sequence (5' – 3')
N701	CAAGCAGAAGACGGCATACGAGATTCGCCTTAGTCTCGTGGGCTCGG
N702	CAAGCAGAAGACGGCATACGAGATCTAGTACGGTCTCGTGGGCTCGG
N703	CAAGCAGAAGACGGCATACGAGATTTCTGCCTGTCTCGTGGGCTCGG
N704	CAAGCAGAAGACGGCATACGAGATGCTCAGGAGTCTCGTGGGCTCGG
N705	CAAGCAGAAGACGGCATACGAGATAGGAGTCCGTCTCGTGGGCTCGG
N706	CAAGCAGAAGACGGCATACGAGATCATGCCTAGTCTCGTGGGCTCGG
N707	CAAGCAGAAGACGGCATACGAGATGTAGAGAGGTCTCGTGGGCTCGG
N708	CAAGCAGAAGACGGCATACGAGATCCTCTCTGGTCTCGTGGGCTCGG
N709	CAAGCAGAAGACGGCATACGAGATAGCGTAGCGTCTCGTGGGCTCGG
N710	CAAGCAGAAGACGGCATACGAGATCAGCCTCGGTCTCGTGGGCTCGG
N711	CAAGCAGAAGACGGCATACGAGATTGCCTCTTGTCTCGTGGGCTCGG
N712	CAAGCAGAAGACGGCATACGAGATTCCTCTACGTCTCGTGGGCTCGG

NRC index 2 adapters are composed by a sequence complementary to the transposase adaptor sequences (green), the index 2 (purple) and a sequence complementary to the flow cell adapter (P7; orange).

1.4.4. Indexing adapters for ChIP-seq

To obtain the binding profiles of several cardiac TFs (GATA4, GATA6 and NKX2.5), we performed ChIP-seq experiments. During the preparation of ChIP-seq libraries, we ligated unique indexing adapters to each library to allow sample multiplexing. In addition to their function in sample indexing, these adapters also carry a consensus sequence complementary to the flow cell adapters that is required for cluster generation and sequencing.

The indexing adapters used in this thesis were obtained from TruSeq® Sample Prep Kits (Illumina®) and are described in **Table 15**. TruSeq indexing adapters are also paired to a TruSeq universal adapter, described in **Table 16**.

Table 15. TruSeq index adapters added to fragmented DNA.

Index	Sequence (5' – 3')
1	GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTTG
2	GATCGGAAGAGCACACGTCTGAACTCCAGTCACCGATGTATCTCGTATGCCGTCTTCTGCTTG
3	GATCGGAAGAGCACACGTCTGAACTCCAGTCACTTAGGCATCTCGTATGCCGTCTTCTGCTTG
4	GATCGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAATCTCGTATGCCGTCTTCTGCTTG
5	GATCGGAAGAGCACACGTCTGAACTCCAGTCACACAGTGATCTCGTATGCCGTCTTCTGCTTG
6	GATCGGAAGAGCACACGTCTGAACTCCAGTCACGCCAATATCTCGTATGCCGTCTTCTGCTTG
7	GATCGGAAGAGCACACGTCTGAACTCCAGTCACCAGATCATCTCGTATGCCGTCTTCTGCTTG
8	GATCGGAAGAGCACACGTCTGAACTCCAGTCACACTTGAATCTCGTATGCCGTCTTCTGCTTG
9	GATCGGAAGAGCACACGTCTGAACTCCAGTCACGATCAGATCTCGTATGCCGTCTTCTGCTTG
10	GATCGGAAGAGCACACGTCTGAACTCCAGTCACTAGCTTATCTCGTATGCCGTCTTCTGCTTG
11	GATCGGAAGAGCACACGTCTGAACTCCAGTCACGGCTACATCTCGTATGCCGTCTTCTGCTTG
12	GATCGGAAGAGCACACGTCTGAACTCCAGTCACCTTGTAATCTCGTATGCCGTCTTCTGCTTG

TruSeq index adapters are composed by a sequence complementary to the TruSeq universal adapter (underlined sequence), a unique index (red), and a sequence complementary to the flow cell adapter (P7; orange).

Table 16. Universal adapter hybridized to unique index adapters added to fragmented DNA.

Adapter	Sequence (5' – 3')
Universal	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT

TruSeq universal adapter is composed by a sequence complementary to the TruSeq index adapters (underlined sequence), a sequence complementary to the flow cell adapter (P5; orange), and a T nucleotide required for adaptor ligation after A-tailing (blue).

1.4.5. Primers to amplify ChIP-seq libraries

Prior to massively parallel sequencing of the ChIP-seq libraries, we PCR-amplified the libraries using the multiplexing PCR primers from Illumina® (**Table 17**). These pair of primers are complementary to P5 and P7 adaptor sequences from the TruSeq indexing adapters detailed above.

Table 17. Multiplexing PCR primers used to amplify ChIP-seq libraries.

Primer	Sequence (5' – 3')
1.0	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
2.0	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT

1.5. Antibodies for ChIP-seq

To obtain the binding profiles of several cardiac TFs (GATA4, GATA6 and NKX2.5), we performed ChIP-seq experiments in iPS-derived cardiomyocytes using the antibodies described in **Table 18**. In addition to the specific information for each antibody, the table also indicates the protein beads used for the subsequent selection of the immunoprecipitated material as explained in Methods section 2.6.3.

Table 18. Antibodies used for ChIP-seq experiments.

Antibody	Origin	Reference	Amount per ChIP	Protein Beads
α-GATA4	Goat	sc-1237x (Santa Cruz® Biotechnology)	20 µg	Magnetic Protein G (ThermoFisher Scientific)
α-GATA6	Rabbit	sc-9055 (Santa Cruz® Biotechnology)	10 µg	Sepharose Protein A
α-NKX2.5	Goat	sc-8697 (Santa Cruz® Biotechnology)	10 µg	Magnetic Protein G

1.6. Public databases and resources

1.6.1. Identification of Regulome-seq regions

To define the candidate non-coding regions possibly regulating BrS-associated genes (Regulome-seq regions), we used information from human cardiomyocytes (HCMs) from several publicly available datasets detailed in **Table 19**.

Of note, when this project was designed, there was no information available of long-range chromatin interactions for cardiac cells and, therefore, we used information from human Embryonic Stem Cells (hESCs). However, all the figures and subsequent information regarding TADs presented in this thesis are based on two recent publications for chromatin interactions in left[245] and right[246] ventricles.

Table 19. Summary of public available data used for the selection of Regulome-seq regions.

Experiment	Information	Cell line	GEO Accession	Published
Hi-C	Long range chromatin interactions	hESCs	GSM862723 (Rep1) GSM892306 (Rep2)	Dixon *et al.,*[80]
DHS-seq	Genome TF binding occupancy	HCMs	GSM736516 (Rep1) GSM736504 (Rep2)	Maurano *et al.,*[247]
ChIP-seq	CTCF binding sites	HCMs	GSM1022657 (Rep1) GSM1022677 (Rep2)	Wang *et al.,*[248]
	H3K4me3	HCMs	GSM945308 (Rep1+2)	Thurman *et al.,*[249]

Rep1 and Rep2 refer to replicas 1 and 2, respectively.

1.6.2. Validation of the Regulome-seq design

To test the validity of our Regulome-seq approach in identifying disease-associated regulatory regions, we downloaded a copy of all those SNPs that have been associated to cardiac conduction defects from GWAS Catalog (https://www.ebi.ac.uk/gwas/).

1.6.3. Annotation of variants

The variants identified in the Regulome-seq regions of 89 BrS individuals were annotated using three different databases: the 1000 Genomes database, the Single Nucleotide Polymorphism database (dbSNP), and the Genome Aggregation database (gnomAD).

1.6.3.1. 1000 Genomes database

The 1000 Genomes Project was the first consortium created to sequence genomes of a large number of people to provide a comprehensive resource on human genetic variation[107]. This database currently contains genetic information from 2,504 individuals from several populations that is openly available to the scientific community[250].

We downloaded all the information available for the human genome GRCh37/hg19 from: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/ALL.wgs.phase3_ shapeit2_mvncall_integrated_v5b.20130502.sites.vcf.gz.

The downloaded information was used for two main purposes: (i) to identify which of the variants reported in this thesis are novel, and (ii) to obtain the AF for American, African, East Asian, European and South Asian populations.

1.6.3.2. Single Nucleotide Polymorphism database

The Single Nucleotide Polymorphism database (dbSNP)[251] is a repository of both single nucleotide substitutions and short indels that was created by the National Center for Biotechnology Information (NCBI; Bethesda, Maryland, US) in collaboration with the National Human Genome Research Institute (NHGRI; Bethesda, Maryland, USA).

We downloaded all the information available at the version dbSNP150 for the human genome GRCh37/hg19 from: ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606_b150_ GRCh37p13/VCF/GATK/All_20170710.vcf.gz.

This information was used to identify which of the variants reported in this thesis are novel.

1.6.3.3. Genome Aggregation database

The Genome Aggregation database (gnomAD)[145] is a resource containing 123,136 exome sequences and 15,496 whole-genome sequences from unrelated individuals that were sequenced as part of various disease-specific and population genetic studies. The gnomAD aggregates information from several databases such as the 1000 Genomes database.

We downloaded all the information available for the chromosomes 3, 7, 10, 11 and 12 for the human genome GRCh37/hg19 from:
https://console.cloud.google.com/storage/browser/gnomadpublic/release/2.0.2/vcf/genomes/?pli=1.

This information was used for two main purposes: (i) to identify which of the variants reported in this thesis are novel, and (ii) to obtain the AF for American, African, East Asian, European and South Asian populations.

1.6.4. Ancestry admixture

To determine the ancestral origin of the 89 BrS and 200 Wellderly individuals, we downloaded the genotypes and AF available for all 2,504 individuals sequenced by the 1000 Genomes Project. We downloaded all the information available for chromosomes 3, 7, 10, 11 and 12 for the human genome GRCh37/hg19 from: ftp://ftp.1000genomes.ebi.ac.uk/../vol1/ftp/release/20130502/.

1.6.5. Variant prioritization

1.6.5.1. DeepBind CTCF model

To predict the CTCF binding effects of the 59 CTCF-overlapping variants identified in 89 BrS individuals, we downloaded the DeepBind tool containing the CTCF binding model from:
http://tools.genes.toronto.edu/deepbind/deepbind-v0.11-linux.tgz.

1.6.5.2. Pre-computed scores

To select which genetic variants identified in the Regulome-seq regions of 89 BrS individuals are more likely to have pathogenic effects, we used two different genome-wide pre-computed scores: the CDTS and the CADD Scores. These scores were downloaded for the human genome GRCh37/hg19 from their respective repositories:

A) CDTS

http://www.hli-opendata.com/noncoding/Pipeline/CDTS_diff_perc_coordsorted_gnomAD_N15496_hg19.bed.gz

B) CADD

Indels were scored directly on CADD website (https://cadd.gs.washington.edu/) under version 1.3, while SNVs were scored from SNVs pre-computed scores downloaded from: http://krishna.gs.washington.edu/download/CADD/v1.3/whole_genome_ SNVs.tsv.gz

2. Methods

2.1. Selection of Regulome-seq regions

To define the candidate Regulome-seq regions, we first extracted the GRCh37/hg19 coordinates of all 6 BrS-associated genes from the UCSC genome browser. Then, we extracted the information of long-range chromatin interactions around the 6 BrS-associated genes for hESCs (Methods **Table 19**) and selected the widest TADs to use from the 2 available replicates (~4-7 Mb surrounding each gene). Once defined the domains, we gathered all peak calling results of DHS-seq, CTCF ChIP-seq and H3K4me3 ChIP-seq for HCMs from the ENCODE Project (Methods **Table 19**). We merged all peak files into a single file using mergeBed (Bedtools v2.26.0) and, for each gene, we selected all peaks within its defined domain. Finally, we extracted the DNA sequence of each gene domain and combined the peak coordinates with the DNA sequence of each gene domain to obtain the final list of 1,293 Regulome-seq regions.

2.2. Target sequencing

We used the NRC Custom Enrichment kit (Illumina®) to genotype the Regulome-seq regions from 89 BrS individuals and the Coriell NA12249. NRC is a targeting approach that uses probes pre-designed by the customer to prepare DNA libraries containing only the regions of interest.

2.2.1. Design of Nextera Rapid Capture probes

Capturing probes used by NRC consist of 80-mer single-stranded biotinylated DNA probes that are complementary to the regions of interest. To design non-overlapping probes, we used the web-based tool DesignStudio™ (Illumina®). Briefly, we selected a DNA assay type using NRC technology for the human genome (hg19), uploaded the coordinates of all the 1,293 Regulome-seq regions and selected a standard center-to-center spacing between adjacent probes of 230 bp (**Figure 53**).

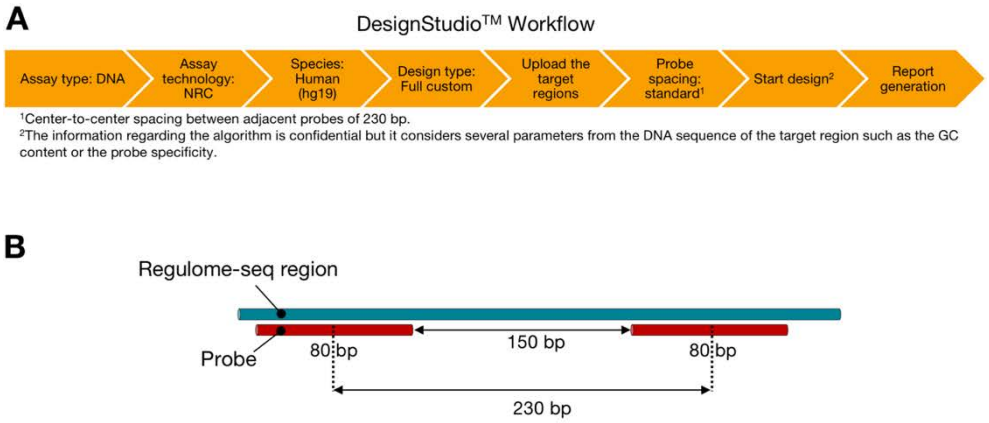


Figure 53. Design of NRC probes. (A) Diagram of the workflow followed on DesignStudio™ and the parameters selected to obtain the NRC probes. **(B)** Schematic representation of the distribution of a hypothetical non-overlapping probes (red) along a Regulome-seq region (blue). The length of the probes and the distance between adjacent probes is represented.

2.2.2. DNA library preparation using Nextera Rapid Capture

DNA library preparation was performed on genomic DNA from the 89 BrS individuals and the Coriell NA12249 (prepared in triplicates). Before starting the protocol, we quantified the genomic DNA of the 89 BrS patients and Coriell sample on a Qubit Fluorimeter (Invitrogen™) using the Qubit® dsDNA Broad Range Assay Kit (Invitrogen™), and visualized its integrity on a 0.8% agarose gel (**Figure 54**).

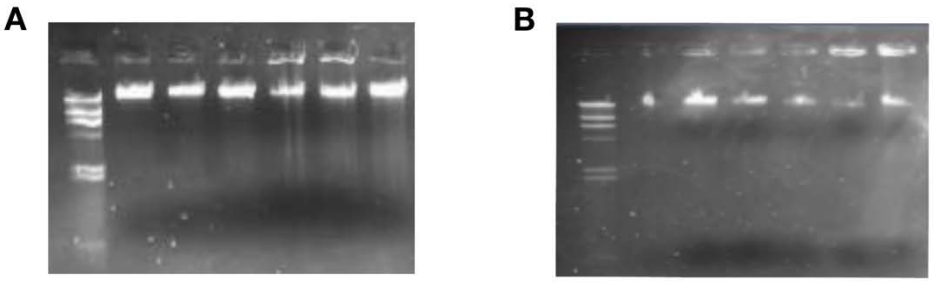


Figure 54. Genomic DNA integrity. Agarose gel showing genomic DNA from different specimens with good **(A)** and low **(B)** quality.

Once confirmed that the genomic DNA from each sample presented a good quality, we started DNA library preparation protocol following the manufacturer instructions, except for a small optimization that we included during DNA fragmentation (described below). Briefly, the NRC protocol can be divided in 5 steps: (i) fragmentation of genomic DNA (tagmentation), (ii) PCR-amplification of fragmented DNA, (iii) hybridization with the NRC probes, (iv) capturing of the NRC probes and, (v) validation of the DNA libraries.

2.2.2.1. Fragmentation of genomic DNA

In the first step, genomic DNA was tagmented with two different Tn5 transposases, enzymes that recognize and cut the AGNTYWRANCT sequence (where N is any nucleotide, R is A or G, W is A or T, and Y is C or T). This sequence is found approximately every 300 bp in the human genome (**Figure 55**). In addition to cut genomic DNA, Tn5 transposases also add an adaptor sequence at both ends of the fragmented DNA that is later used for the PCR amplification step.

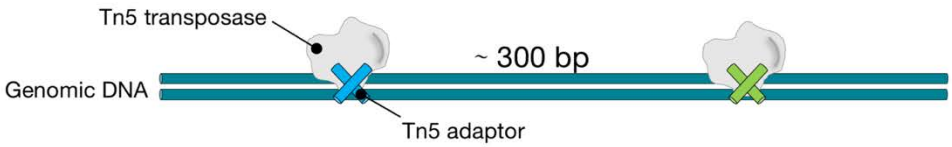


Figure 55. Genomic DNA fragmentation. Schematic representation of two Tn5 transposases with their corresponding adaptor sequences.

To set up the genomic DNA tagmentation reaction, we incubated 10 µL of DNA at 5 ng/µL (50 ng final) with 25 µL of Tagment DNA Buffer, 10 µL of Tagment DNA Enzyme 1, and 5 µL of nuclease-free water for 10 minutes at 58ºC. After the incubation, the tagmentation reaction was stopped by adding 15 µL of Stop Tagment Buffer and incubating for 4 minutes at room temperature.

Fragmented genomic DNA was then purified by adding 65 µL of magnetic sample purification beads. Magnetic beads with the DNA bound were captured in a magnetic rack and cleaned with 80% ethanol. Finally, fragmented genomic DNA was eluted in 20 µL of Resuspension Buffer.

We loaded 1 µL of the eluted DNA on an Agilent Technologies 2100 Bioanalyzer using an Agilent High Sensitivity DNA Chip. We checked that the size of DNA fragments was broadly distributed (around 150 bp-1 kb), with a peak at ~250 bp (**Figure 56**).

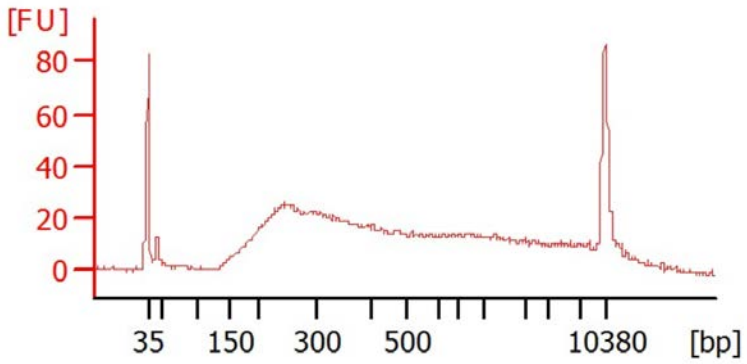


Figure 56. Tagmentation quality control. Example of a Bioanalyzer High Sensitivity DNA Chip for on of the BrS samples showing the amount of DNA (in fluorescence units; y-axis) at each length (x-axis).

2.2.2.2. First PCR amplification

In this step, purified fragmented genomic DNA is PCR-amplified, and specific index sequences (index 1 and index 2) are incorporated into each DNA fragment (Materials **Tables 13** and **14**). Flow cell adapters P5 and P7 required for cluster generation and sequencing are also introduced in this step (**Figure 57**).

Figure 57. PCR amplification of fragmented genomic DNA. Schematic representation of a genomic DNA fragment containing the transposase adaptors added by the Tn5, and the primers for PCR-amplification containing the specific index sequences (index 1 and index 2) and the flow-cell adapters (P5 and P5).

For the first amplification, we added 5 µL of index 1 adapter and 5 µL of index 2 adapter to the 20 µL of fragmented genomic DNA. Different combinations of index 1 and index 2 adapters were introduced for each sample prepared. PCR amplification was performed on a thermal cycler under the following conditions: (i) 3 minutes at 72ºC; (ii) 30 seconds at 98ºC; (iii) 10 cycles of 10 seconds at 98ºC, 30 seconds at 60ºC and 30 seconds at 72ºC; (iv) 5 minutes at 72ºC; and (v) hold at 10ºC.

We purified the PCR-amplified DNA by adding 90 µL of magnetic sample purification beads. Magnetic beads were captured in a magnetic rack and cleaned with 80% ethanol. Finally, PCR-amplified DNA was eluted in 25 µL of Resuspension Buffer.

2.2.2.3. First hybridization with Nextera Rapid Capture probes

In this step, DNA library is mixed with the previously designed NRC probes (**Figure 58**).

Figure 58. Hybridization of indexed genomic DNA with NRC probes. Schematic representation of a genomic DNA fragment hybridizing with a previously designed NRC probe. The probe is labelled with biotin to facilitate the subsequent capturing of the target region using magnetic streptavidin beads.

This step of the protocol requires pooling of a maximum of 12 samples. Therefore, we divided the 89 BrS samples and the triplicates of the Coriell NA12249 in 8 pools: 7 pools of 12 samples and 1 pool of 8 samples. Each pool requires a final DNA concentration of 9 µg, and it is critical

to add the same amount of each sample to avoid any overrepresentation. Hence, we quantified each PCR-amplified sample on a Qubit Fluorimeter (Invitrogen™) using the Qubit® dsDNA Broad Range Assay Kit (Invitrogen™). Then, we calculated the amount of DNA to be added considering the number of samples included in the pool. For the pools of 12 samples, we mixed 750 ng of DNA (9 µg/12 samples = 750 ng) in a final volume of 40 µL. For the pool of 8 samples, we mixed 1.13 µg of DNA (9 µg/8 samples = 1.13 µg) in a final volume of 40 µL.

To set up the hybridization, we added 50 µL of Enrichment Hybridization Buffer and 10 µL of NRC probes to each pool. The hybridization reaction was performed on a thermal cycler under the following conditions: (i) 10 minutes at 95ºC; (ii) 18 cycles of 1 minute incubations starting at 94ºC and decreasing 2ºC per cycle; and (iii) hold at 58ºC.

2.2.2.4. Capturing of Nextera Rapid Capture probes

To capture the NRC probes hybridized to the targeted regions of interest, we added 250 µL of magnetic streptavidin beads to the hybridization reaction and incubated them shaking at 1,200 rpm for 5 minutes at room temperature, followed by a subsequent incubation for 25 minutes at room temperature without shaking.

Streptavidin magnetic beads bound to the NRC probes were placed on a magnetic rack and washed twice in 200 µL of Enrichment Wash Solution for 30 minutes at 50ºC. Finally, the NRC probes hybridized with the target DNA were incubated with 28.5 µL of Enrichment Elution Buffer 1 and 1.5 µL of NAOH [2N] at for 2 minutes at room temperature. After the incubation, we placed the samples on the magnetic rack and transferred 21 µL of the eluted NRC probes to a new tube. We added 4 µL of Elute Target Buffer 2 and 15 µL of Resuspension Buffer.

After this first capture, we repeated the hybridization step with the NRC probes, but in this case holding the reaction at 58ºC overnight (for at least 14.5 hours). Hybridization was also followed by a second capture of the NRC probes. The secondly captured NRC probes hybridized with the target DNA were purified by adding 45 µL of magnetic sample purification beads. The magnetic beads with the DNA bound were captured in a magnetic rack and cleaned with 80% ethanol. Finally, the targeted DNA was eluted in 25 µL of Resuspension Buffer.

2.2.2.5. Second PCR amplification

In this step, the DNA library containing the targeted DNA is PCR-amplified for subsequent sequencing. For this amplification, we added 5 µL of the PCR Primer Cocktail and 20 µL of the Nextera Enrichment Amplification Mix. PCR amplification was performed on a thermal cycler

under the following conditions: (i) 30 seconds at 98ºC; (ii)12 cycles of 10 seconds at 98ºC, 30 seconds at 60ºC and 30 seconds at 72ºC; (iii) 5 minutes at 72ºC; and (iv) hold at 10ºC.

PCR-amplified DNA was purified with 90 µL of magnetic sample purification beads and cleaned with 80% ethanol. Finally, targeted DNA was eluted in 25 µL of Resuspension Buffer.

2.2.2.6. Validation of the DNA libraries

We validated the quality of the DNA libraries by loading 1 µL of the post-enriched library on an Agilent Technologies 2100 Bioanalyzer using an Agilent High Sensitivity DNA Chip. The size of DNA fragments was expected to be distributed around 200 bp-1 kb, with a peak at ~350 bp (**Figure 59**).

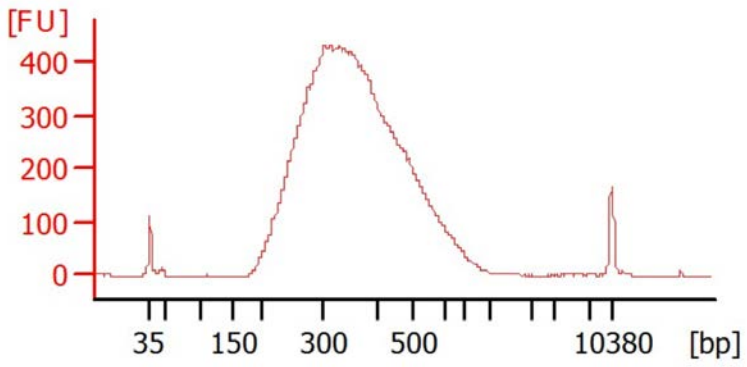


Figure 59. NRC library quality control. Example of a Bioanalyzer High Sensitivity DNA Chip for one pool containing 12 different BrS samples. The amount of DNA (in fluorescence units) is represented on the y-axis, while the length of the fragments is shown on x-axis.

2.2.3. Sequencing Nextera Rapid Capture libraries

We sequenced the resulting DNA libraries in two consecutive sequencing runs: a first run of 6 pools (72 samples), and a second run of the remaining 2 pools (20 samples). To mix the pools to be sequenced together, each pool was first diluted to a final concentration of 10 nM. Then, we mixed 10 µL of each pool into a new tube (one for 72 samples and another for 20 samples) ready for sequencing. These tubes were sent to the Center for Genomic Regulation (CRG, Barcelona, Spain), where it was sequenced on a HiSeq2500 following a paired-end protocol of 100 cycles.

2.3. Read alignment and variant discovery

To identify genetic variants from the 89 BrS individuals and the triplicates of the Coriell NA12249 sample, we followed a three-step analysis, starting from raw reads obtained from the HiSeq2500 and finishing with a list of genetic variants stored in a VCF (**Figure 60**).

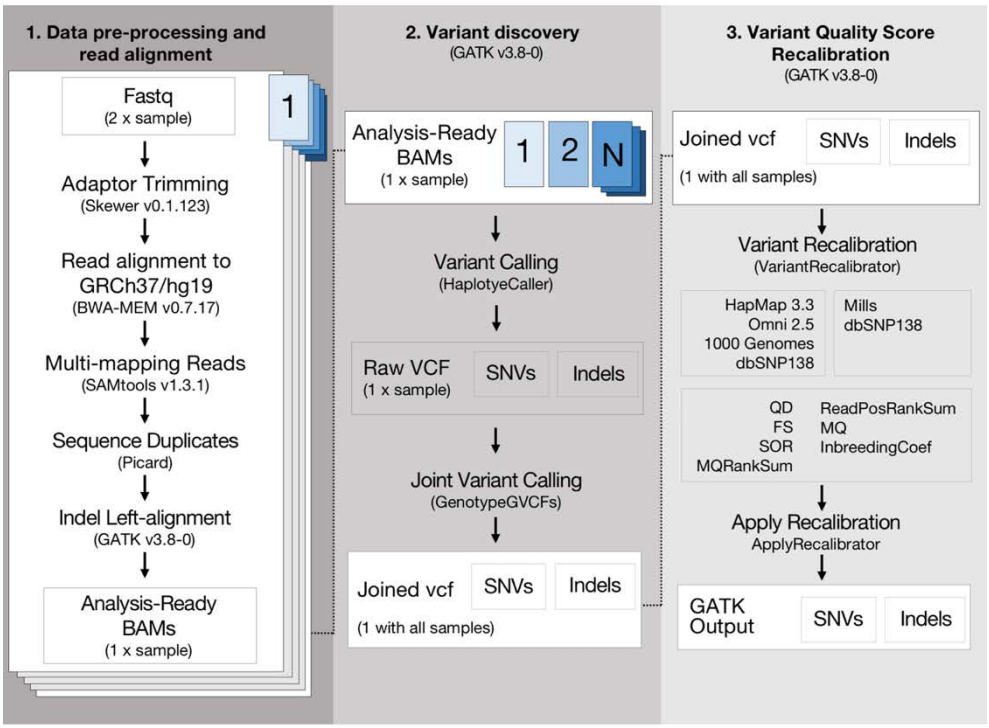


Figure 60. Bioinformatic pipeline followed for Regulome-seq variant discovery. Three-step bioinformatic pipeline followed to identify non-coding variants present in the Regulome-seq regions of 89 BrS individuals and the Coriell NA12249. Figure adapted from GATK website (software.broadinstitute.org/gatk/).

2.3.1. Data pre-processing and read alignment

This step is the first phase in any application designed to identify genetic variants. Since our DNA libraries were sequenced following a paired-end protocol, we obtained 2 Fastq files for each sample corresponding to the forward and reverse reads. The 2 Fastq files from each sample were pre-processed and the reads were aligned to the human reference genome (GRCh37/hg19) to obtain analysis-ready BAM files (1 per sample), which are the starting material for variant discovery.

For Fastq pre-processing, we used Skewer (v0.1.123)[252] to remove adaptor sequences (added during DNA library preparation). We also trimmed-off low quality nucleotides using an in-house perl algorithm.

For read alignment, we used the Burrows-Wheeler Aligner (BWA-MEM; v0.7.17)[253]. After alignment, we used the Sequence Alignment/Map tools (SAMtools; v1.3.1)[254] to remove multi-mapping reads and Picard (v2.18.9)[255] to remove sequencing duplicates.

Finally, as recommended by standard pipelines, we used the Genome Analysis Toolkit (GATK; v3.8-0)[256] to left-align indels at their left-most position possible and avoid multiple representations of the same indel.

The specific commands used for each tool are detailed below:

Skewer:

```
~/Skewer-0.1.123 -x nextera_adaptors.fa -m pe first_fastq second_fastq
-o output_name -t 6
```

BWA-MEM:

```
~/Bwa-0.7.17/bwa mem ucsc.hg19.fasta -M -t 6 trimmed_first_fastq
trimmed_second_fastq -R '@RG\tID:dip1\tSM:dip1' > name.bwa.sam
```

SAMtools:

```
~/Samtools-1.3.1/samtools view -bh -F 256 name.bwa.sam I ~/samtools-
1.3.1/samtools sort -o name.bwa.nomulti.bam
```

Picard:

```
Java -jar Picard/picard.jar MarkDuplicates I=name.bwa.nomulti.bam
O=name.bwa.nomulti.nodup.bam M=name.bwa.duplicate_metrics.log
REMOVE_DUPLICATES=true ASSUME_SORTED=true
```

GATK:

```
java -Xmx20g -jar GenomeAnalysisTK.jar \
-T LeftAlignIndels \
-R ucsc.hg19.fasta \
-I name.bwa.nomulti.nodup.bam \
-o name.bwa.nomulti.nodup.leftAlindel.bam
```

2.3.2. Variant discovery

For variant discovery, we used GATK (v3.8-0)[256] and followed its best practices recommendations (online guide). We first ran the HaplotypeCaller, which uses local *de novo* assembly of haplotypes. Specifically, when the tool encounters a region showing signs of genetic variation, it discards the existing aligned information and completely reassembles the reads in that region. This reassembly allows the HaplotypeCaller to be more accurate when calling regions with SNVs and indels close to each other.

After running the HaplotypeCaller, we merged all genotype VCFs (gVCFs) obtained (1 per sample) into a single VCF (1 for all 89 BrS samples and Coriell NA12249 triplicates) using the GenotypeGVCFs. This tool performs a multi-sample joint aggregation step and merges the records together in a sophisticated manner: at each position of the input gVCFs, it combines all spanning records, produces correct genotype likelihoods, re-genotypes the newly merged record, and then re-annotates it.

The specific commands used for each tool are detailed below:

HaplotypeCaller:

```
java –Xmx20g –jar GenomeAnalysisTK.jar \
-T HaplotypeCaller \
-R ucsc.hg19.fasta \
-I name.bwa.nomulti.nodup.leftAlindel.bam \
--emitRefConfidence GVCF \
-L Regulome-seq.genomic.positions \
-o sample_name.raw.snps.indels.g.vcf
```

GenotypeGVCFs:

```
java –Xmx20g –jar GenomeAnalysisTK.jar \
-T GenotypeGVCFs \
-R ucsc.hg19.fasta \
--variant sample1.raw.snps.indels.gvcf \
--variant … \
--variant sample92.raw.snps.indels.g.vcf
-o Reglome.raw.vcf \
```

2.3.3. Variant Quality Score Recalibration

As part of the best practices, GATK recommends to run a variant recalibration to obtain a more confident variant call set. Briefly, during variant quality score recalibration (VQSR), an adaptive error model is built based on several parameters (**Table 20**) from public resources of known variation (**Table 21**). Once built, the error model is applied to the variants identified after GenotypeGVCF to estimate the probability that each variant is a true genetic variant or a sequencing artifact.

Table 20. Parameters used to build the adaptive error model during variant recalibration.

Parameter	Description
QualByDepth (QD)	Variant confidence obtained from HaplotypeCaller and GenotypeGVCF
FisherStrand (FS) and StrandOddsRatio (SOR)	Measure of strand bias. Both parameters are complementary
MappingQualityRankSumTest (MQRankSum)	Rank sum test for mapping qualities of reads with variants
ReadPosRankSumTest (ReadPosRankSum)	Rank sum test to calculate the distance of variants from the end of the reads
RMSMappingQuality (MQ)	Estimation of the overall mapping quality of reads supporting a variant call
InbreedingCoeff	Evidence of inbreeding in the samples analyzed

Table 21. Public resources used to build the adaptive error model during variant recalibration.

Resource	Application	Type of variant
HapMap_3.3	True variants used for training the recalibration model	SNVs
Omni_2.5	True variants used for training the recalibration model	SNVs
1000 Genomes_phase1	True variants and false positives used for training the recalibration model	SNVs
Mills	True variants used for training the recalibration model	Indels
dbSNP_138	Variants used to stratify output metrics. Not used for training the recalibration model	SNVs and Indels

To generate the adaptive error model, we used the VariantRecalibrator tool, from which we obtain a recalibration file that is then applied to the variants using the ApplyRecalibration tool. SNVs and Indels were recalibrated separately with the same pipeline.

The specific commands used for each tool are detailed below:

VariantRecalibrator (SNVs):

```
java –Xmx20g –jar GenomeAnalysisTK.jar \
-T VariantRecalibrator \
-R ucsc.hg19.fasta \
-input Regulome.raw.vcf \
-resource:hapmap,known=false,training=true,truth=true,prior=15.0
hapmap_3.3.hg19.sites.vcf \
-resource:omni,known=false,training=true,truth=false,prior=12.0
1000G_omni2.5.hg19.sites.vcf \
-resource:1000G,known=false,training=true,truth=false,prior=10.0
1000G_phase1.snps.high_confidence.hg19.sites.vcf \
-resource:dbsnp,known=true,training=false,truth=false,prior=2.0
dbsnp_138.hg19.vcf
-an QD –an MQ –an MQRankSum –an ReadPosRankSum –an FS –an SOR –an
InbreedingCoeff \
-mode SNP \
-recalFile Regulome.snvs.recal\
-tranchesFile Regulome.snvs.tranches \
```

VariantRecalibrator (Indels):

```
java –Xmx20g –jar GenomeAnalysisTK.jar \
-T VariantRecalibrator \
-R ucsc.hg19.fasta \
-input Regulome.raw.vcf \
-resource:mills,known=false,training=true,truth=true,prior=12.0
Mills_and_1000G_gold_standard.indels.hg19.sites.vcf
-resource:dbsnp,known=true,training=false,truth=false,prior=2.0
dbsnp_138.hg19.vcf
-an QD –an MQ –an MQRankSum –an ReadPosRankSum –an FS –an SOR –an
InbreedingCoeff \
-mode INDEL \
-recalFile Regulome.inels.recal \
-tranchesFile Regulome.indels.tranches \
```

ApplyRecalibration (SNVs):

```
java –Xmx20g –jar GenomeAnalysisTK.jar \
-T ApplyRecalibration \
-R ucsc.hg19.fasta \
-input Regulome.raw.vcf \
-recalFile Regulome.snvs.recal \
-tranchesFile Regulome.snvs.tranches \
-o Regulome.recalibrated.snvs.vcf \
-mode SNP
```

ApplyRecalibration (Indels):

```
java –Xmx20g –jar GenomeAnalysisTK.jar \
-T ApplyRecalibration \
-R ucsc.hg19.fasta \
-input Regulome.recalibrated.snvs.vcf \
-recalFile Regulome.indels.recal \
-tranchesFile Regulome.indels.tranches \
-o Regulome.recalibrated.vcf \
-mode INDEL
```

2.4. Variant call quality analysis and data curation

2.4.1. Quality analysis

To determine the quality of the variants obtained following our variant discovery pipeline, we used the sequencing results from the Coriell NA12249 triplicates. We downloaded the list of all SNVs available at the 1000 Genomes Project database for this Coriell and kept only those SNVs within our Regulome-seq regions, obtaining a list of SNVs that was named “public dataset”. In parallel, we also extracted the list of SNVs present in each Coriell NA12249 triplicate from our variant call set (“regulome dataset”). Then, we compared the two datasets and assessed the number of true positives (TP), false positives (FP) and false negatives (FN): TP correspond to those SNVs that are present in both datasets, FP correspond to those SNVs that are only present in the regulome dataset, and FN correspond to those SNVs that are present in the public dataset but we were not able to identify.

Finally, we also measured the sensitivity and the positive predictive value (PPV) of our variant discovery pipeline. The **sensitivity** (TP/(TP + FN) measures the proportion of positive variants that are correctly identified as such. The **PPV** (TP/TP + FP) measures the odds of having a true variant in the final variant call set.

2.4.2. Curation of the BrS variant call set

In order to increase the quality of the final variant call set, we applied several consecutive filters to the VCF obtained following our variant discovery pipeline (**Figure 61**).

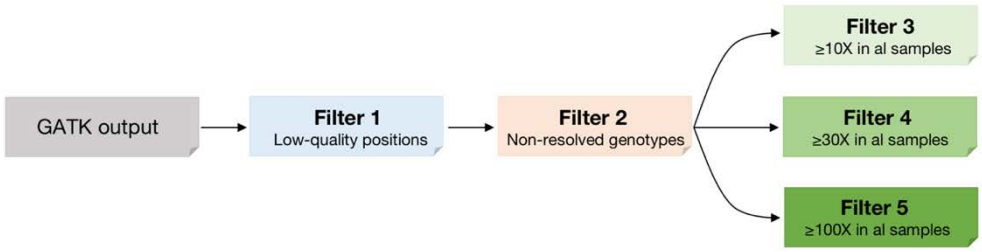


Figure 61. BrS variant call set curation. Diagram showing the consecutive filters applied to increase the quality of the final variant call set.

Each filter was applied to the VCF containing the information of the 89 BrS individuals and the Coriell NA12249 triplicates. However, to determine the quality of the variant call set after each filter, we only used the data obtained from the Coriell. As described above, we also calculated the number of TP, FP and FN and measured the **sensitivity** and **PPV** after applying each filter.

The first filter (**Filter 1**) was applied to the VCF obtained from GATK (GATK output). Here, we removed genomic positions tagged as low-quality after VQSR. In the second filter (**Filter 2**), applied to the Filter 1-VCF output, we removed all genomic positions with a non-resolved genotype–even if the genotype was not resolved in a single sample, the position was removed in all samples–. **Filters 3**, **4**, and **5** consisted of removing all genomic positions that were not covered ≥10X, ≥30X or ≥100X in all samples, respectively.

2.4.3. Curation of the Wellderly variant call set

We received an VCF containing all the genetic variants identified in the Regulome-seq regions in 200 Wellderly individuals (Wellderly raw).

To compare this Wellderly variant call set to our previously-curated BrS call set, we applied several consecutive filters to the raw VCF obtained from the group of Dr. Eric Topol (Wellderly raw; **Figure 62**). To assess the validity of each filter, we tested whether the average number of variants per individual in both cohorts was similar.

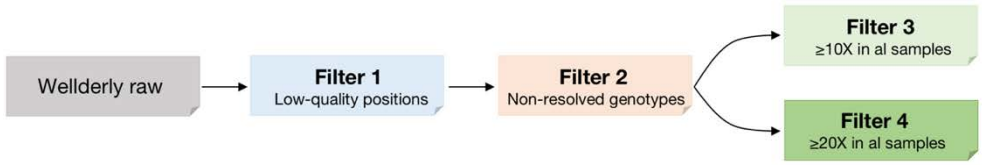


Figure 62. Wellderly variant call set curation. Diagram showing the consecutive filters applied to increase the quality of the final variant call set.

The first and second filters (**Filter 1** and **Filter 2**) were directly applied to the raw Wellderly call set because it was generated following the same variant discovery pipeline as ours. In Filter 1, we removed genomic positions tagged as low quality after VQSR. In Filter 2, applied to the Filter 1-VCF output, we removed all genomic positions with a non-resolved genotype–even if the genotype was not resolved in a single sample, the position was removed in all samples–. **Filters 3** and **4** consisted in removing all genomic positions that were not covered ≥10X, ≥20X in all samples, respectively.

2.5. Analysis of BrS and Wellderly ancestry admixture

To determine the ancestral origin of the 89 BrS and 200 Wellderly individuals, we compared the SNVs in the Regulome-seq regions from these two cohorts with the SNVs in the Regulome-seq regions of 661 African, 347 American, 504 East Asian, 503 European and 489 South Asian individuals, downloaded from the 1000 Genomes database (Materials section 1.6.4). Specifically, we compared the AF of all the SNVs shared by all 7 cohorts. We placed this information in a tab-delimited file containing all 1,983 shared SNVs (columns) and all 2,793 individuals (rows), which we analyzed using the R package Rtnse. Briefly, Rtsne is an R wrapper around the **t-Distributed Stochastic Neighbor Embedding** (**t-SNE**)[257], a technique for dimensionality reduction that is particularly well suited for the visualization of high-dimensional datasets. To visualize the Rtsne results, we used the R package ggplot2.
The commands used in R for t-SNE analysis and visualization are detailed below:

```
#Import the tab-delimited file
mydata <- read.table("tsne.txt", header=TRUE, sep="\t", row.names=1)
matrix <- as.matrix(as.data.frame(mydata[,1:1983]))

#Run Rtsne
library(Rtsne)
set.seed(9)
tsne_model_1 <- Rtsne(matrix, dims=2, perplexity=30, theta=0.0,
check_duplicates=FALSE, pca=FALSE, max_iter=5000, eta=10)

#Plot Rtsne results
library(ggplot2)
d_tsne_1 <- as.data.frame(tsne_model_1$Y)
Population <- mydata$Population
p <- ggplot(d_tsne_1, aes(x=V1, y=V2, colour=Population)) +
geom_point(size=0.5) + xlab("tSNE_1") + ylab("tSNE_2") +
theme_classic(base_size=11) + theme(panel.background =
element_rect(fill="white", color="black")) + theme(legend.direction =
"vertical", legend.position = "right")
```

2.6. ChIP-seq experiments in iPS-derived cardiomyocytes

To profile the binding pattern of several cardiac TFs (GATA4, GATA6 and NKX2.5), we performed ChIP-seq experiments in iPS-derived cardiomyocytes (Materials section 1.2.2).

As explained in the introduction, the ChIP-seq protocol is divided in four steps: (i) cross-linking or fixation of proteins to DNA, (ii) chromatin shearing, (ii) immunoprecipitation using antibodies against the TF/histone modification of interest, and (iv) DNA library preparation to obtain sequencing-ready fragments (**Figure 63**).

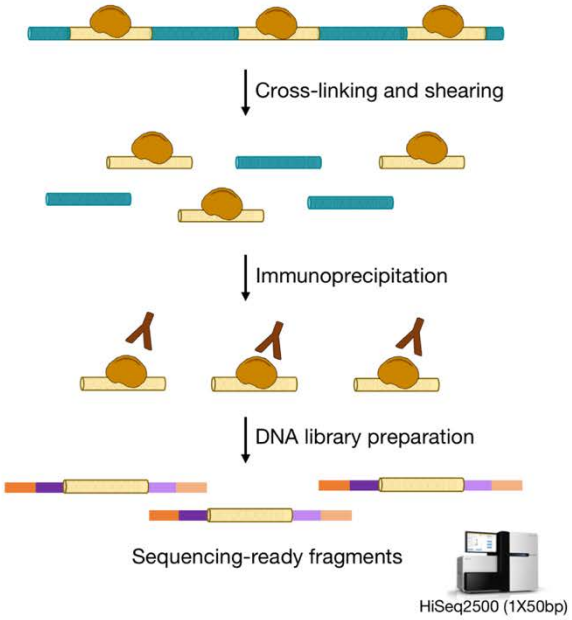


Figure 63. Overview of the ChIP-seq protocol.

2.6.1. Cross-linking

To cross-link the iPS-derived cardiomyocytes (grown in 6-well plates), we incubated the cells for 20 minutes at room temperature in 1 mL of Cross-linking Solution (1% formaldehyde in Phosphate Buffered Saline; PBS). After the incubation, cells were quickly rinsed twice with ice-cold PBS, and the cross-linking reaction was stopped by adding 500 µL of freshly prepared Stop Cross-linking Solution (100 mM TrisHCl pH 9.4 and 10 mM DTT). Cells were harvested using a cell scraper and centrifuged at 4,000 g for 1 minute. Cell pellets were then recovered, washed with 1 mL of PBS and centrifuged at 4,000g for 1 minute.

2.6.2. Chromatin shearing

We followed two different chromatin shearing protocols depending on the type of protein beads required for the immunoprecipitation: sepharose protein A or magnetic protein G (Materials **Table 18**). The two protocols differ in two main aspects: (i) composition of the Lysis Buffer and; (ii) overnight incubation with the antibody and protein beads in the case of protein G samples.

2.6.2.1. Chromatin shearing for sepharose protein A samples

This protocol was applied to all cell pellets immunoprecipitated with rabbit antibodies against GATA6. Cell pellets were re-suspended with 300 µL of Lysis Buffer (0.3% SDS, 10 mM EDTA, 50 mM TrisHCl pH 7.8 and ½ tablet of protease inhibitors) and sonicated with a Bioruptor® NGS (Diagenode) under different number of cycles (30 seconds ON/30 seconds OFF at high frequency). To determine the number of sonication cycles, we took a 5 µL aliquot of sheared chromatin and incubated it with 1µL of Proteinase K (ThermoFisher Scientific) for 5 minutes at 50ºC, followed by an incubation of 15 minutes at 95ºC. We visualized the size of sheared chromatin fragments in a 1% agarose gel that we ran for 30 minutes at 100V: if the chromatin was not properly fragmented, the original sample was sonicated a few more cycles. Then, another 5 µL aliquot was separated and the process was repeated until the strongest signal showed up around 1 Kb on the agarose gel.

Once the sheared chromatin fragments reached the desired size, we centrifuged the chromatin at 14,000 rpm for 10 minutes at 4ºC and divided the supernatant into two new tubes (150 µL each). Then, we diluted the sheared chromatin with 1.35 mL of Dilution Buffer (1% Triton X-100, 2 mM EDTA, 150 mM NaCl, 200 mM TrisHCl pH 7.8 and ½ tablet of protease inhibitors), added the corresponding antibody and incubated the samples overnight at 4ºC under gentle rotation.

2.6.2.2. Chromatin shearing for magnetic protein G samples

This protocol was applied to all cell pellets immunoprecipitated using goat antibodies against GATA4 and NKX2.5. Cell pellets were re-suspended in 300 µL of RIPA Lysis Buffer (500 mM TrisHCl pH 7.4, 1% IGEPAL CA-630, 0.25% Na-Deoxycholate, 150 mM NaCl, 1 mM EDTA, 0.1% SDS and 0.5 mM DTT) and sonicated with a Bioruptor® NGS (Diagenode) under different number of cycles (30 seconds ON/30 seconds OFF at high frequency). To determine the number of sonication cycles, we followed the same workflow as explained for sepharose protein A samples.

Once the sheared chromatin fragments reached the desired size, we centrifuged the chromatin at 14,000 rpm for 10 minutes at 4ºC and divided the chromatin into two new tubes (150 µL each), already containing 30 µL of magnetic protein G with the corresponding antibody. We diluted the sheared chromatin with 1.35 ml of RIPA Lysis Buffer (500 mM TrisHCl pH 7.4, 1% IGEPAL CA-630, 0.25% Na-Deoxycholate, 150 mM NaCl, 1 mM EDTA, 0.1% SDS and 0.5 mM DTT) and incubated the samples overnight at 4ºC under gentle rotation.

In parallel, we washed the 30 µL of magnetic protein G twice with 1 mL of Blocking Solution (0.5% Bovine Serum Albumin in PBS) by inverting the tube and placing it on the magnetic rack. Then, we pre-incubated the magnetic protein G with the corresponding antibody at room temperature for 1 hour. Finally, we added the sheared chromatin to the tubes and incubated the samples overnight at 4ºC under gentle rotation.

2.6.3. Chromatin immunoprecipitation

Here, we also followed two different immunoprecipitation protocols depending on the type of protein beads required for this step: sepharose protein A or magnetic protein G (Materials **Table 18**). The two protocols differ in two main aspects: (i) the composition of the solutions used and; (ii) the manner by which the washings are performed.

2.6.3.1. Chromatin immunoprecipitation for sepharose protein A samples

After overnight incubation with the antibody, we incubated the samples with 50 µL of sepharose protein A beads under rotation for 5-7 hours at 4ºC. Prior to their utilization, we washed the sepharose protein A beads five times with TE by inverting the tube and centrifuging the beads at 3,000 rpm for 30 seconds.

After the incubation with the sepharose protein A beads, we centrifuged the samples for 30 seconds at 2,000 rpm and washed the beads twice with 1.4 mL of Wash Solution (1% Triton X-100, 2 mM EDTA, 150 mM NaCl and 20 mM TrisHCl pH 7.8). Washing steps were performed by incubating the beads under rotation for 15 minutes at room temperature followed by a centrifugation at 2,000 rpm for 30 seconds. We also performed two extra washes with 1.4 mL of TE 1X but this time only inverting the tubes and centrifuging at 2,000 rpm for 30 seconds. Finally, we decross-linked the samples overnight at 65ºC with 300 µL of Decross-link Solution (1% SDS in TE 1X).

2.6.3.2. Chromatin immunoprecipitation for magnetic protein G samples

After overnight incubation with the antibody, samples were washed three times with 150 µL of Wash Buffer I (20mM Tris-HCl pH 7.4, 150 mM NaCl, 0.1% SDS, 1% Triton X-100, 2 mM EDTA) and two times with 150 µL of TET Buffer (0.2% Tween 20 in TE 1X). Washing steps were performed by inverting the tube for 30 seconds and placing them on the magnetic rack.

Then, we eluted the DNA three times with 40 µL of Elution Buffer (1% SDS, 10 mM TrisHCl pH 8.5, 270 mM NaCl and 1 mM EDTA). This elution step was performed by incubating the samples for 5 minutes at room temperature and placing them into the magnetic rack. Finally, we decross-linked the samples overnight at 65ºC with 40 µL of Elution Buffer.

2.6.4. ChIP-seq library preparation

From this point forward, we applied the same protocol to sepharose protein A and magnetic protein G samples with the exception that, before starting with the library preparation protocol, magnetic protein G samples were digested for 30 minutes at 37ºC with 0.5 µL of RNase A (10 mg/mL; ThermoFisher Scientific) and incubated for an additional 1-2 hours at 55ºC with 1 µL of Proteinase K (20 mg/mL; Five Prime®).

To prepare the ChIP-seq libraries ready for massively parallel sequencing, we used the KAPA Library Preparation Kit for Illumina series KK8200 (KAPA Biosystems). Briefly, this protocol consists of 4 main steps: (i) End Repair, (ii) A-tailing, (iii) Adaptor ligation, and (iv) PCR-amplification. We also performed 2 size-selections: one before and one after PCR-amplification.

2.6.4.1. End Repair

This step is used to convert the overhangs generated during chromatin shearing into blunt ends (**Figure 64**).

Figure 64. End Repair of immunoprecipitated DNA fragments. Schematic representation showing 5' and 3' overhangs filling, resulting in DNA fragment with blunt ends.

We purified immunoprecipitated DNA with the QIAquick PCR Purification Kit (QIAGEN) following the manufacturer instructions, and eluting the DNA with 87 µL of Elution Buffer. We quantified eulted DNA on a Qubit Fluorimeter (Invitrogen™) using the Qubit® dsDNA High

Sensitivity Assay Kit (Invitrogen™). When possible, we used 400 ng of DNA for End Repair reaction. Otherwise, we used 85 µL of the eluted DNA.

We prepared the End Repair Mix (10 µL of 10X End Repair Buffer, 5 µL of End Repair Enzyme, 400 ng or 85 µL of DNA, and nuclease-free water until reach a final volume of 100 µL) and incubated the reactions 30 minutes at 20ºC. We purified the end repaired DNA using the MinElute PCR Purification Kit following manufacturer instructions, eluting the DNA with 42 µL of Elution Buffer.

2.6.4.2. A-tailing

This step adds an A nucleotide to the 3' ends of the blunt fragments, which provides a complementary overhang for the T nucleotide found at the 3' end of the TruSeq universal adapter (Materials **Table 16** and **Figure 65**).

Figure 65. A-tailing of end repaired DNA fragments. Schematic representation showing the addition of an A nucleotide to 3' ends of blunt DNA fragments.

We prepared the A-tailing Mix (5 µL of 10X A-tailing Buffer, 3 µL of A-tailing Enzyme and 42 µL of End Repaired DNA) and incubated the reactions 30 minutes at 30ºC. We purified the A-tailed DNA using the MinElute PCR Purification Kit, eluting the DNA with 34 µL of Elution Buffer.

2.6.4.3. Adapter ligation

In this step, TruSeq indexing adapters (Materials **Table 15**) are added to the ends of the DNA fragments. TruSeq indexing adapters are paired with the TruSeq universal adapter (Materials **Table 16**) by the 12 nucleotides located at their 5' end, leaving Y-shaped overhangs (**Figure 66**). The TruSeq universal adapter carries a T nucleotide at the 3' end that is used to ligate the Y-shaped adapters to the A-tailed DNA fragments.

Indexing adapters also carry the flow cell adapters P5 and P7 required for cluster generation and sequencing.

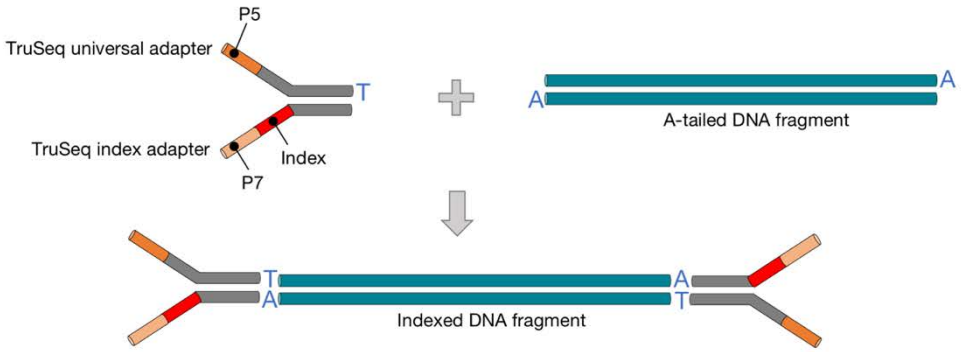


Figure 66. Adapter ligation to A-tailed DNA fragments. Schematic representation showing the Y-shaped adapters (composed by TruSeq universal adapters and TruSeq index adapters; top left) and the A-tailed DNA (top right). The resulting indexed DNA fragment is shown at the bottom of the figure.

We prepared the Adapter ligation Mix (10 µL of 5X Ligation Buffer, 5 µL of DNA Ligase, 1 µL of DNA adapters and 34 µL of A-tailed DNA) and incubated the reactions overnight at 16ºC on a water bath. We purified the adapter ligated DNA using the MinElute PCR Purification Kit, and eluting the DNA with 16 µL of Elution Buffer.

2.6.4.4. Size-selection

In this step, we select DNA fragments between 200-400 bp. We ran the 16 µL of Adaptor ligated DNA with 3 µL of ethidium bromide (Invitrogen™) on a 2% pure agarose gel (UltraPure™ Agarose; ThermoFisher Scientific) prepared in TAE 1X buffer (40 mM Tris-acetate and 1 mM EDTA). As a marker for the size selection, we used the TrackIt 1 Kb Plus DNA Ladder (ThermoFisher Scientific).

Importantly, running of the agarose gel required several critical points: (i) we loaded the samples using agarose gels not covered by the TAE 1X buffer of the electrophoresis tray; (ii) we let the samples ran on the dry gel for 5 minutes at 50V; and (iii) we added the remaining TAE 1X to cover the gel and let the samples ran at 100V until they reached three-quarters of the agarose gel.

To cut the 200-400 bp bands, we placed the gel on top of a UV light source and cut the bands using a scalpel. Finally, we purified the DNA from the gel using the QIAquick Gel Extraction Kit (QIAGEN) following manufacturer instructions, eluting the DNA with 46 µL of Elution Buffer.

2.6.4.5. PCR amplification

In this PCR amplification step, those DNA fragments that have adapter molecules on both ends are selectively enriched and the amount of DNA of the final library is amplified.

We prepared the PCR-amplification Mix (25 µL of 2X KAPA HiFi HotStart Ready Mix, 2 µL of primers (Materials **Table 17**) and 23µL of DNA) and amplified the PCR reactions on a thermal cycler (BIO-RAD). Thermal cycler conditions were the following: (i) 45 seconds at 98ºC; (ii) 16-20 cycles of 15 seconds at 98ºC, 30 seconds at 60ºC and 30 seconds at 72ºC; (iii) 1 minute at 72ºC; and (iv) hold at 4ºC. In this PCR-amplification, the number of cycles was defined based on the band intensity observed in the size selection.

The remaining size-selected DNA that was not used for PCR-amplification was stored at -20ºC in case this step had to be repeated.

2.6.4.6. Second size-selection

In this step, DNA fragments are again selected in a range between 200-400 bp. For this size selection we followed the same protocol detailed in section 2.6.4.4 with the exception that we eluted the DNA in 15 µL of Elution Buffer.

2.6.5. Sequencing of ChIP-seq libraries

We sequenced the GATA4, GATA6 and NKX2.5 ChIP-seq libraries in a single sequencing run performed at the CRG. To pool the corresponding libraries, we first diluted them to a final concentration of 5 nM, and then mixed 5 µL of each library into a new tube that was ready for sequencing. Independently of the facility where the run took place, ChIP-seq libraries were sequenced on a HiSeq2500 following a single-read protocol of 50 cycles.

2.6.6. Analysis of ChIP-seq data

The analysis of the ChIP-seq data was performed by Dr. Daria Merkurjev at UCLA, CA, USA. Briefly, the ChIP-seq Fastq files were aligned to the human reference genome hg18 using Bowtie2 (v2.3.4.1)[258]. After the alignment, SAMtools (v1.3.1) was used to convert the Sequence Alignment Map (SAM) file to bed format and remove duplicated reads. Then, HOMER (v4.8)[97] was applied to find the ChIP-seq peaks (regions with TF binding) and create files that could be directly visualized at the UCSC genome browser. Finally, for each ChIP-seq sample, TF motifs were identified using HOMER (v4.8) to validate the quality of the whole ChIP-seq technique. We

only accepted those ChIP-seqs where the highest enriched motif corresponded to the known binding motif for the TF immunoprecipitated for subsequent analysis.

2.7. Obtention of DeepBind CTCF predictions

To predict the effects of the 59 CTCF-overlapping variants on CTCF binding, we followed two main steps. First, we defined a list of CTCF binding sites from HCMs genome-wide. Then, we integrated this information with the genetic variants from the 89 BrS individuals and obtained binding predictions using DeepBind[159] (Materials section 1.6.5.1).

2.7.1. Defining a list of CTCF binding sites

For this purpose, we downloaded the peak files of the two CTCF ChIP-seq replicas from HCMs (Materials **Table 19**) and merged them into a single file. We also accepted peaks found only in one replica (private peaks) because, even these private peaks can derive from technical artifacts, they can also derive from biological aspects (i.e. genetic variants from HCMs affecting the CTCF binding).

After the merged file was obtained, we performed a motif analysis using HOMER (v4.8). Here, we only accepted those peaks with a CTCF motif to facilitate direct association of genetic variants to binding effects. Finally, we redefined the peaks as 36 bp sequences surrounding the CTCF motif.

2.7.2. DeepBind scores

The obtention of CTCF binding scores for each of the 59 CTCF-overlapping variant was performed by Bernat del Olmo and Dr. Jesús Matés at the CGC using an in-house perl wrapper (DeepBindTK v1.0) around DeepBind. Briefly, DeepBindTK combined the CTCF peak file with the variants identified in 89 BrS individuals and created a list of haplotypes that were loaded into DeepBind.

DeeBind was used under the CTCF model in default parameters and DeepBind scores were visualized in a variation map using GNUplot (v4.6).

2.8. Generation of vectors for luciferase reporter assays

To experimentally validate DeepBind predictions in a luciferase reporter assay, we prepared the pMIR-E-hCTCF-VP64 and the pGL4.23_variant vectors (Materials section 1.3).

2.8.1. Generation of pMIR-E-hCTCF-VP64 vector

This vector expresses the human CTCF fused to the VP64 trans-activator to suppress the CTCF context-dependent activity: since the fusion protein will act as a potent transactivator when bound to the CTCF motif, the designed system allows to only assess the CTCF DNA binding capacity. To generate this vector, we mutated the hCTCF stop codon from the vector pMIR-E-hCTCF and digested the mutated vector. In parallel, we PCR-amplified VP64 from the vector lenti dCAS-VP64. Finally, we ligated the mutated hCTCF vector with the PCR-amplified VP64.

2.8.1.1. pMIR-E-hCTCF site-directed mutagenesis

We mutated the CTCF stop codon TGA from the vector pMIR-E-hCTCF to GGA (Glycine) using the QiaQuick® Lightning Site-Directed Mutagenesis kit (Agilent). We prepared the site-directed mutagenesis reaction as follows: 5 µL of 10X reaction buffer, 1 µL of 100 ng of vector, 125 ng of t4250g-Fw and t4250g-Rv primers (Materials **Table 9**), 1 µL of dNTP mix, 1.5 µL of QuickSolution reagent, 1 µL of QuikChange Lightning Enzyme, and distilled water until a final volume of 50 µL. The reaction was incubated on a thermal cycler under the following conditions: (i) 2 minutes at 95ºC; (ii) 18 cycles of 20 seconds at 95ºC, 10 seconds at 60ºC, 5 minutes at 68ºC; (iii) 10 minutes at 68ºC; and (iv) hold at 10ºC.

Then, we digested the supercoiled parental DNA by incubating the site-directed mutagenesis reactions for 5 minutes at 37ºC with 2 µL of DpnI enzyme. After digestion, we transformed 45 µL of XL10-Gold Ultracompetent Cells (Agilent) with 2 µL of the DpnI-treated DNA, and incubated for 30 minutes on ice. We heat-shocked the cells for 30 seconds in a 42ºC water bath and placed the tubes for 2 minutes on ice. We recovered the cells by adding 0.5 mL of preheated (42ºC) NZY^{+} broth and incubating them at 250 rpm for 1 hour at 37ºC. Finally, we harvested all the recovered volume into LB-Agar plates (35 g/L) containing 100 µg/mL ampicillin, and incubated the plates overnight at 37ºC.

The colonies that grew on the LB-Agar plates were clonal-amplified (as detailed in Methods section 2.9) and checked by Sanger sequencing using the primers from Materials **Table 12**.

2.8.1.2. Digestion of mutated pMIR-E-hCTCF

The mutated vector pMIR-E-hCTCF was digested with NotI enzyme (GC/GGCCGC; New England Biolabs®) in a reaction containing: 2 µL of the mutated pMIR-E-hCTCF (final concentration of 2 µg), 1 µL of NotI, 2 µL of Buffer 3.1 (New England Biolabs®), 1 µL of Alkaline Phosphatase (CIP; New England Biolabs®) and nuclease-free water until 20 µL. We incubated

the digestion reaction for 4 hours at 37ºC and then purified the linearized vector using the MinElute PCR Purification Kit (QIAGEN), eluting the DNA with 40 µL of Elution Buffer.

2.8.1.3. PCR-amplification of VP64

The VP64 trans-activator was PCR-amplified from the lenti dCAS-VP64 vector using the primers described in Materials **Table 10**. The PCR reaction was prepared as follows: 2.5 µL of VP64-Fw and VP64-Rv primers (10 µM final concentration), 25 µL of Phusion® High-Fidelity PCR Master Mix (New England Biolabs®), 1 µL of lenti dCAS-VP64 vector (10 ng final concentration) and nuclease-free water until 50 µL. The PCR reaction was placed on a thermal cycler under the following conditions: (i) 30 seconds at 98ºC; (ii) 35 cycles of 10 seconds at 98ºC, 30 seconds at 60ºC and 30 seconds at 72ºC; (iii) 5 minutes at 72ºC; and (iv) hold at 4ºC.

We loaded the PCR reaction on a 1% agarose gel with 3 µL of ethidium bromide (Invitrogen™), which ran at 100V for 1 hour. Then, we cut the band with a scalpel on top of a UV light source and purified the DNA using the QIAquick Gel Extraction Kit (QIAGEN), eluting the DNA with 50 µL of Elution Buffer.

2.8.1.4. Digestion of PCR-amplified VP64

Since the primers used for VP64 PCR amplification also introduced the NotI restriction site, we digested the amplified VP64 in a reaction containing 17 µL of the VP64, 1 µL of NotI and 2 µL of Buffer 3.1. We incubated the digestion reaction for 4 hours at 37ºC and then purified the DNA using the MinElute PCR Purification Kit (QIAGEN), eluting the DNA with 40 µL of Elution Buffer.

2.8.1.5. Ligation of pMIR-E-hCTCF with VP64

To obtain the final pMIR-E-hCTCF-VP64, we ligated the pMIR-E-hCTCF with VP64. Ligation reactions were as follows: 1 µL of pMIR-E-hCTCF (final concentration of 50 ng), 1.4 µL of VP64 (final concentration of 10 ng), 1 µL of T4 DNA ligase (New England Biolabs®), 1.5 µL of T4 DNA ligase buffer (New England Biolabs®) and 10.5 µL of nuclease-free water. Ligation reactions were incubated overnight at 16ºC.

Ligation reactions were clonal-amplified (as detailed in Methods section 2.9) and checked by Sanger sequencing using the primers from Materials **Table 12**.

2.8.2. Generation of pGL4.23_variant vectors

The pGL4.23 vector was used to clone the 36 bp sequences from which we obtained the CTCF binding predictions using DeepBind (see Methods section 2.7 for more details). To clone these sequences into the pGL4.23 vector, we designed 36 bp single-stranded oligonucleotides containing the sequences tested in DeepBind (Materials section 1.4.1.3). These oligonucleotides were annealed to generate double-stranded DNA fragments that were ligated with the previously digested pGL4.23 vector. Oligonucleotides were ligated into the pGL4.23 vector Multiple Cloning Site, located upstream of the minimal promoter regulating the expression of the luciferase reporter gene.

2.8.2.1. Oligonucleotide annealing

The annealing reaction of single strand oligonucleotides was prepared as follows: 1 µL of both forward and reverse oligonucleotides, 1 µL of T4 Ligase Buffer (New England Biolabs®), 0.5 µL of PNK Enzyme (New England Biolabs®) and 6.5 µL of nuclease-free water. Annealing reaction was carried out on a thermal cycler under the following conditions: (i) 30 minutes at 37ºC; (ii) 5 minutes at 95ºC; (iii) 12 seconds incubations starting at 95ºC and decreasing 1ºC until 25ºC; and (iv) hold at 4ºC.

Then, 1 µL of annealed oligonucleotides was diluted 250-fold with nuclease-free water and stored at -20ºC.

2.8.2.2. pGL4.23 vector digestion

The pGL4.23 vector was double-digested with KpnI (GGTAC/C) and NheI (G/CTACC) enzymes (New England Biolabs® in a reaction containing: 2 µL of pGL4.23 (final concentration of 2 µg), 1 µL of KpnI, 1 µL of NheI, 2 µL of Buffer 1.1 (New England Biolabs®) and nuclease-free water until 20 µL. Digestion was performed for 4 hours at 37ºC and checked in a 1.2% agarose gel. The digested vector was purified using QIAquick Gel Extraction kit (QIAGEN), eluting the vector in 40 µL of Elution Buffer.

2.8.2.3. Ligation of pGL4.23 with each annealed oligonucleotide

To generate the final pGL4.23_variant vectors, we ligated digested pGL4.23 vectors with each of the annealed double-stranded oligonucleotides. The ligation reaction was prepared as follows: 1 µL of pGL4.23 (final concentration of 50 ng), 1 µL of diluted double-stranded oligonucleotides, 1 µL T4 ligase (New England Biolabs®), 1 µL of T4 Ligase Buffer (New England

Biolabs®) and nuclease-free water until 10 µL. Ligations were performed for 1 hour at room temperature.

Ligation reactions were clonal-amplified (as detailed in Methods section 2.9) and checked by Sanger sequencing using the primers from Materials **Table 12**.

2.9. Clonal amplification of luciferase expression vectors

2.9.1. Transformation of ligation reactions into competent cells

All ligation reactions (Materials sections 2.8.1.1, 2.8.1.5 and 2.8.2.3) were transformed into DH5α Max Efficiency Competent Cells (Invitrogen™) following manufacturer instructions. Briefly, we transformed 1 µL of ligation reactions into 16 µL of DH5α cells. Bacterial cells were incubated for 30 minutes on ice, heat-shocked for 30 seconds in a 42ºC water bath and placed for another 2 minutes on ice. We recovered the cells by adding 250 µL of S.O.C. Medium (Invitrogen™) and incubating them shaking at 220 rpm for 1 hour at 37ºC. Once recovered, we plated the transformed cells into LB-Agar plates (35 g/L) containing 100 µg/mL ampicillin and incubated overnight at 37ºC.

2.9.2. Miniprep

From the colonies that grew after transformation into competent cells, we picked up two from each LB-agar plate and incubated them in 4 mL of LB-Broth (20 g/L) containing 100 g/mL ampicillin shaking at 220 rpm overnight at 37ºC.

Plasmid DNA was purified from 3 mL of the culture, and the remaining 1 mL was stored at 4ºC for the next step. To purify plasmid DNA, we used the reagents from EndoFree® Plasmid Midi Kit (QIAGEN) but with a different protocol. Briefly, we spun down the culture, and re-suspended the cell pellets with 100 µL of Buffer P1 by vortexing until no clumps remained. Bacterial cells were lysed by incubating them in 150 µL of Buffer P2 for 5 minutes at room temperature. We stopped the lysis reaction by adding 100 µL of Buffer P3. We centrifuged the cells at maximum speed for 5 minutes and transferred the supernatant into a new tube. DNA was then precipitated by adding 350 µL of isopropanol and centrifuging the samples at maximum speed for 5 minutes. After removing supernatant, we washed the precipitated DNA with 200 µL of 70% ethanol and centrifuging at maximum speed for 5 minutes. Finally, supernatants were removed and the precipitated DNA was air-dried to be then re-suspended in 20 µL of Elution Buffer.

The resulting vector DNA was analyzed by Sanger sequencing using the primers detailed in Materials **Table 12**.

2.9.3. Midiprep

Once the correct sequences were confirmed, the remaining 1 mL bacterial culture that was not used for Miniprep was incubated by shaking it at 220 rpm for 6 hours at 37ºC in 4 mL of sterile LB-Broth (20 g/mL) with 100 µg/mL of ampicillin. The total volume was transferred into a glass erlenmeyer containing 100 mL of sterile LB-Broth (20 g/L) with 100 µg/mL of ampicillin and incubated at 220 rpm overnight at 37ºC.

The grown bacterial culture was used to purify the vector DNA of interest using the QIAGEN Plasmid Midi Kit (QIAGEN) following manufacturer instructions, eluting the DNA with 50 µL of endotoxin-free Buffer TE.

2.9.4. Glycerol stocks

Before starting with the Midiprep, we placed 750 µL of the overnight culture in a cryo-tube containing 250 µL of sterile 60% glycerol. We stored the tubes at -80ºC as a long-term vector stocks.

2.10. Maintenance and subculture of H9c2 cells

Luciferase reporter assays were performed in H9c2 cells. Cells were subcultured every two days and long-term storage vials were also prepared to guarantee H9c2 maintenance. The sub-culturing protocol and the protocols to freeze and defrost long-term storage vials are detailed below.

2.10.1. Subculture of H9c2 cells

H9c2 cells were grown in Dulbecco's Modified Eagle's Medium (DMEM) supplemented with 10% fetal bovine serum, 1% penicillin/streptomycin and 1% Glutamax (Invitrogen™). Cells were seeded in 75 cm^2 flasks (to grow as an adherent monolayer) and incubated at 37ºC with 5% CO_2. When they reached a 70-80% confluence, we removed the medium and rinsed the cells with 10 mL of PBS. We detached the cells from the flask by incubating them with 0.8 mL of Tripsin-EDTA 0.05% (Gibco™) for 2 minutes at 37ºC. Then, the tripsin reaction was stopped with 10 mL of supplemented DMEM medium to avoid the reduction of cell viability due to

prolonged Tripsin. Finally, we re-seeded the cell suspension in a new 75 cm^2 flask in a 1:5 proportion (cell suspension: supplemented DMEM medium).

2.10.2. Freezing of H9c2 cells in liquid nitrogen

To ensure a long term maintenance of the cells, we prepared 1 mL stocks that were stored in liquid nitrogen.

To prepare the cells for freezing, we first removed the medium from the H9c2 cell culture maintained in 75 cm^2 flasks and rinsed the cells with 10 mL of PBS. Cells were detached from the flask with tripsin as described above. After cell re-suspension, we measured the cell concentration using an automated cell counter (Scepter™ 2.0 Handheld Automated Cell Counter, Merck Millipore). We centrifuged the cells at 1,500 rpm for 5 minutes and we re-suspended the cell pellet with the Cell Freezing Medium-Serum-free to a final concentration of 4,000,000 cells/mL. Then, we divided the cell suspension in 1 mL cryo-tube aliquots that we placed at -80ºC in a freezing recipient with propanol (Nalgene® Mr. Frosty, ThermoFisher Scientific). This procedure guarantees a gradual decrease of the temperature inside the cell cryo-tube to avoid the formation of intracellular ice crystals that could harm the cells. After 24 hours, we placed the cells in liquid nitrogen for long-term storage.

2.10.3. Thawing of H9c2 cells stored in liquid nitrogen

We rapidly thawed the H9c2 long-term stocks on a 37ºC water bath to avoid the intracellular ice crystal formation. Thawed cells were carefully re-suspended with 10 mL of supplemented DMEM medium and centrifuged at 1,500 rpm for 5 minutes. We re-suspended again the cell pellet with 10 mL of supplemented DMEM medium and seeded the cells in a 75 cm^2 flask. After 24 hours, we replaced the old medium with fresh supplemented DMEM medium.

2.11. Luciferase reporter assay

To measure the binding effects of the 59 CTCF-overlapping variants identified in the Regulome-seq regions of 89 BrS individuals, we performed luciferase reporter assays. Plasmids transfected for luciferase reporter assays are detailed in Materials section 1.3.

2.11.1. Cell transfection

The day before transfection, we counted H9c2 cells using a Scepter™ 2.0 Cell Counter (Merck Millipore) and seeded 1 mL of cells (final concentration of 60,000 cells/mL) in 12-well plates. Cells were placed in a CO_2 incubator for 24 hours.

On transfection day, we prepared a different transfection reaction for each of the 36 bp sequences to be tested. Each DNA transfection tube contained 2 µL of pGL4.23_variant (final concentration of 200 ng), 2 µL of pMIR-E-hCTCF-VP64 (final concentration of 200 ng), 1.5 µL of EF-1α-renilla (final concentration of 25 ng) and 100 µL of Optimem (Gibco™). In parallel, for each transfection reaction, 0.82 µL of Lipofectamine® 2000 (Invitrogen™) were mixed with 100 µL of Optimem. The volume of Lipofectamine® 2000 was determined as a ratio of 2 µL of lipofectamine: 1µg of DNA lipofectamine. Both tubes were incubated for 5 minutes at room temperature. Then, the tubes were mixed an incubated for another 20 minutes at room temperature. During the 20 minutes incubation, the medium was changed to antibiotic-free DMEM medium (supplemented with 10% Fetal Bovine Serum and 1% Glutamax) to avoid the antibiotic to interfere with the DNA-lipofectamine complex. Approximately 6 hours post-transfection, the antibiotic-free medium was replaced by supplemented DMEM medium.

Each transfection reaction was prepared in triplicates.

2.11.2. Measurement of luciferase activity

The firefly and renilla luciferase activities were measured 48 hours after transfection in a GloMax-96 luminometer (Promega) using a Dual Luciferase Reporter Assay System (Promega). First, we washed the cells twice with 2 mL of PBS. Then, cells were lysed with 100 µL of Passive Lysis Buffer by incubating them at room temperature shaking for 15 minutes. After incubation, we harvested the cells using a cell scraper and centrifuged them for 2 minutes at 4ºC at maximum speed. We transferred the supernatant into a new tube and used 10 µL of this supernatant to measure the firefly and renilla luciferase activities. Measurements were recorded in 96-well plates specific for the GloMax-96 luminometer, where 50 µL of both the Luciferase Assay Reagent and the Stop&Glo Reagent were automatically injected.

To measure the CTCF binding activity at each 36 bp sequences tested, the firefly luciferase activity for each sequence was normalized to the renilla activity.

2.12. Statistical analysis

All the statistical analyses except the permutations were conducted with R (v 3.3.3). Differences were considered significant at $p \leq 0.05$ (*), $p \leq 0.01$ (**), and $p \leq 0.001$ (***).

2.12.1. One-way ANOVA

We conducted a one-way ANOVA to compare the mean of one group to the mean of every other group analyzed. If the differences between groups were significant, the homogeneity of variances was tested using Levene's test for homogeneity of variance. When the variances between groups were homogeneous, the analysis continued with a Tukey Honest Significant Differences test. Otherwise, one-way ANOVA test was discarded and the comparison was followed by a Welch pairwise t-test with Benjamini-Hochberg p-value correction for multiple testing. The commands introduced in R for each test are detailed below:

One-way ANOVA:

```
#Import the data
mydata <- read.table("ANOVA.txt", header=TRUE, sep="\t")

#Run ANOVA
ANOVA <- aov(mydata$numerical_variable ~ mydata$categorical_variable,
data=mydata)
summary(ANOVA)
```

Levene's test for homogeneity of variance:

```
#Continue with the same data previously imported
#Run Levene's test
library(car)
leveneTest(mydata$numerical_variable ~ mydata$categorical_variable,
data=mydata)
```

Tukey Honest Significant Differences:

```
#Continue with the same data previously imported

#Run Tukey test using the previous computed ANOVA as an argument
TukeyHSD(ANOVA)
```

Welch pairwise t-test with Benjamini-Hochberg p-value correction:

```
#Continue with the same data previously imported

#Run Welch test
pairwise.t.test(mydata$numerical_variable, mydata$categorical_variable,
p.adjust.method="BH", pool.sd=FALSE)
```

2.12.2. Two-tailed t-test

For pairwise comparisons of the mean of one group to the mean of another group, a two-tailed t-test assuming unequal variances was conducted. The command introduced in R is detailed below:

Two-tailed t-test:

```
#Import the data
mydata <- read.table("t_test.txt", header=TRUE, sep="\t")

#Run t-test
t.test(mydata$group1, mydata$group2, alternative=c("two.sided"), mu=0,
paired=FALSE, var.equal=FALSE)
```

2.12.3. Permutations

To identify which of the variants shared between BrS and Welldelry cohorts were significantly enriched among BrS individuals, we performed 100 permutations using an in-house perl script. Each permutation consisted on a random selection of 89 Wellderly individuals and the measurement of the AF of each variant. AF was calculated as: (nº of homozygous x2 + nº of heterozygous) / total nº of alleles.

For each variant, a significance p-value was obtained using the formula: nº of times the AF was higher in BrS cohort / total nº of permutations.

2.12.4. Pearson's chi-squared test

To identify which of the variants specific for the BrS cohort (BrS-specific) were significantly enriched among BrS individuals, a Perason's chi-squared test was conducted. For each variant, we compared the number of reference and alternative alleles in each cohort. Given that the variants tested are not present in the Wellderly cohort, its number of reference alleles is 400 (200 individuals x2 alleles each) and its number of alternative alleles is 0. The command introduced in R, together with an example are detailed below:

Pearson's chi-squared test:

```
chisq.test(data.frame(alt=c(nº_BrS,nº_Wellderly), ref=c(nº_BrS,
nº_Wellderly)))

#Example
chisq.test(data.frame(alt=c(5,0), ref=c(173,400)))
```

IV. Results

1. Defining the Regulome-seq regions

In order to catalogue genetic variants at non-coding regulatory regions of six BrS-associated genes (*SCN5A, SCN2B, SCN3B, CACNA1C, CACNB2* and *CACNA2D1*), we designed a strategy based on a pre-selection of candidate regions to selectively capture and sequence these regions in 89 BrS individuals (**Figure 67**). We called this strategy **Regulome-seq**, and we therefore refer to the pre-selected non-coding regulatory regions as **Regulome-seq regions**.

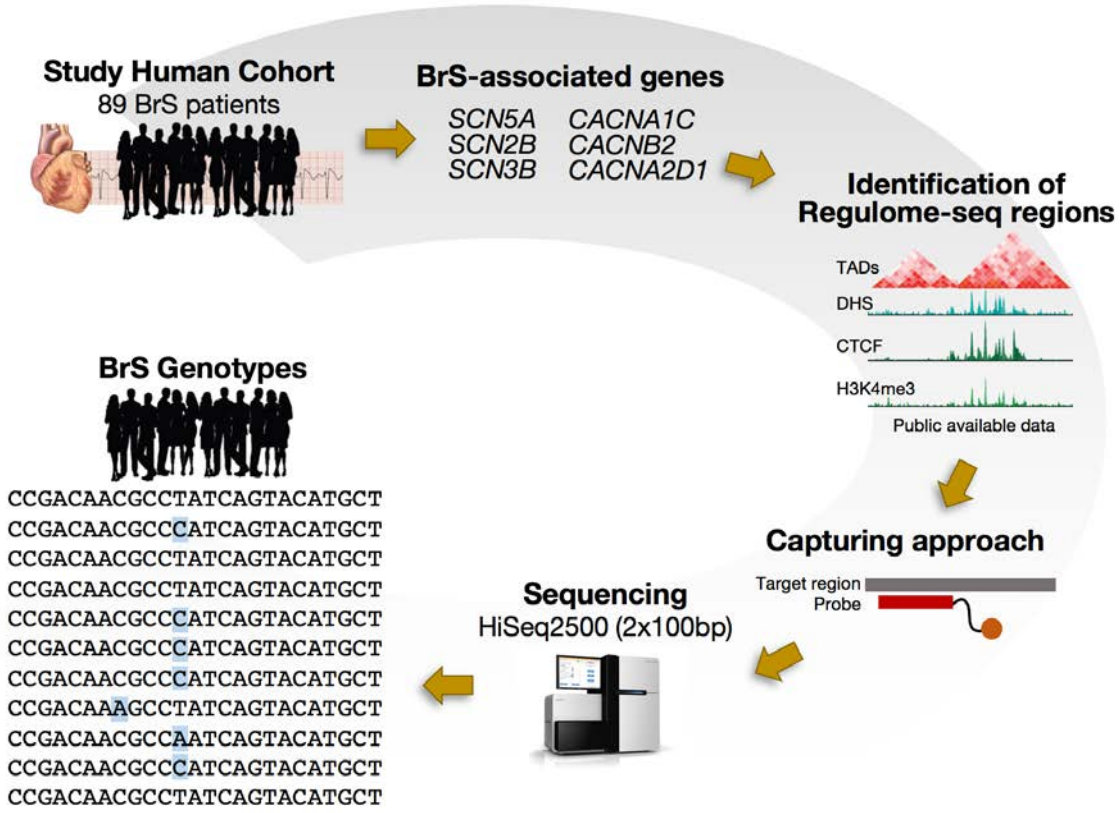


Figure 67. Summary of the Regulome-seq approach.

1.1. Using available information of long-range chromatin interactions

The exact location of non-coding regulatory regions linked to BrS-associated genes remains currently unknown. We therefore conceived an approach to pre-select the regions that putatively regulate BrS-associated genes based on currently available information (Materials **Table 19**).

As it has been observed in recent Hi-C data analysis, the human genome is organized in TADs that are involved both in the co-regulation of genes and in blocking the interaction of regions between neighboring TADs[28]. Hence, regulatory elements found in the same TAD of a

particular gene are more likely to be regulating that gene than other regulatory elements found in neighboring TADs.

When this project was designed, Hi-C information from cardiac tissue was not available. Hence, and given that TADs are invariant across cell types[85], we used published Hi-C data for hESCs to establish the potential regulatory limits of transcriptional regulatory regions associated to each candidate gene (Materials **Table 19**). For each BrS-associated gene, we selected the widest TADs from the 2 available replicates, corresponding to the specific TAD in which each gene is embedded, plus an extended region upstream and downstream (~4-7 Mb surrounding each gene; **Figure 68** and Methods section 2.1). The most extensive region was found at the *CACNA2D1* locus (6.8 Mb) while the least extensive region was found at the *SCN5A* locus (3.8 Mb). In total, for the six BrS-associated genes, the resulting defined regions span a total of 28.4 Mb, which represent less than 1% of the whole human genome (**Figure 69**).

It is important to note that, at a more advanced phase of the project, Hi-C information for left[245] and right[246] ventricles was published and therefore, we based all the figures presented in this thesis on this recent data.

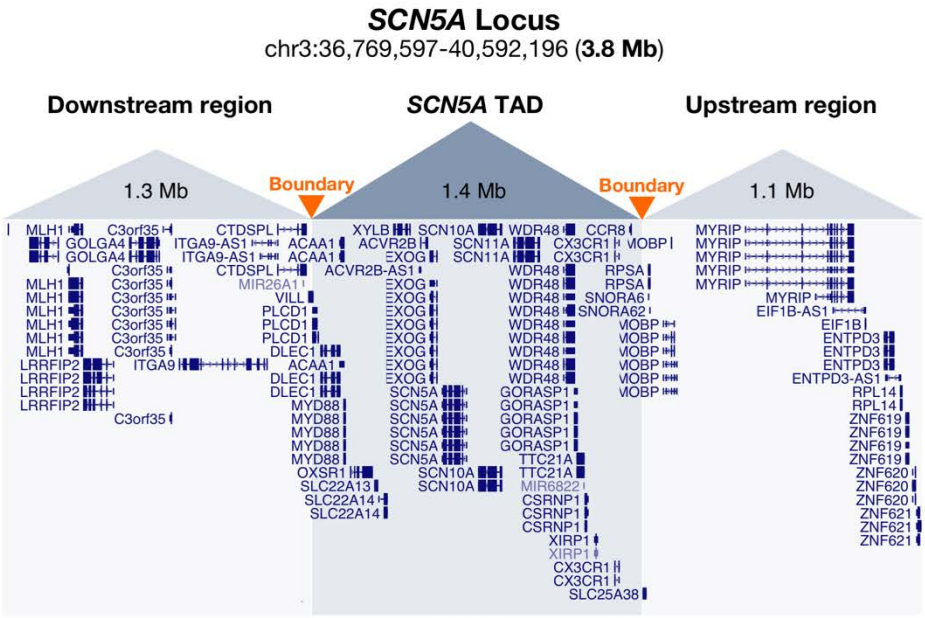


Figure 68. Example of the *SCN5A* TAD selection and the extended upstream and downstream regions. UCSC genome browser tracks showing all gens embedded in the 3.8 Mb selected surrounding the *SCN5A* gene (found in the middle of the figure).

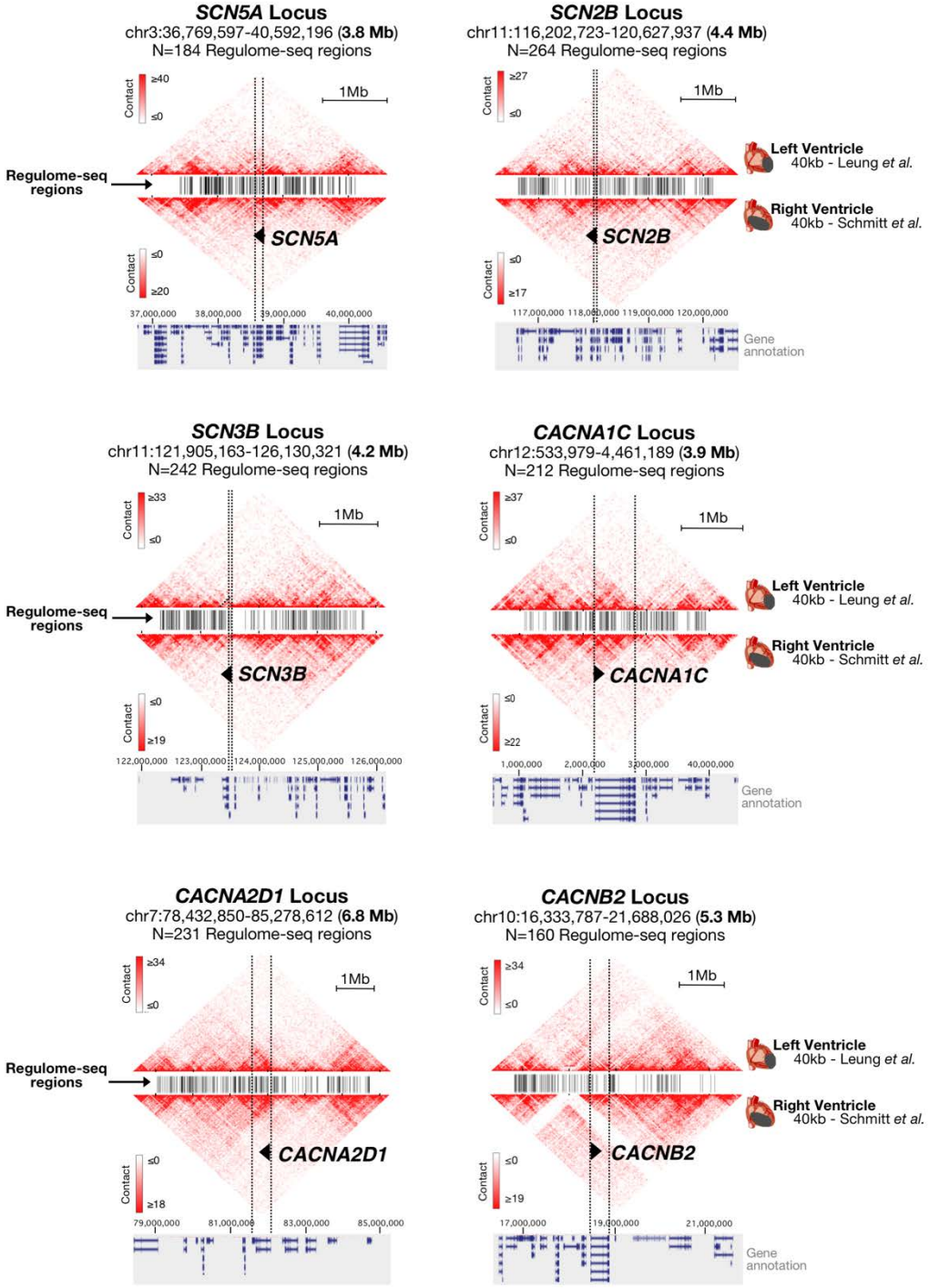


Figure 69. Chromatin interactions at ~4-7 Mb surrounding each BrS-associated gene in human heart. Normalized Hi-C interaction frequencies at a 40 kb resolution, displayed as a two-dimensional heat maps for left (top) and right ventricles (bottom). Heat maps are overlayed on all RefSeq genes present in each loci but only the position and transcriptional direction of the gene of interest is highlighted. Separating the two-dimensional heat maps, Regulome-seq regions are also represented. Heat maps were generated using the 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. Hi-C data for left ventricle was extracted from Leung *et al*.,[245] while right ventricle data was from Schmitt *et al.*[246].

1.2. Using available information of *cis*-regulatory elements

Once defined the TADs, we determined the location of potential *cis*-regulatory regions within each TAD using data of chromatin accessibility (DHS-seq), histone marks (H3K4me3 ChIP-seq), and TF binding (CTCF ChIP-seq) from HCMs obtained from the ENCODE Project (Materials **Table 19**). More specifically, DHS are genomic regions sensitive to enzymatic cleavage due to nucleosome displacement by TF binding. Since regulatory regions act via TF binding, we used DHS to identify active *cis*-regulatory regions. H3K4me3 profiles were used to reveal the presence of active promoters. Finally, CTCF profiles were used to identify TAD boundaries and other classes of *cis*-regulatory regions (**Figure 70** and Methods section 2.1).

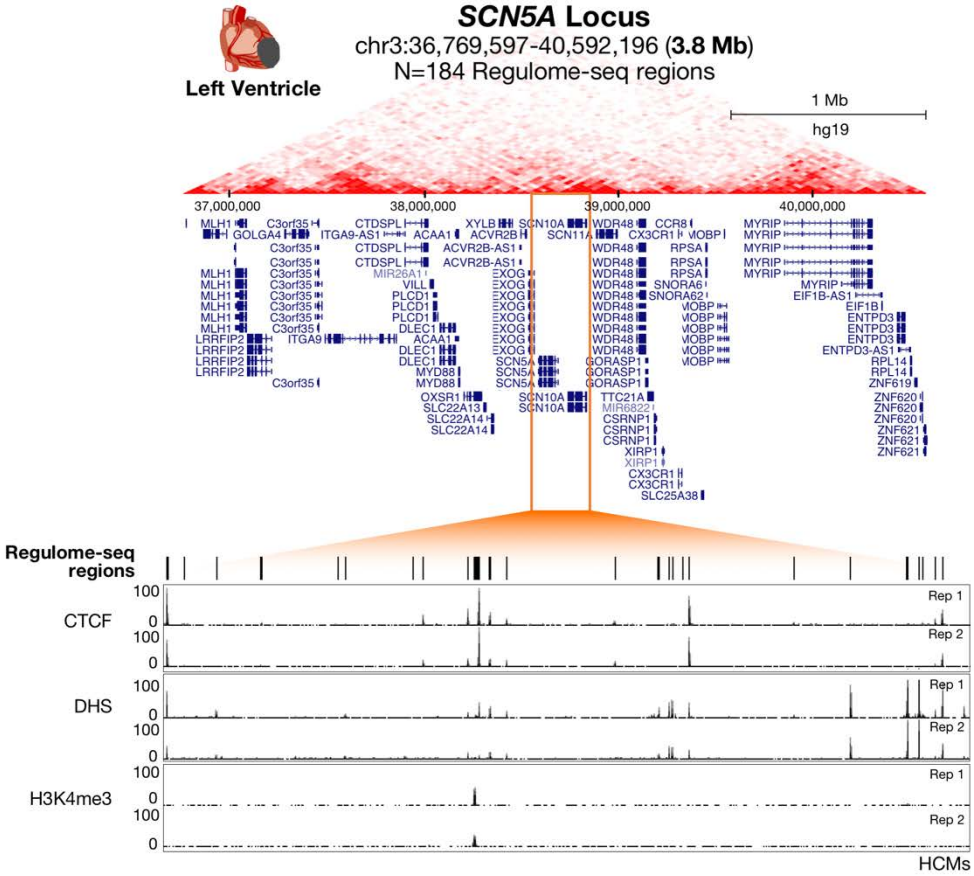


Figure 70. Example of the identification of Regulome-seq regions within the *SCN5A* locus. Top: Normalized Hi-C interaction frequencies at a 40 kb resolution in left ventricle[245]. The UCSC genome browser track of all genes is also shown. **Bottom:** Zoom in of *SCN5A* and *SCN10A* genes showing several Regulome-seq regions selected based on CTCF ChIP-seq, DHS-seq and H3K4me3 ChIP-seq signals from HCMs.

This analysis identified n=1,293 putative *cis*-regulatory regions for the six BrS-associated genes (Regulome-seq regions): 184 regions were found within the *SCN5A* locus, 264 within the

SCN2B locus, 242 within the *SCN3B* locus, 212 within the *CACNA1C* locus, 231 within the *CACNA2D1* locus and 160 within the *CACNB2* locus (**Figure 69** and **Annex 3**).

Although the selection of Regulome-seq regions was based on three different regulatory features (DHS, H3K4me3 and CTCF), not all the regions selected showed the three features together (**Figure 71** and **Annex 3**). In particular, 796 (61.5%) Regulome-seq regions were found to be enriched only in one of the features: 603 (46.6%) DHS, 184 (14.2%) CTCF and 9 (0.7%) H3K4me3; 436 (33.7%) Regulome-seq regions were found to be enriched in two features: 316 (24.4%) DHS-CTCF, and 120 (9.3%) DHS- H3K4me3; and 61 (4.72%) regions were found to be enriched in DHS-CTCF-H3K4me3.

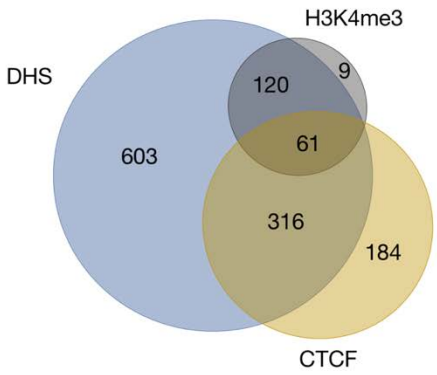


Figure 71. Distribution of regulatory features in the Regulome-seq regions. Venn diagram representing the number of Regulome-seq regions enriched in DHS, CTCF and H3K4me3.

In total, the 1,293 Regulome-seq regions account for 1.13 Mb of the whole human genome. Approximately 77.26% of the Regulome-seq regions are smaller than 1 kb in length, although they cover a broad spectrum of lengths, ranging from 150 to 10,100 bp. These differences are due to two different elements. First, the regulatory feature considered influences the final length of the region selected (**Figure 72**). Specific TF binding sites are small DNA regions compared to histone marks that tend to be broad. For this reason, Regulome-seq regions that are only enriched in CTCF correspond to the smallest regions (median length of 150 bp), followed by regions only enriched in DHS (median length of 755 bp), and regions only enriched in H3K4me3 (median length of 897 bp). Second, being enriched by more than one regulatory feature also influences in the final length of the Regulome-seq regions (**Figure 72**). That is because in some of the cases in which more than one regulatory feature is present, regulatory features may be partially overlapping. In these cases, we therefore extended the selected region to include all regulatory features. Indeed, those Regulome-seq regions that are enriched in all three features (DHS, CTCF and H3K4me3) are the longest regions, with a median length of 2,093 bp.

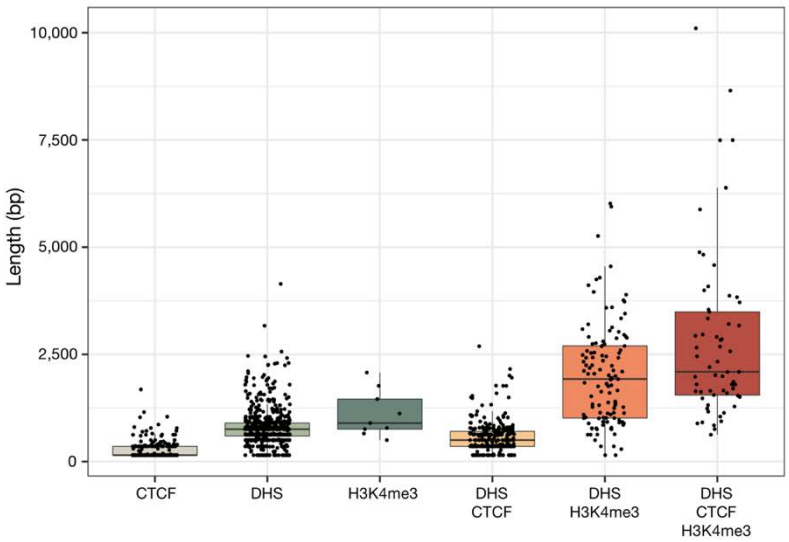


Figure 72. Length of Regulome-seq regions. Boxplots showing the influence of the type of regulatory feature and their combinations (y-axis) in the final length of the Regulome-seq regions (y-axis).

To further examine the regulatory function of the selected regions, we used HOMER to annotate the 1,293 Regulome-seq regions and obtained the following distribution: 55.6% were found in intronic regions, 31.99% in intergenic regions, 4.9% in promoter-transcription start sites (TSS), 2.86% in exons, 1.2% in transcription termination sites (TTS), 1.47% in 3' Untranslated Regions (UTR) and 1.31% in 5' UTR (**Figure 73**).

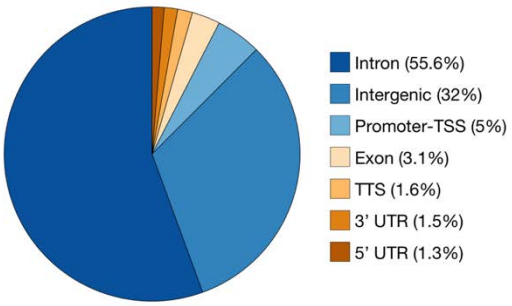


Figure 73. HOMER annotation of Regulome-seq regions. Pie chart showing the proportion of Regulome-seq regions spanning each annotated genomic feature.

1.3. Validation of the Regulome-seq design

To test the validity of our Regulome-seq approach in identifying disease-associated regulatory regions, we obtained a list of all GWAS SNPs associated to cardiac conduction defects from the GWAS Catalog (Materials section 1.6.2). When we merged the GWAS information with the list of our Regulome-seq regions, we observed that the rs6801957 polymorphism, which has been previously associated to QRS duration and ventricular conduction[259], is embedded in one of the Regulome-seq regions from the *SCN5A* locus (**Figure 74**). This polymorphism is found in a binding site for multiple TFs in different cell types and, more importantly, it is described to disrupt Tbx5 binding in cardiomyocytes, resulting in aberrant *SCN5A* gene expression[44,225].

These observations suggest that our Regulome-seq regions are indeed putative *cis*-regulatory regions related to electrical diseases of the heart such as BrS.

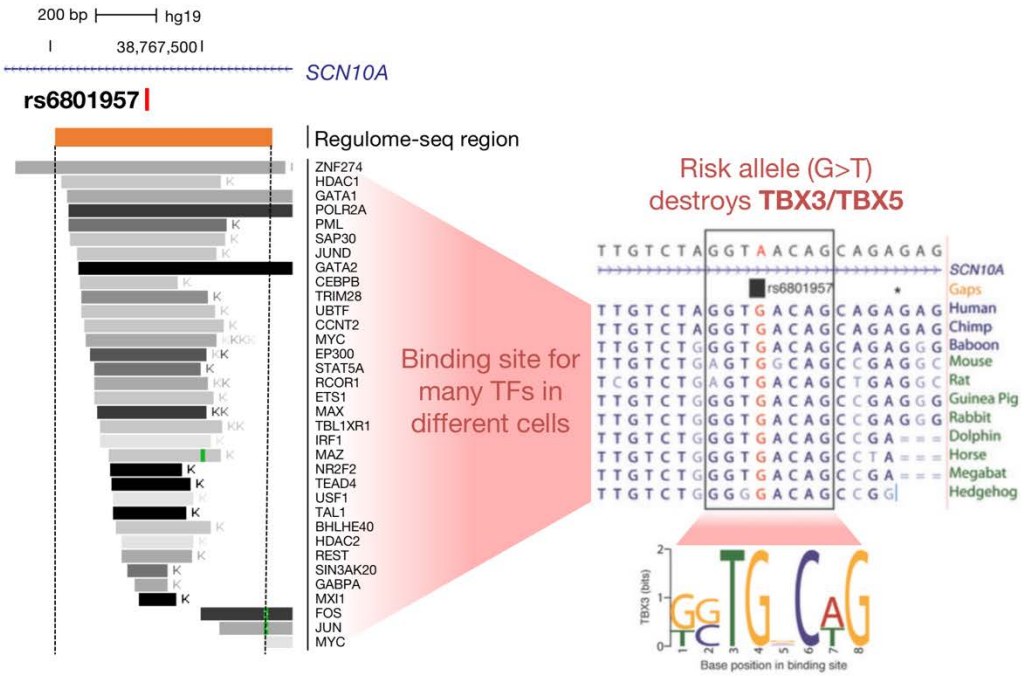


Figure 74. Validation of the Regulome-seq approach. Left: sequence conservation between species around rs6801957 as well as the nucleotide affected in the TBX5 binding motif (red). **Right:** The genomic localization of rs6801957 and the position in the Regulome-seq region is shown. UCSC browser tracks of many TF binding sites in different cells are also represented. Figure adapted from van den Boogaard *et al*.[44].

2. Sequencing of Regulome-seq regions in 89 BrS individuals

Once defined the list of 1,293 Regulome-seq regions, we designed a capturing approach to prepare DNA libraries containing only these regions (Methods section 2.2.1). DNA libraries were prepared from genomic DNA of the selected 89 BrS individuals and a Coriell sample (Materials **Tables 5** and **6**, respectively) using NRC from Illumina®, as described in Methods section 2.2.2. For the Coriell sample, we prepared three different DNA libraries that were treated as three different samples. The resulting 92 DNA libraries that were sequenced in a massively-parallel fashion as detailed in Methods section 2.2.3.

2.1. Design of Nextera Rapid Capturing probes

Using the DesignStudio™ tool (Illumina®), we designed a total of 5,546 NRC capturing probes to selectively capture our Regulome-seq regions in 89 BrS individuals (**Figure 75** and Methods **Figure 53**). We designed the probes to be non-overlapping, with a spacing between adjacent probes of 150 bp. The parameters on which DesignStudio™ is based to design the probes are patented and confidential. However, we know that, to obtain the most specific probes, it considers several parameters in addition to the spacing between probes. Therefore, when possible, DesignStudio™ will design probes separated by 150 bp but, in some cases, the most specific probe can be farther than 150 bp. In this case, the spacing criteria will not be accomplished and the added spacing will appear in the final report as a gap. Our design resulted in 346 gaps that represented 0.8% of the total Regulome-seq length, indicating that 99% of the Regulome-seq regions were covered. Moreover, in any case, the length of the gaps was greater than the spacing between adjacent probes, which guaranteed that all fragments would be captured by at least one probe.

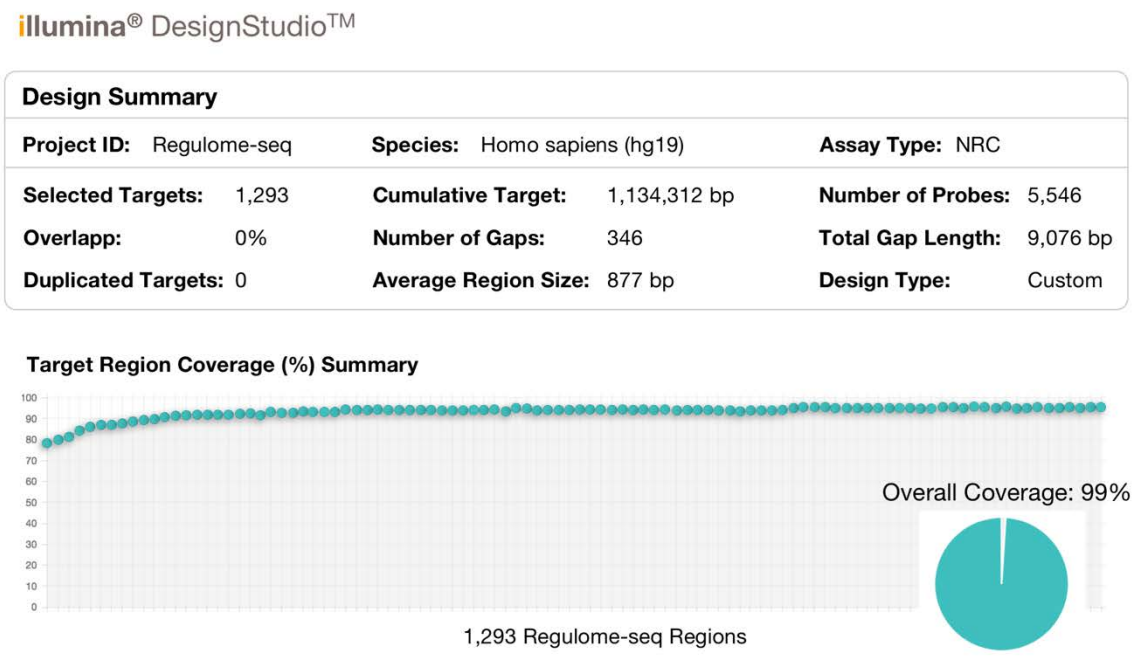


Figure 75. DesignStudio™ detailed report. Top: Summary table of the Regulome-seq regions, NRC probes and gaps. **Bottom:** Graph showing the percentage of coverage (y-axis) for each Regulome-seq region (x-axis). The overall coverage is also shown.

The designed 5,546 NRC capturing probes were synthesized by Illumina®, and were used to prepare DNA libraries from the 89 BrS individuals and the Coriell sample as described in Methods section 2.2.2.

2.2. Sequencing performance

The resulting DNA libraries from the 89 BrS individuals and the Coriell sample were sequenced on a HiSeq2500 following a paired-end protocol of 100 cycles. The paired-end sequencing resulted in 2 Fastq files per sample corresponding to the forward and reverse DNA strands (total of 184 Fastqs).

We analyzed the quality of our Fastqs using the FastQC package developed by the Bioinformatics group of the Babraham Institute (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). We considered the following parameters: (i) base quality, (ii) sequence content, (iii) GC content distribution and, (iv) duplication rate.

2.2.1. Base quality

The base quality module shows an overview of the base quality scores achieved at each position of the sequencing read. If the main goal of the sequencing is the identification of genetic variants, it is crucial to reach high quality base scores and avoid sequence ambiguity.

Overall, our reads presented a high quality base call even though a small drop could be observed at both ends of the reads (**Figure 76**). The decrease in quality at the 5' end of our reads is the result of technical biases derived from DNA fragmentation using transposases during library preparation. In contrast, the decrease in quality at the 3' end of our reads is common for most sequencing platforms and it is the result of the decrease in the sequencing efficiency as the sequencing cycles are succeeding. This decrease in base quality at both ends of our reads produced a warning error in this module, which was further corrected before variant discovery by a process called trimming.

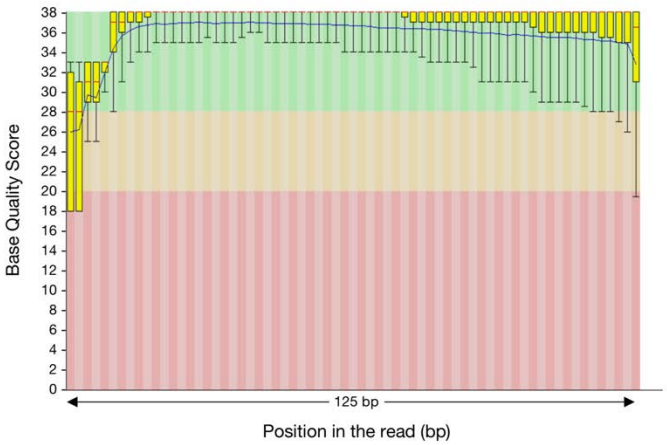


Figure 76. Base quality results for a particular BrS sample. Box plot showing the quality scores (y-axis) for each position (x-axis). The central red line in each box signals the median value, the yellow box represents the inter-quartile range (25-75%), the upper and lower whiskers represent the 10% and 90% points and the blue line represents the mean quality. The background of the graph divides the y-axis into very good quality calls (green), calls of reasonable quality (orange), and calls of poor quality (red).

2.2.2. Sequence content

The sequence content module measures the proportion of each of the four DNA bases (A, T, C and G) found at each position of the sequencing read. In this module, it is expected to obtain reads with a balanced proportion of bases. In our case, we obtained a well balanced proportion of A-T and G-C along our reads with the exception of the 5' end of the read (**Figure 77**). This small fluctuation also results from technical biases caused by DNA fragmentation using

transposases and also produced a warning error in this module, which was further corrected by trimming.

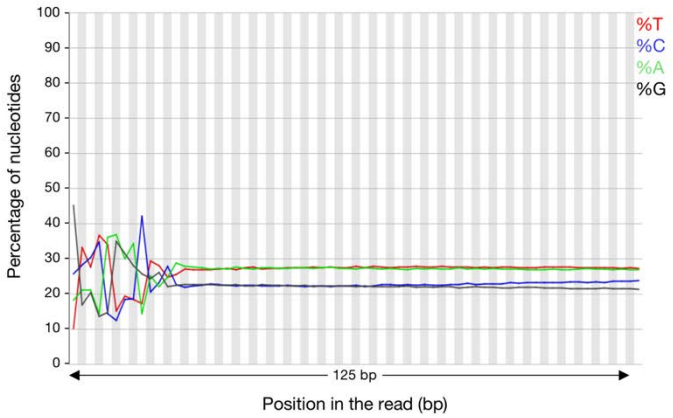


Figure 77. Sequence content results for a particular BrS sample. Percentage of each of the four DNA bases (A, T, C and G) that have been called for each base position in a Fastq file.

2.2.3. GC content distribution

The GC content distribution module builds a theoretical distribution model of the GC content of the human genome, and then compares the GC content of the sequencing reads with the theoretical distribution model. In this module, it is expected to obtain sequencing reads with a roughly normal distribution of the GC content, where the central peak corresponds to the overall GC content of the underlying genome. Our sequencing reads presented the expected normal distribution of the GC content, with a central peak around 45% (**Figure 78**).

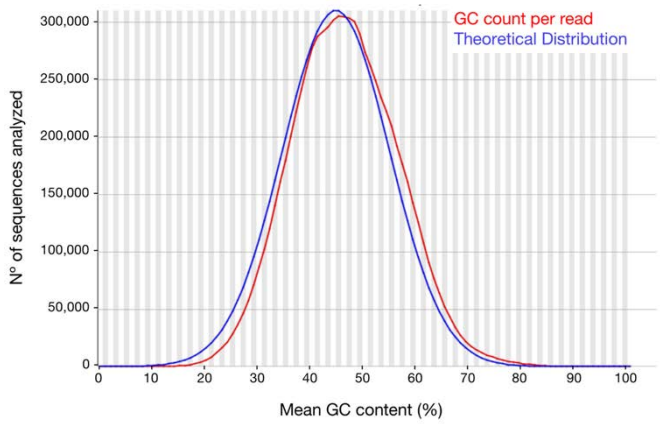


Figure 78. GC distribution results for a particular BrS sample. GC distribution (y-axis) over all the sequences analyzed (x-axis). The red line represents the GC count per read from this particular sample, while the blue line represents the theoretical GC content distribution of the genome, calculated from the observed data.

2.2.4. Duplication rate

Finally, the duplicated sequences module measures the different degrees of duplication for all sequencing reads. It also estimates the proportion of sequencing reads that will remain at each duplication level after removing all sequencing duplicates (process named deduplication). In our case, we obtained that 95% of our sequencing reads are only duplicated once, and that this percentage drops to values near 0% as we screen for sequences duplicated more than once (**Figure 79**). Therefore, when deduplicating our DNA library, almost 100% of the sequences are still considered for subsequent analysis.

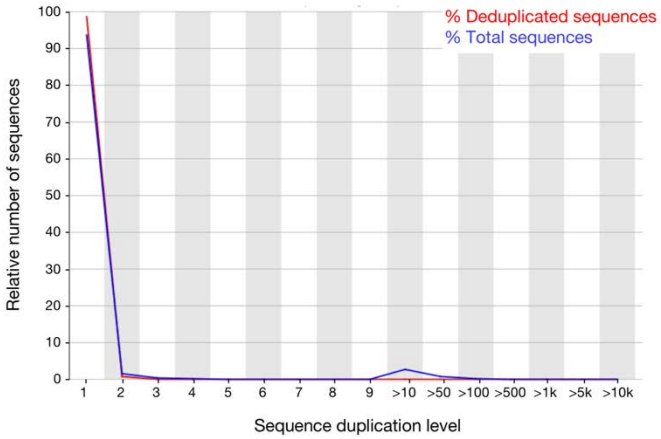


Figure 79. Sequence duplicate results for a particular BrS sample. Relative number of sequences with different degrees of duplication. The blue line represents the proportion of the total sequences present in the DNA library that fall in each duplication bin. The red line represents the proportion of sequences that will remain at each duplication bin after deduplication of the DNA library.

2.3. Capturing performance

In addition to the sequencing quality analysis, we also examined the performance of our capturing approach. Overall, we obtained an average of 5.9 million reads per sample, from which 3.8 million ± 948,849 corresponded to the Regulome-seq regions (**Table 22**). This result shows that 64.79% of total reads obtained from sequencing derive from our regions of interest (reads of interest; **ROI**), while the remaining 35% of reads derive from off-target regions.

Off-target events are typical in capturing approaches, and occur during DNA library preparation because, although the probes are designed to efficiently hybridize to the target regions, sometimes they can partially hybridize with other genomic regions similar in sequence than the targets. When sequencing DNA libraries prepared from small targeted Illumina sequencing panels, the expected percentage of off-target reads ranges from 60-40%. The 35% off-target reads obtained in our sequencing is even lower than expected, suggesting that our capturing design was efficient.

Once removed the off-target reads, we measured the **density of coverage**. This is another parameter that we used to measure the capturing performance, as it is indicative of the number of times a given base is sequenced. On average, for all samples, each base of our Regulome-seq regions showed a coverage of **384X ± 149** (**Table 22**).

The coverage is highly influenced by the **GC content** of the region to be captured, given that GC paired nucleotides are bound by 3 hydrogen bonds, which make more difficult the DNA denaturation and affect the subsequent hybridization with the probes. This phenomenon is observed as a drop in the average coverage for those Regulome-seq regions displaying a GC content ≥ 60% (**Figure 80**). Nevertheless, only 112 of the 1,293 Regulome-seq regions were enclosed in the category of GC content ≥60% and yet, they presented a high coverage (average of **240X ± 92**).

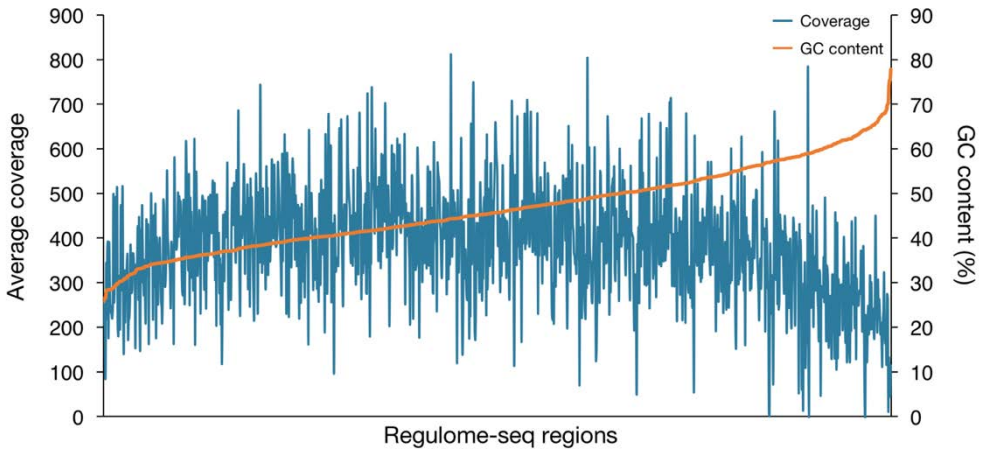


Figure 80. Coverage versus GC content. Graph showing the average coverage obtained for each Regulome-seq region (blue) in all BrS and NA12249 samples. Regulome-seq regions are displayed from lower to higher GC content (orange).

Finally, we complemented the information provided by the percentage of enrichment and the density of coverage with the **call rate**, which represents the fraction of bases from the regions of interest that reached a specific coverage. The call rate is tightly related to the robustness of the genetic variants that will be identified from sequencing data because, the more times a given position has been sequenced, the more confidence we have that a variant identified is real. For this reason, the call rate is very important for clinical diagnosis, and all those genetic variants with a call rate <30X are recommended to be validated by Sanger sequencing to be accepted as true genetic variants.

We measured the percentage of the total sequenced bases that would be recovered at different call rates. As expected, we observed that the number of bases recovered at distinct call rates decreases at stricter thresholds (**Figure 81**). Importantly, 96.01% of the Regulome-seq bases sequenced were covered ≥30X (**Figure 81** and **Table 22**), which is relatively close to the 99% obtained for validated gene panels used for diagnostic of several cardiovascular diseases in our laboratory.

Table 22. Sequencing statistics of BrS and NA12249 samples.

	Mean ± SD	1st Q	Median	3rd Q
Raw Reads	5,900,974 ± 2,148,873	4,612,000	4,985,000	5,816,000
ROI Reads	3,823,490 (64.79%) ± 948,849	3,089,000	3,544,000	4,061,000
Density of coverage	384X ± 149	301X	383X	460X
Call rate at ≥30X	96.01 ± 0.85	95.46	95.94	96.62

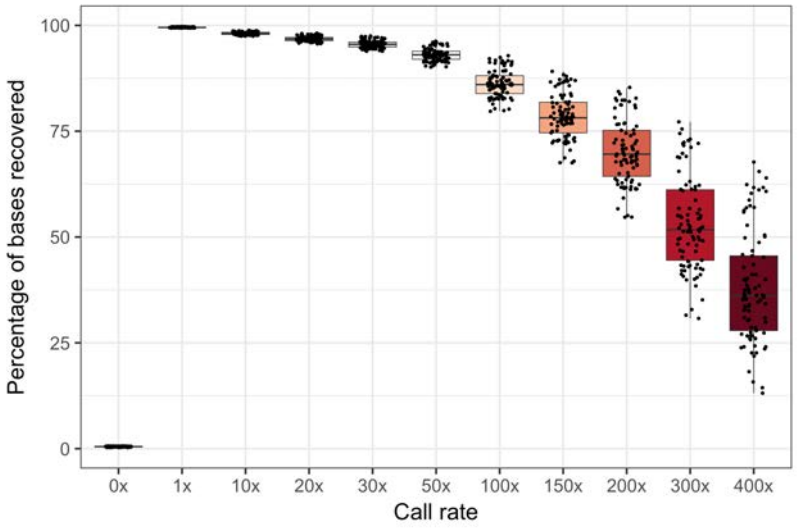


Figure 81. Percentage of bases recovered at different call rate thresholds. Boxplots showing the percentage of Regulome-seq bases (y-axis) that would be recovered at different call rate thresholds (x-axis).

3. Variant discovery

After validating the quality of our sequencing data, we proceeded with the identification of genetic variants using GATK (Methods section 2.3). We applied our variant discovery pipeline to the 89 BrS individuals and the Coriell sample together and obtained a list of 7,913 genetic variants that we refer to as GATK output.

We used the variants from the Coriell sample present in the GATK output to measure the quality of the variants identified, but also to determine several filters in order to obtain a final high confidence subset of variants (Methods section 2.4).

3.1. BrS variant call quality analysis

To measure the performance of our variant discovery pipeline, we compared the SNVs identified for the three replicas of the Coriell (NA12249 A-C; regulome dataset) with the SNVs present in the 1000 Genomes database for the same Coriell (public dataset).

In total, we identified 96% (1,337) of the 1,391 SNVs present in the 1000 Genomes dataset, resulting in an **average sensitivity of 0.961** and an **average PPV of 0.954** (**Table 23**). These values are nearly identical to those obtained by Hwang *et al*.,[260] when comparing different variant calling pipelines for targeted sequencing (sensitivity = 0.99 and PPV = 0.96), which reflect a good performance of our variant discovery pipeline.

Table 23. Number of TP, FP, FN identified in each of the three Coriell replicas with our variant discovery pipeline. Sensitivity and PPV are calculated for each sample.

	TP	FP	FN	Sensitivity	PPV	Sample
GATK output	1331	59	61	0.956	0.958	NA12249A
	1341	60	51	0.963	0.957	NA12249B
	1339	76	53	0.962	0.946	NA12249C

TP (True Positives), FP (False Positives), FN (False Negatives), PPV (Positive Predictive Value).

3.2. Curation of the BrS variant call set

Although the performance of our variant discovery pipeline was robust, we optimized it to obtain the lowest possible number of FP and the highest possible number of TP possible. Our goal was to increase the precision of our variant discovery pipeline without significantly compromising the sensitivity.

We applied several consecutive filters to the GATK output, which contained the list of genetic variants found in the 89 BrS individuals as well as the 3 replicas of the Coriell sample (Methods **Figure 61**). After each filter, the SNVs from the Coriell sample were used to measure the sensitivity and PPV of each filter as explained in Methods section 2.4.2. The results obtained are summarized in **Table 24.**

In the first filter (**Filter 1**), we removed 650 genetic variants labelled as low quality after GATK variant recalibration, applied during variant discovery to increase the confidence of the variants. With this filter, the average sensitivity was 0.930 and the average PPV was 0.986.

The second filter (**Filter 2**) was applied to the results from Filter 1 and removed 327 genetic variants from which the genotype was not resolved for some samples. With this filter, the average sensitivity was 0.930 and the average PPV was 0.988.

The third filter (**Filter 3**) was applied to the results from Filter 2 and removed 413 genetic variants that did not reach a coverage of ≥10X in all samples. With this filter, the average sensitivity was 0.912 and the average PPV was 0.987.

The fourth filter (**Filter 4**) was applied to the results from Filter 2 and removed 992 genetic variants that did not reach a coverage of ≥30X in all samples. With this filter, the average sensitivity was 0.862 and the average PPV was 0.990.

Finally, the fifth filter (**Filter 5**) was applied to the results from Filter 2 and removed 5,666 genetic variants that did not reach a coverage of ≥100X in all samples. With this filter, the average sensitivity was 0.231 and the average PPV was 0.99.

Table 24. Filters applied to BrS variant call set. Initial number of variants, filtered variants and remaining variants in the call set of 89 BrS individuals after applying each filter. The number of TP, FP and FN as well as the average sensitivity and PPV measured for the 3 Coriell replicas are also shown. The final filter selected is highlighted in bold.

Filter	Initial variants	Filtered variants	Remaining variants	TP	FP	FN	Sensitivity	PPV
Filter 1 (Low quality)	7,913	650	7,263	1,294	19	98	0.930	0.986
Filter 2 (non-resolved genotypes)	7,263	327	6,936	1,294	16	98	0.930	0.988
Filter 3 (≥10X in all samples)	6,936	413	6,523	1,269	16	123	0.911	0.988
Filter 4 (≥30X in all samples)	**6,936**	**992**	**5,944**	**1,200**	**12**	**192**	**0.862**	**0.990**
Filter 5 (≥100X in all samples)	6,936	5,666	1,270	380	3	1,261	0.231	0.993

At a global level, we observed that the total number of variants removed increases considerably as more filters are being applied (**Table 24**). The critical point resides in the tradeoff between the number of FP and the number of TP that are being removed. In our case, every filter accomplished the objective to decrease the number of FP but this achievement was also related to a reduction in the number of TP recovered. Therefore, and not surprisingly, after each filter, the sensitivity was reduced while the PPV was increased.

In the light of these results, the filter combination chosen to curate our variant call set was **Filter 1+2+4**, given that it was showing the highest PPV (0.990) without excessively compromising the sensitivity (0.862).

4. Characterization of Regulome-seq variants

After applying the corresponding filters to the variants from the 89 BrS and the 3 Coriell replicas, we obtained a total of 5,944 genetic variants. From these, 5,902 variants are exclusive to BrS individuals and 42 variants are exclusive to the Coriell sample.

It is important to note that 339 of these variants were found at multi-allelic positions. A multi-allelic position is a given genomic position that displays different variants in distinct individuals. As it has been described in the introduction, considering the large amount of differences found after a pairwise comparison of two individual genomes, it is expected to observe different variants in the same position when taking into account more than two genomes. However, we realized that the vast majority of the multi-allelic variants identified corresponded to ambiguous indels found at repetitive regions. When sequencing data is obtained from short read sequencing (as it is in this case), the exact genotype of these type of indels is difficult to solve and, in some cases, variant callers might be assigning different genotypes to the same indel.

Therefore, we removed the 42 variants exclusive to the Coriell and the 339 variants found in multi-allelic positions. The resulting list contained a total of **5,508 genetic variants** identified at the Regulome-seq regions of 89 BrS individuals.

The 5,508 genetic variants are distributed as follows: 90.22% (4,969) SNVs and 9.78% (539) indels. From the 539 indels, 235 are insertions and 304 deletions (**Table 25**). These results, in which we observe that the vast majority of the variants identified are SNVs, are in agreement with previous reports showing that genetic variants affecting a single nucleotide are the most frequent variant found in the human genome[107]. Similarly, although we identified insertions spanning 82 bp and deletions spanning 42 bp, the most frequent indel identified are small indels affecting only 1 bp (62.13% for 1 bp insertions and 42.43% for 1 bp deletions; **Figure 82**), which is also the most common indel length in the human genome[261].

Table 25. Summary of the types of variants identified in the 89 BrS individuals. The average number of variants per individual and the median length of indels are also represented.

Type	Count	Average/individual	Median length
SNVs	4,969 (90.22%)	1,148 ± 54	-
Insertions	235 (4.27%)	63 ± 5.20	1 bp
Deletions	304 (5.52%)	72 ± 5.88	2 bp

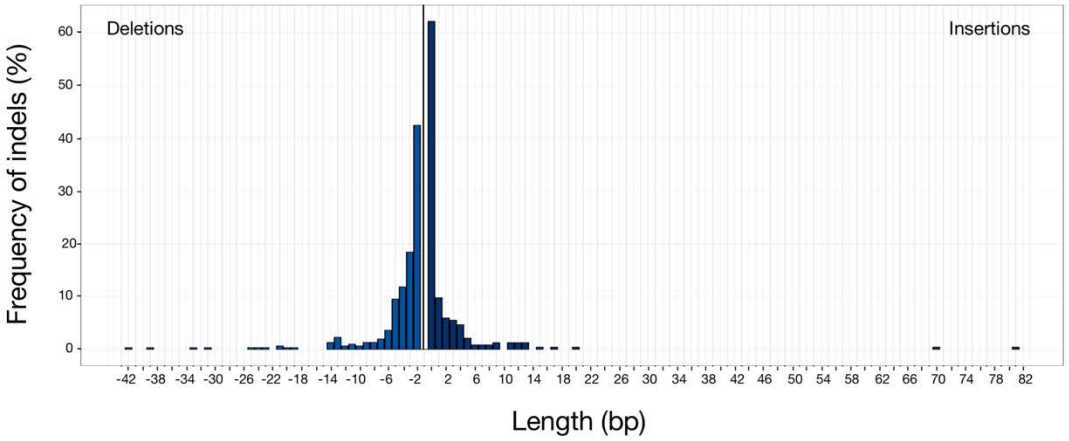


Figure 82. Patterns of indels in the 89 BrS individuals sequenced. Frequency distribution of indels (y-axis) by length (x-axis).

Independently of the type of genetic variant, we observed that SNVs and indels are homogeneously distributed among the 89 BrS individuals, with an average of 1,283 ± 59.3 variants per individual: 1,148 ± 54 SNVs, 63 ± 5.20 insertions, and 72 ± 5.88 deletions (**Table 25** and **Figure 83**).

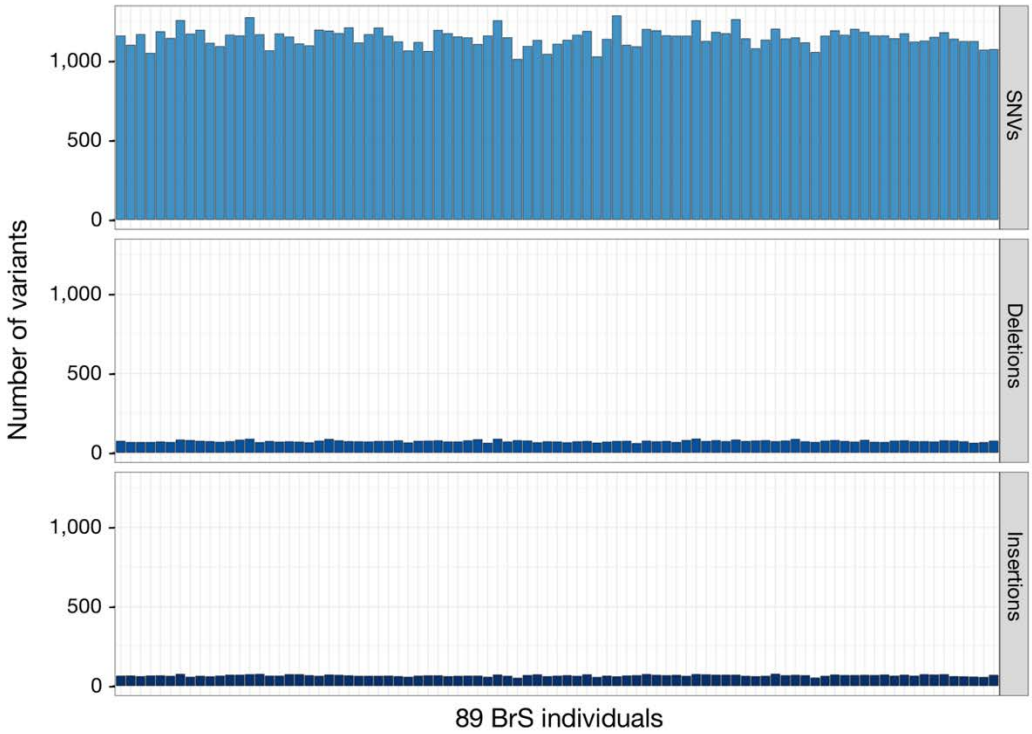


Figure 83. Total number of variants per individual. Number of SNVs, deletions and insertions variants (y-axis) identified in each BrS patient (x-axis).

Although each BrS individual presents a similar number of variants, we observed that they do not share the same variants. In fact, only 2.89% of total variants are shared by all 89 BrS individuals (**Figure 84**). The remaining variants can be classified in three different groups depending on the proportion of variants shared. The first group, comprises private variants of each BrS individual and includes 33.29% of total variants; the second group, comprises variants shared by 2 to 10 individuals and contains 26.76% of total variants; and, the third group, comprises variants shared by 10 to 88 individuals and contains 37.05% of total variants.

When analyzing which type of variant is more prevalent at each category, we observed that SNVs are more frequent within the first group of private variants than indels (33.84% vs 28.93%). In contrast, indels become more prevalent than SNVs as the variants are more shared between individuals (**Figure 84**).

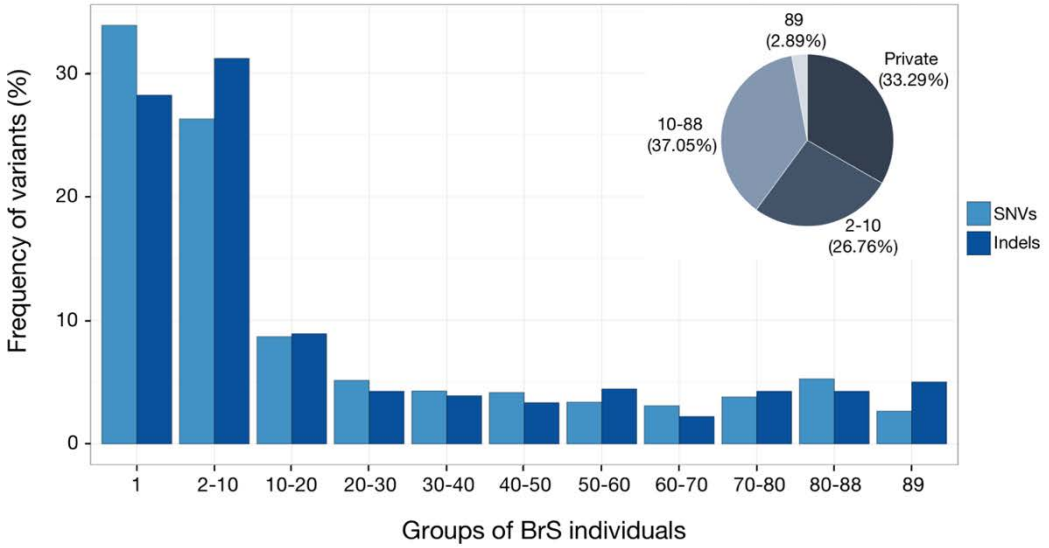


Figure 84. Distribution of shared variants in the 89 BrS individuals. Frequency of variants in percentage (y-axis) found at each sharing group (x-axis).

4.1. Annotation of Regulome-seq variants

Our analysis shows that, from the total 5,508 genetic variants, 2.89% (159 variants) are shared among all 89 BrS patients studied. However, it is unlikely that the BrS phenotype could be explained by this 2.89%. Most probably, these highly shared variants will be related to a common feature of that particular group of individuals such as the ancestral origin, rather than the disease phenotype itself. For this reason, and to facilitate the identification of possible variants that may contribute to the BrS phenotype, we removed these 159 variants from the subsequent analysis, resulting in a total number of **5,349 variants**.

Next, we sought to identify which of these final variants reported are described for the first time (novel variants). For this purpose, we annotated the 5,349 variants using dbSNP150, gnomAD (release 2.0.2) and 1000 Genomes databases (Materials section 1.6.3). We found that 366 of the variants (304 SNVs and 62 indels: 23 insertions and 39 deletions) were not present in any of the three datasets and hence, could be considered novel variants. Notably, we also found that approximately 93.44% of these novel variants correspond to private variants. This observation is in agreement with the recent discovery that the number of newly reported variants keeps growing with every new genome sequenced[262].

5. Identification of functionally relevant variants to BrS

After the characterization of genetic variants present in the Regulome-seq regions of 89 BrS individuals, we aimed to identify which of these variants may be functionally relevant to BrS (BrS-candidate variants). To address this issue, we used two different approaches: (i) an approach focused on the identification of genetic variants that might be affecting the binding of TFs and, (ii) an approach focused on the identification of genetic variants significantly enriched in BrS individuals compared to a cohort of healthy-aging individuals (Wellderly).

The final objective was to integrate all the information obtained to identify BrS-candidate variants, which will be functionally validated in future experiments.

5.1. BrS variants affecting transcription factor binding

In humans, gene expression programs that establish and maintain specific cell states are controlled by thousands of TFs that bind to regulatory elements to activate or repress transcription of specific genes. In the last few years, it has been observed that the presence of genetic variants affecting the binding of TFs to any type of regulatory element might lead to a misregulation of these cell-specific gene expression programs, resulting in a broad range of diseases[26,147–149]. Similarly, we hypothesized that genetic variants affecting the binding of TF that regulate cardiac ion channel expression may be related to BrS. Therefore, we focused on identifying those variants overlapping binding sites of cardiac TFs (GATA4, GATA6 and NKX2.5) involved in the regulation of cardiac ion channels[172,263]. We also included the possibility that genetic variants affecting the binding of CTCF at boundary elements could disrupt the insulation of BrS-associated genes and result in aberrant expression of these genes[28,75].

When this strategy was designed, there was not any available information for TF binding in human cardiomyocytes other than CTCF. Hence, to identify genetic variants affecting the binding of cardiac TFs, we set ourselves the goal of profiling the binding patterns of GATA4, GATA6 and NKX2.5 using ChIP-seq in iPS-derived cardiomyocytes (Materials section 1.2.2). To identify genetic variants affecting CTCF binding, we used the information already available from HCMs (Methods **Table 19**).

5.1.1. ChIP-seq in iPS-derived cardiomyocytes

We performed ChIP-seq experiments in iPS-derived cardiomyocytes using antibodies against GATA4, GATA6 and NKX2.5 as described in Methods section 2.6.

After the analysis of ChIP-seq data, we obtained a list of 315 peaks for GATA4, 1,971 peaks for GATA6 and 10,801 peaks for NKX2.5 (**Table 26**). We already expected that the number of peaks obtained would be lower for cardiac TF than for TFs with broader functions such as CTCF (>100,000 peaks in the available dataset). However, the number of peaks obtained for GATA4 is not consistent with the importance of this TF in heart physiology.

We proceeded with the analysis of our ChIP-seq data by interrogating the regions of the genome where the peaks were found. The genomic localization is indicative of the reliability of the experiment. For example, all peaks found at satellite repeats are considered artifacts of the technique. In our case, 38.41% of GATA4 peaks, 2.38% of GATA6 peaks and 1.14% of NKX2.5 peaks were found in satellite repeats. The observation that almost 40% of GATA4 peaks were found in satellite repeats, together with the low number of peaks detected for this factor, suggested that our ChIP-seq for GATA4 did not work. Still, we removed all peaks found at satellite repeats for the three factors and continued the analysis with a final list of 194 peaks for GATA4, 1,924 peaks for GATA6 and 10,678 peaks for NKX2.5 (**Table 26**).

To investigate whether the accepted ChIP-seq peaks directly correlate with DNA binding sites for GATA4, GATA6 and NKX2.5, we performed a motif analysis using HOMER. This analysis showed that 70.62% of GATA4 peaks contain the GATA4 motif, 43.40% of GATA6 peaks contain the GATA6 motif and 26.69% of NKX2.5 peaks contain the NKX2.5 motif (**Table 26**). The presence of the motif confirms that the immunoprecipitated DNA is bound by the TF of interest, although we cannot conclude that the remaining peaks without motif are not bound by the TF analyzed. One possible explanation for these peaks without motif could be the binding of TFs through other TFs or co-factors. Indeed, GATA4, GATA6 and NKX2.5 are known to interact with each other and co-occupy several regulatory elements to modulate cardiac gene expression[168,264]. To examine the co-occupancy of these cardiac TFs in our cardiomyocyte cell model, we assessed whether the peaks identified overlap. We observed that, although the number of peaks identified was low, we still could detect co-occupancy between them: 56.70% of GATA4 peaks overlap with GATA6 and NKX2.5 peaks, 15.85% of GATA6 peaks overlap with GATA4 and NKX2.5 peaks, and 2.42% of NKX2.5 overlap with GATA4 and GATA6 peaks (**Figure 85**).

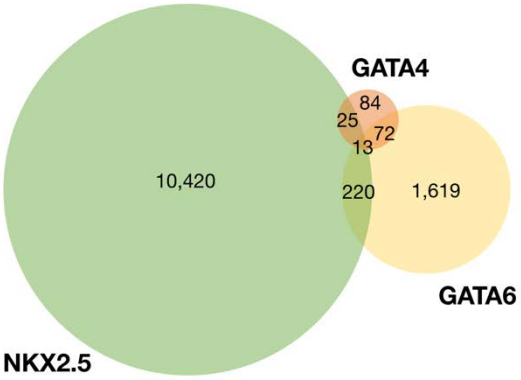


Figure 85. Co-occupancy of cardiac TFs. Venn diagram representing the number of overlapping GATA4, GATA6 and NKX2.5 peaks.

To identify genetic variants that might be affecting the binding of GATA4, GATA6 or NKX2.5, we intersected the ChIP-seq peaks with the Regulome-seq regions. We observed that only 2 GATA4 peaks, 14 GATA6 peaks and 66 NKX2.5 peaks were overlapping any Regulome-seq region (**Table 26**). In view of these results, we manually inspected each ChIP-seq peak on the UCSC genome browser. We observed that the quality of all the peaks was very low, and less than 1% of them showed the motif of the TF that was being analyzed. Therefore, we concluded that our ChIP-seq experiments did not have enough quality to be used for our purposes.

Table 26. ChIP-seq results for cardiac TFs in iPS-derived cardiomyocytes.

TF	Raw peaks	Accepted peaks	Peaks with motif	Regulome peaks
GATA4	315	194	137 (70.62%)	2
GATA6	1,971	1,924	835 (43.40%)	14
NKX2.5	10,801	10,678	2.850 (26.69%)	66

5.1.2. BrS variants overlapping CTCF binding sites

Since our ChIP-seq data for cardiac TFs could not be used, we took advantage of the already available information of CTCF binding from HCMs. As mentioned earlier, one of our hypothesis was that the aberrant expression of BrS-associated genes linked to BrS phenotype could be caused by a disruption of CTCF binding at boundary elements.

CTCF ChIP-seq data from HCMs was extracted from the ENCODE Project repository (Materials **Table 19**). In order to facilitate the direct association of genetic variants to CTCF binding effects, we only accepted those 56,264 binding sites from HCMs (from a total of ~120,000) where a CTCF motif was found after HOMER motif analysis. Additionally, we

redefined CTCF binding sites as 36 bp regions surrounding the CTCF motif as explained in Methods section 2.7.1. The reason to use 36 bp was merely because the algorithm used in the section below to predict the binding effects of CTCF-overlapping variants (DeepBind) required sequences of this length.

From the 5,349 Regulome-seq variants surveyed, we identified a total of 59 variants (52 SNVs and 7 indels) overlapping 54 different CTCF binding sites (**Annex 4**). These **59 CTCF-overlapping variants** are shared by different number of individuals: 42.37% (25) are private variants and 57.65% (34) are shared by 2 to 75 BrS individuals.

Next, we interrogated the number of variants affecting each position of the CTCF binding site. We observed that the 59 CTCF-overlapping variants are affecting almost all positions, although important nucleotides tend to show a reduced number of variants (**Figure 86**). These important nucleotides, also referred to as core nucleotides, correspond to those nucleotides that are more frequently found in the binding site of a given TF and hence, they are assumed to be crucial for the binding of the TF.

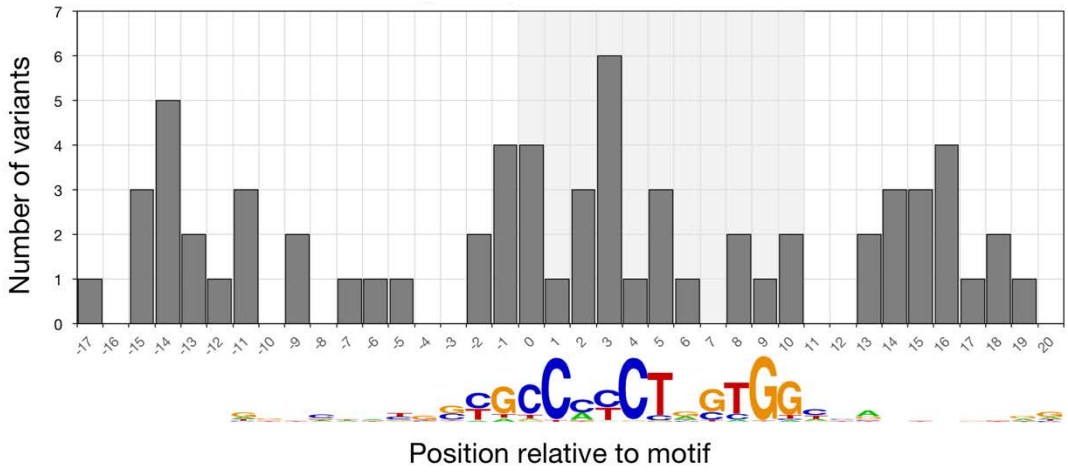


Figure 86. Relative position of the 59 CTCF-overlapping variants along the consensus CTCF motif. Number of variants identified (y-axis) at a given position of the CTCF binding site (x-axis). CTCF motif logo with important nucleotides represented with a larger size is also shown. The 11-bp central motif is highlighted in gray shading.

5.1.2.1. Binding effects of CTCF-overlapping variants: DeepBind predictions

TFs have different binding affinities for their target sequences, which constitutes a biological advantage because it allows them to recognize a broader spectrum of binding sites. This property, however, challenges the identification of variants affecting TF binding. To overcome this issue, new computational algorithms based on deep learning have emerged. Among these algorithms, **DeepBind**[159] was designed to: (i) learn the important features that influence TF

binding from millions of ChIP-seq sequences and, (ii) create a binding model to interrogate binding affinity of a particular TF to any other sequence. DeepBind scores (ranging from -5 to 30) obtained for two identical sequences except for a given variant, can be used to predict the possible effects of these variants in TF binding.

To comprehensively identify BrS variants with potential functional consequences for CTCF binding, we predicted the effect of the 59 CTCF-overlapping variants on CTCF binding using DeepBind (Methods section 2.7.2). For each variant analyzed, we obtained a binding prediction for the 36 bp motif-centered binding sites, containing either the reference or alternative alleles (118 predictions in total; **Table 27**). In order to facilitate the interpretation of DeepBind scores, we visualized the data using a variation map, which illustrates the effect of all possible nucleotide substitutions at each position of the 36 bp sequence on the CTCF binding score (**Figure 87**).

Table 27. DeepBind scores for all 59 CTCF-overlapping variants.

Chr	Pos	Ref	Alt	DB-Ref	DB-Alt	DB difference
chr3	37953719	T	A	8.3082	2.7873	-5.5209
chr3	38045036	T	A	7.4822	3.7850	-3.6972
chr3	38521504	A	G	15.2501	14.2140	-1.0361
chr3	38780342	G	A	7.4553	9.6975	2.2422
chr3	39540276	G	A	11.0922	9.6211	-1.4712
chr3	39854224	A	G	2.1961	0.5391	-1.6570
chr3	39953594	C	T	-0.3996	0.2113	0.6110
chr7	79677353	C	T	9.9340	9.7280	-0.2061
chr7	80288990	ATGT	A	3.2642	3.2642	0.0000
chr7	80549633	C	G	-2.3403	-3.9181	-1.5778
chr7	80625526	G	A	4.3940	2.6399	-1.7540
chr7	81076865	TC	T	2.6504	-4.5854	-7.2358
chr7	81076870	A	T	2.6504	-0.9966	-3.6470
chr7	81087389	T	C	3.3678	3.7228	0.3550
chr7	81130725	C	T	-0.0611	0.9565	1.0177
chr7	82110992	T	C	2.8331	2.0573	-0.7758
chr7	82702421	G	A	0.6306	0.2144	-0.4162
chr7	82900741	T	G	-1.0576	-1.3973	-0.3398
chr10	18883940	C	T	0.9719	1.3090	0.3371
chr10	19457714	C	G	-2.1596	-2.5036	-0.3440
chr10	21147418	G	A	9.4909	10.9514	1.4605
chr11	117122398	A	T	-2.1662	-2.2417	-0.0755
chr11	117924278	C	T	15.0846	14.8785	-0.2061
chr11	118042498	C	T	16.2888	16.6618	0.3729

Chr	Pos	Ref	Alt	DB-Ref	DB-Alt	DB difference
chr11	118549803	C	T	1.1932	-3.5073	-4.7005
chr11	118560953	G	GC	5.5580	-4.6901	-10.2481
chr11	118886117	C	T	1.9024	2.7677	0.8653
chr11	118889370	C	G	-4.0154	-5.2260	-1.2106
chr11	118889378	G	A	-4.0154	-4.2119	-0.1965
chr11	119287550	G	A	0.0317	-1.5674	-1.5991
chr11	119352239	G	A	12.6934	13.2055	0.5121
chr11	119404719	C	T	3.8218	2.2103	-1.6115
chr11	119600183	A	T	5.7462	5.1608	-0.5854
chr11	119612173	C	T	17.3878	15.3234	-2.0644
chr11	120053724	C	T	3.3906	-24.2525	-27.6431
chr11	120100704	T	G	6.0067	11.3597	5.3531
chr11	120173911	C	CTGAAG	16.8103	2.6049	-14.2054
chr11	120173914	C	CGGG	12.3167	-3.7030	-16.0198
chr11	120173917	C	CT	16.8103	-4.2313	-21.0416
chr11	122659691	T	C	0.9251	0.7121	-0.2131
chr11	123036887	G	C	-0.8743	-5.2260	-4.3517
chr11	123105986	G	T	5.6479	3.5463	-2.1016
chr11	123118102	CT	C	15.5834	6.6664	-8.9170
chr11	123132370	G	C	14.7207	13.7907	-0.9300
chr11	123425811	A	C	16.9775	14.9987	-1.9788
chr11	123447866	A	G	5.4157	5.5082	0.0926
chr11	123511861	A	AC	14.2739	14.2739	0.0000
chr11	123511862	C	T	14.2739	14.4386	0.1646
chr11	124539211	G	A	10.6385	8.5742	-2.0644
chr11	124648367	T	G	-0.0981	-3.9124	-3.8143
chr11	124984629	G	T	5.2290	2.1235	-3.1054
chr12	2469918	T	A	0.8303	-1.9177	-2.7481
chr12	2733860	C	T	13.4411	10.1164	-3.3247
chr12	3053956	G	A	6.9876	1.2724	-5.7153
chr12	3384802	G	A	26.9163	27.0687	0.1524
chr12	3451418	C	T	-0.4915	-3.9848	-3.4934
chr12	3764916	G	C	-5.0597	-5.1132	-0.0535
chr12	3767720	G	T	4.8477	0.1653	-4.6825
chr12	3913211	G	A	5.0002	5.1649	0.1646

Chr (Chromosome), Pos (Position), Ref (reference allele), Alt (Alternative allele), DB (DeepBind), DB-Ref (DeepBind prediction for reference allele), DB-Alt (DeepBind prediction for alternative allele). Green highlights variants predicted to increase binding both by DeepBind and luciferase activity, red highlights variants predicted to decrease binding both by DeepBind and luciferase activity, orange highlights variants predicted to have no effects on CTCF binding both by DeepBind and luciferase activity, and grey highlights a variant that could not be cloned for luciferase activity measurements (see next section for further details).

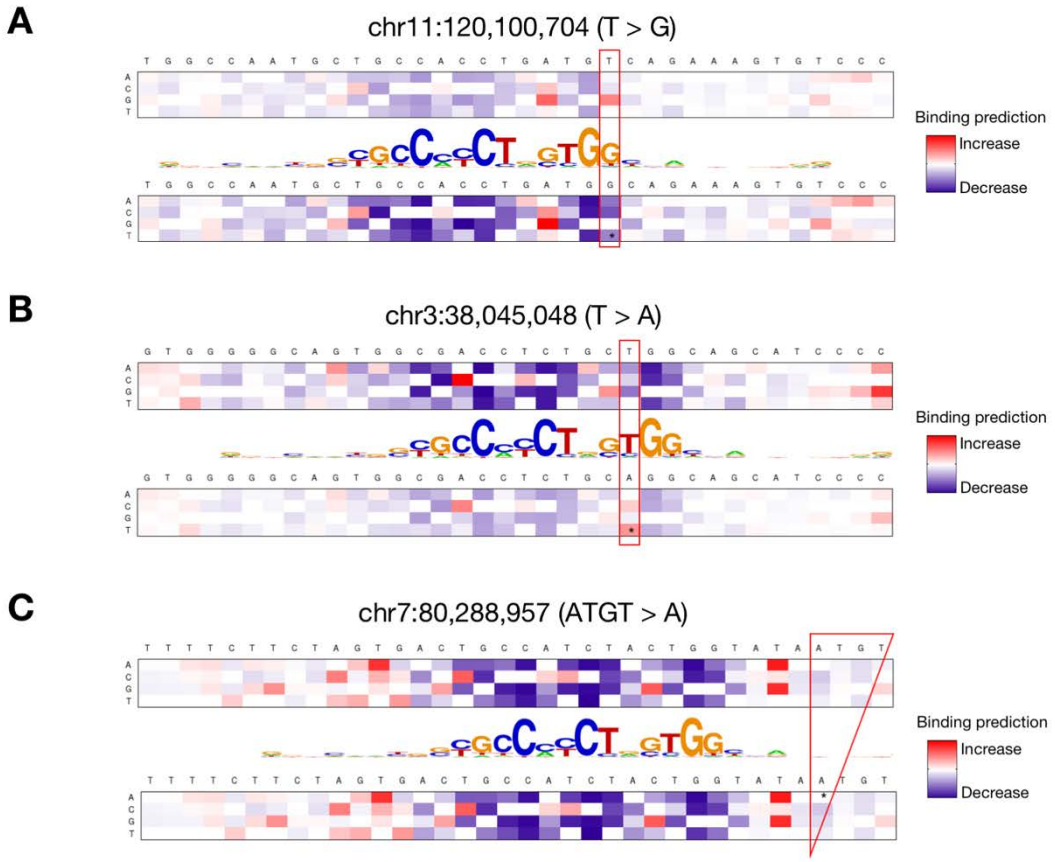


Figure 87. Example of CTCF binding predictions using DeepBind. Variation maps of three different variants analyzed by DeepBind. The top map corresponds to the reference sequence and the bottom map corresponds to the alternative sequence. The intensity of DeepBind scores is color-coded: red (increase on binding), blue (decrease on binding) or white (no effects on binding). **(A)** T > A substitution at chr3:38,945,036 predicted to decrease CTCF binding by disrupting the CTCF consensus motif. This diminished binding in the alternative sequence results in softer binding effects in front of other possible variants (less intense colors). **(B)** T > G substitution at chr11:120,100,704 predicted to increase CTCF binding by restoring the CTCF consensus motif. This increased binding in the alternative sequence results in larger binding effects in front of other possible variants (more intense colors). **(C)** ATGT > A deletion at chr7:80,288,957 predicted to not have any effect on CTCF binding as it is not affecting the CTCF motif. This neutral effect results in both sequences showing the same binding effects in front of other possible variants (equally intense colors). In this case the reference and alternative sequences are identical because the following nucleotides to the deletion are also TGT.

In summary, DeepBind analysis of the 59 CTCF-overlapping variants showed that **43/59** variants will **decrease**, **14/59** will **increase**, and **2/59** will have **no effect** on CTCF binding (**Table 27**). We also analyzed the difference in binding between reference and alternative sequences predicted by DeepBind based on the relative position of the variant from the CTCF motif. As it is shown in **Figure 88**, variants were classified as being far, closer or inside the CTCF motif. Variants falling inside the motif were further subdivided as non-core and core, depending on the nucleotide affected.

To determine whether the differences in binding between each category were significant, we used a pairwise t-test with Benjamini-Hochberg p-value correction for multiple comparisons as explained on Methods section 2.12.1. The differences between far and closer variants were found to be not significant. However, the differences in binding when the variants are inside the motif were found significant (**Figure 88**). Moreover, we observed that, although not significant, variants affecting core nucleotides show higher binding effects than variants affecting non-core nucleotides. Altogether, these results suggest that DeepBind predictions may recapitulate possible functional effects of CTCF-overlapping variants.

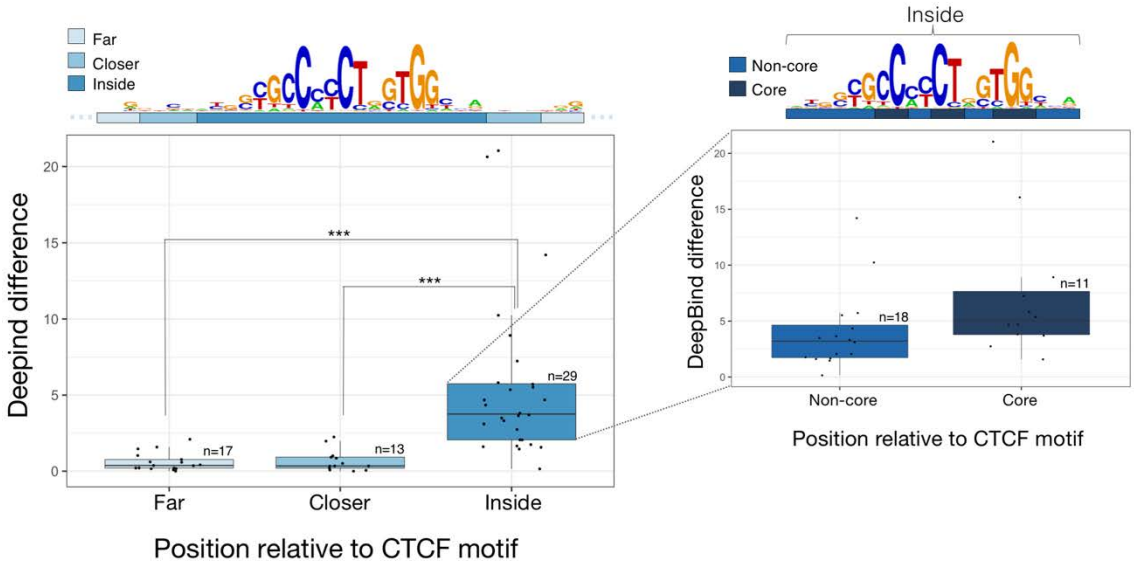


Figure 88. Positional binding effects predicted by DeepBind. Boxplots showing the absolute binding difference (y-axis) for each group of variants, classified according to their position relative to the CTCF motif (x-axis). For indels affecting more than one position, only the first position affected was considered for variant classification. ***p ≤0.001.

5.1.2.2. Binding effects of CTCF-overlapping variants: Luciferase assays

To experimentally measure the effects of the 59 CTCF-overlapping variants, we designed a **luciferase reporter-based assay** (Methods section 2.11). We cloned the 36 bp sequences, from which a DeepBind score was obtained, into a pGL4.23 vector containing the luciferase gene under the control of a minimal promoter. In total, we cloned 118 sequences (59 variants x 2 alleles; **Annex 2** and Materials **Figure 52**) as detailed in Methods section 2.8.2. Each of these constructs was co-transfected into H9c2 cells together with a vector expressing the human CTCF fused to the VP64 trans-activator–used to increase the potential of CTCF to act as an activator–. Forty-eight hours post-transfection, we quantified the luciferase activity of each variant to evaluate CTCF binding.

Previous to the aforementioned experiments, we tested whether this luciferase strategy was a valid method to measure CTCF binding. We therefore performed luciferase assays by co-transfecting H9c2 cells with a pGL4.23 vector containing either the CTCF consensus motif or a scramble sequence, together with increasing amounts of the CTCF-VP64 expression vector. We observed that, when co-transfecting the scramble sequence together with CTCF-VP64, luciferase activity was undetectable, indicating that CTCF is not able to bind to this sequence. In contrast, we detected luciferase activity (i.e. CTCF binding) when the reporter vector containing the CTCF consensus motif was transfected. Importantly, luciferase activity increased with the amount of CTCF-VP64 expression vector transfected in a dose-dependent manner (**Figure 89**). Of note, we still detected luciferase activity when we transfected the CTCF motif sequence in the absence of CTCF-VP64. This could be due to the presence of endogenous CTCF, constitutively expressed in H9c2 cells.

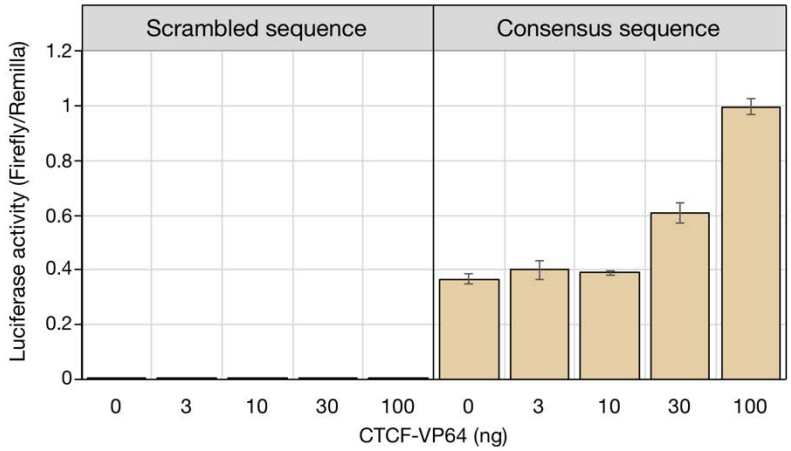


Figure 89. Verification of luciferase assay strategy to quantify CTCF binding. Luciferase assays in H9c2 cells transfected either with the scrambled or consensus sequences together with increasing amounts of the expression vector containing CTCF-VP64 (0, 3, 10, 30 and 100 ng). Firefly luciferase values were normalized by renilla activity (average ± SD, n=3).

Once demonstrated that this reporter strategy was valid to quantify CTCF binding for a specific sequence, we performed luciferase assays for the 59 CTCF-overlapping variants (**Table 28**). The luciferase results show that the number of variants significantly affecting CTCF binding was lower than those predicted by DeepBind. In particular, **21/59** variants were found to significantly **decrease** binding (43 according to DeepBind), **10/59** variants were found to significantly **increase** binding (14 according to DeepBind), and **28/59** variants were found to **not significantly affect** binding (2 according to DeepBind). Still, about 38% of the variants show the same effects in both methods: 16 decreasing binding, 5 increasing binding and 1 having no effects in binding (**Figure 90A**).

Table 28. Normalized luciferase activity for the 59 CTCF-overlapping variants. Firefly luciferase values were normalized by renilla activity (average ± SD, n=3).

Chr	Pos	Ref	Alt	Lucif-Ref	Lucif-Alt	Lucif-difference	Sig
chr3	37953719	T	A	0.8433 ± 0.05	0.7425 ±0.04	-0.1008	ns
chr3	38045036	T	A	0.7441 ± 0.08	0.5199 ± 0.08	-0.2242	*
chr3	38521504	A	G	0.7092 ± 0.07	0.6954 ± 0.02	-0.0137	ns
chr3	38780342	G	A	1.3549 ± 0.07	0.2517 ± 0.05	-1.1032	***
chr3	39540276	G	A	0.6293 ± 0.06	0.6484 ± 0.03	0.0192	ns
chr3	39854224	A	G	1.1747 ± 0.09	1.1781 ± 0.01	0.0034	ns
chr3	39953594	C	T	0.3621 ± 0.01	0.3987 ± 0.02	0.0366	ns
chr7	79677353	C	T	0.7606 ± 0.02	0.8164 ± 0.03	0.0558	ns
chr7	80288990	ATGT	A	0.5803 ± 0.01	0.6943 ± 0.03	0.1140	*
chr7	80549633	C	G	0.8832 ± 0.03	0.2496 ± 0.01	-0.6336	***
chr7	80625526	G	A	1.1249 ± 0.05	0.9235 ± 0.04	-0.2014	**
chr7	81076865	TC	T	0.7604 ± 0.03	0.4164 ± 0.02	-0.3441	***
chr7	81076870	A	T	0.7131 ± 0.04	0.4730 ± 0.05	-0.2401	**
chr7	81087389	T	C	0.2148 ± 0.01	0.1820 ± 0.01	-0.0328	*
chr7	81130725	C	T	0.0197 ± 0.00	0.1034 ± 0.00	0.0837	***
chr7	82110992	T	C	0.4407 ± 0.04	0.4008 ± 0.01	-0.0399	ns
chr7	82702421	G	A	0.3235 ± 0.01	0.3427 ± 0.01	0.0192	ns
chr7	82900741	T	G	0.2831 ± 0.00	0.4699 ± 0.01	0.1869	***
chr10	18883940	C	T	0.5733 ± 0.05	0.8599 ± 0.05	0.2866	**
chr10	19457714	C	G	0.8149 ± 0.06	0.4767 ± 0.01	-0.3382	**
chr10	21147418	G	A	1.5786 ± 0.06	1.3267 ± 0.05	-0.2519	**
chr11	117122398	A	T	0.6206 ± 0.03	0.6601 ± 0.03	0.0395	ns
chr11	117924278	C	T	0.8137 ± 0.06	0.7764 ± 0.06	-0.0373	ns
chr11	118042498	C	T	1.1930 ± 0.07	1.1931 ± 0.01	0.0001	ns
chr11	118549803	C	T	1.2375 ± 0.09	0.5183 ± 0.05	-0.7192	**
chr11	118560953	G	GC	0.8015 ± 0.05	0.3978 ± 0.02	-0.4037	**
chr11	118886117	C	T	0.3587 ± 0.02	0.4069 ± 0.02	0.0482	*
chr11	118889370	C	G	NP	NP	NP	NP
chr11	118889378	G	A	0.4326 ± 0.02	0.2379 ± 0.01	-0.1947	**
chr11	119287550	G	A	0.4676 ± 0.03	0.4745 ± 0.01	0.0069	ns
chr11	119352239	G	A	0.5085 ± 0.04	0.4831 ± 0.03	-0.0254	ns
chr11	119404719	C	T	1.3241 ± 0.07	1.2333 ± 0.01	-0.0908	ns
chr11	119600183	A	T	0.8935 ± 0.07	0.9345 ± 0.04	0.0410	ns
chr11	119612173	C	T	0.7499 ± 0.06	0.7502 ± 0.08	0.0003	ns
chr11	120053724	C	T	0.9914 ± 0.07	0.2438 ± 0.02	-0.7476	**
chr11	120100704	T	G	0.5773 ± 0.04	1.0202 ± 0.04	0.4429	***
chr11	120173911	C	CTGAAG	1.6252 ± 0.04	0.8501 ± 0.07	-0.7751	***
chr11	120173914	C	CGGG	0.7605 ± 0.07	0.3123 ± 0.01	-0.4482	**
chr11	120173917	C	CT	1.1069 ± 0.10	0.1985 ± 0.01	-0.9084	**

Chr	Pos	Ref	Alt	Lucif-Ref	Lucif-Alt	Lucif-difference	Sig
chr11	122659691	T	C	0.2473 ± 0.03	0.2443 ± 0.01	-0.0031	ns
chr11	123036887	G	C	1.6756 ± 0.09	0.2084 ± 0.01	-1.4673	**
chr11	123105986	G	T	0.5612 ± 0.05	0.6049 ± 0.07	0.0437	ns
chr11	123118102	CT	C	0.8194 ± 0.04	0.7303 ± 0.05	-0.0891	ns
chr11	123132370	G	C	0.8975 ± 0.10	0.7772 ± 0.08	-0.1203	ns
chr11	123425811	A	C	0.8335 ± 0.03	0.9113 ± 0.04	0.0778	ns
chr11	123447866	A	G	0.6002 ± 0.06	0.6135 ± 0.02	0.0133	ns
chr11	123511861	A	AC	0.1973 ± 0.01	0.2203 ± 0.02	0.0229	ns
chr11	123511862	C	T	0.2386 ± 0.36	0.1864 ± 0.01	-0.0521	**
chr11	124539211	G	A	1.9183 ± 0.12	2.2308 ± 0.04	0.3125	*
chr11	124648367	T	G	0.9603 ± 0.04	0.2863 ± 0.00	-0.6740	**
chr11	124984629	G	T	0.5025 ± 0.06	0.5927 ± 0.08	0.0902	ns
chr12	2469918	T	A	0.3104 ± 0.04	0.4633 ± 0.04	0.1529	*
chr12	2733860	C	T	0.8095 ± 0.08	1.0525 ± 0.05	0.2430	*
chr12	3053956	G	A	1.5154 ± 0.10	1.6204 ± 0.04	0.1050	ns
chr12	3384802	G	A	0.4721 ± 0.04	0.5220 ± 0.04	0.0499	ns
chr12	3451418	C	T	1.1073 ± 0.05	0.3697 ± 0.01	-0.7376	**
chr12	3764916	G	C	0.3922 ± 0.01	0.4029 ± 0.02	0.0107	ns
chr12	3767720	G	T	1.4192 ± 0.09	1.2380 ± 0.04	-0.1812	ns
chr12	3913211	G	A	0.5778 ± 0.04	0.8207 ± 0.10	0.2429	*

Chr (Chromosome), Pos (Position), Ref (reference llele), Alt (Alternative allele), Lucif (Luciferase activity), Lucif-Ref (Luciferase activity for reference allele), Lucif-Alt (Luciferase activity for alternative allele), sig (significance). Green highlights variants predicted to increase binding by luciferase activity and DeepBind, red highlights variants predicted to decrease binding by luciferase activity and DeepBind, orange highlights variants predicted to have no effects on CTCF binding by luciferase activity and DeepBind, and grey highlights a variant that could not be cloned for luciferase activity measurements. The statistical significance was measured using a two-tailed t-test. *p≤0.05, **p≤0.01 and ***p≤0.001

We then compared, for each of the 118 variants analyzed (reference and alternative), the luciferase activities (in absolute numbers) with the DeepBind prediction score. In this analysis, we observed that luciferase activities show a poor linear correlation with DeepBind scores (r^2 = 0.084; **Figure 90B**). It must be taken into account that luciferase activity and DeepBind scores were obtained from different strategies with their corresponding limitations. On one hand, luciferase activity was measured from *in vitro* experiments that do not take into account the original genomic context where the CTCF binding site tested is embedded. On the other hand, even DeepBind overcomes the genomic context limitation by being trained using *in vivo* experimental data, the scores outputted are based on computational predictions that cannot be accurate for a given variant. Therefore, due to the distinct nature of the results obtained, it is not surprising that luciferase activity and DeepBind scores show a poor linear correlation.

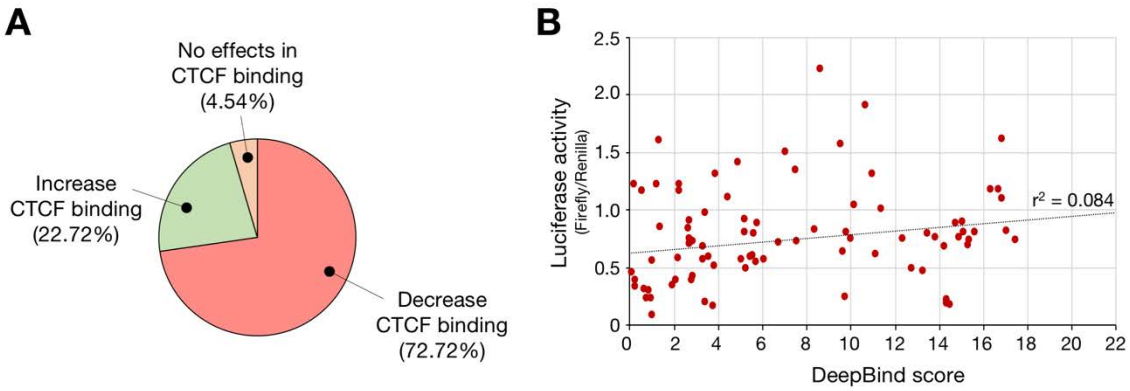


Figure 90. Comparison between luciferase assays and DeepBind scores. (A) Pie chart showing, from those variants have the same luciferase and DeepBind and predictions, the percentage of variants that decrease, increase or have no effects on CTCF binding. **(B)** DeepBind scores and luciferase activity (Firefly/Renilla) correlation for the 118 variants analyzed.

Based on the aforementioned observations, we speculated that DeepBind might be useful to predict variants affecting CTCF binding but not to quantify the effect the variants on binding. This hypothesis is confirmed by the observation that the binding effects measured in the luciferase assay (i.e. the absolute difference between reference and alternative variants) showed the same tendency than DeepBind predictions (**Figure 91**). Again, not all the categories show statistically significant differences in binding (analyzed by a pairwise t-test with Benjamini-Hochberg p-value correction for multiple comparisons as explained on Methods section 2.12.1). However, similarly to DeepBind, we observed that: (i) binding effects are increased as the variants are closer to the motif and, (ii) variants at core nucleotides show greater binding effects than variants affecting non-core nucleotides.

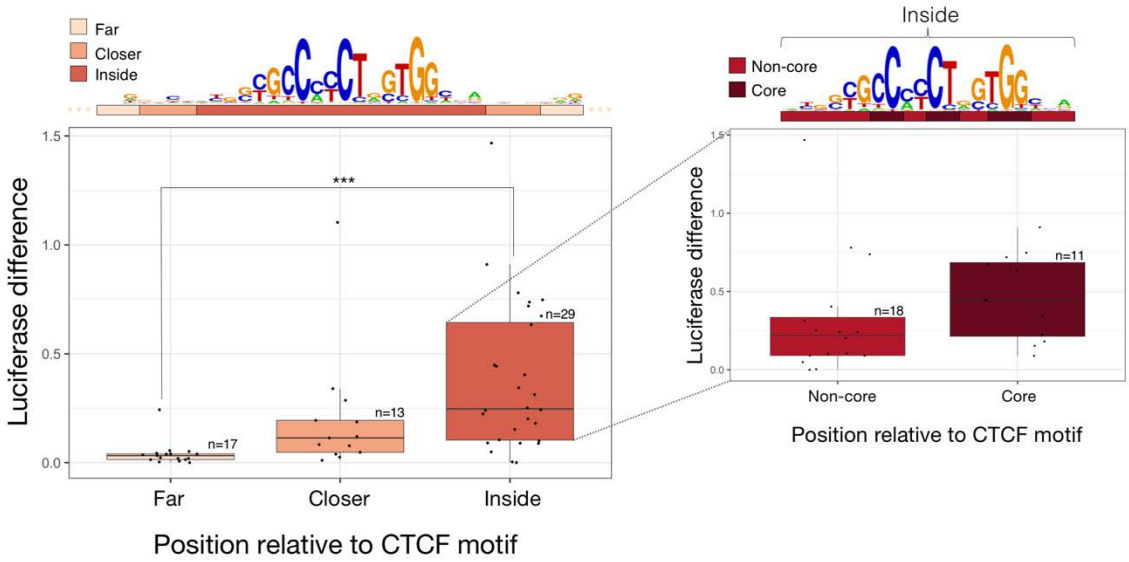


Figure 91. Positional binding effects obtained from luciferase assays. Boxplots showing the absolute binding difference (y-axis) for each group of variants, classified according to their position relative to the CTCF motif (x-axis). For indels affecting more than one position, only the first position affected was considered for variant classification. ***p≤0.001.

5.1.2.3. Candidate variants based on DeepBind and Luciferase results

After integrating the results obtained by DeepBind predictions and luciferase assays for the 59 CTCF-overlapping variants, we obtained a list of **21 candidate variants** potentially altering CTCF binding (**Table 29**). These 21 variants show significant effects on CTCF binding in luciferase reporter-based assays and exhibit the same predicted effects by DeepBind. As mentioned earlier, most of these variants (76.2%) are expected to decrease CTCF binding, while the remaining 23.8% are expected to increase CTCF binding. Independently of their effect, at least one variant was found in each of the 6 locus considered, although almost half of the variants were found in loci surrounding sodium channel regulatory beta subunits (*SCN2B* and *SCN3B*).

All described variants were already present in dbSNP150, gnomAD and 1000 Genomes databases at the time this analysis was performed.

Table 29. CTCF-overlapping variants affecting binding according to DeepBind predictions and luciferase activity.

Chr	Position	Ref	Alt	DB and Lucif. classification	Sig	Locus
chr3	38045036	T	A	Decrease	*	*SCN5A*
chr7	80549633	C	G	Decrease	***	*CACNA2D1*
chr7	80625526	G	A	Decrease	**	*CACNA2D1*
chr7	81076865	TC	T	Decrease	***	*CACNA2D1*
chr7	81076870	A	T	Decrease	**	*CACNA2D1*
chr7	81130725	C	T	Increase	***	*CACNA2D1*
chr10	18883940	C	T	Increase	**	*CACNB2*
chr10	19457714	C	G	Decrease	**	*CACNB2*
chr11	118549803	C	T	Decrease	**	*SCN2B*
chr11	118560953	G	GC	Decrease	**	*SCN2B*
chr11	118886117	C	T	Increase	*	*SCN2B*
chr11	118889378	G	A	Decrease	**	*SCN2B*
chr11	120053724	C	T	Decrease	**	*SCN2B*
chr11	120100704	T	G	Increase	***	*SCN2B*
chr11	120173911	C	CTGAAG	Decrease	***	*SCN2B*
chr11	120173914	C	CGGG	Decrease	**	*SCN2B*
chr11	120173917	C	CT	Decrease	**	*SCN2B*
chr11	123036887	G	C	Decrease	**	*SCN3B*
chr11	124648367	T	G	Decrease	**	*SCN3B*
chr12	3451418	C	T	Decrease	**	*CACNA1C*
chr12	3913211	G	A	Increase	*	*CACNA1C*

Chr (chromosome), Pos (position), Ref (reference allele), Alt (alternative allele), DB (DeepBind), Lucif (luciferase), Sig (significance). *p≤0.05. **p≤0.01 and ***p≤0.001.

5.2. Comparison between BrS and Wellderly cohorts

To identify BrS-candidate variants, we also used a second strategy. The main goal of this approach was to compare the frequency of genetic variants present in the Regulome-seq regions of the 89 BrS individuals with the frequency of genetic variants present in the Regulome-seq regions of 200 Wellderly individuals (Materials **Table 7**). The genetic variants from the Wellderly cohort had been previously obtained from WGS by the group of Dr. Eric Topol[241]. Wellderly cohort consists of healthy individuals >80 years old with no chronic diseases and not taking chronic medications at the moment of recruitment.

5.2.1. BrS and Wellderly ancestry admixture

The frequency of genetic variants in a given human genome is related to the ancestral origin of the genome studied. For this reason, and before making any assumptions when comparing two different populations, it is very important that both cohorts share the same ancestral origin. To determine whether BrS and Wellderly individuals derive from the same ancestry, we used the information from 2,504 individuals from 5 super-populations (African, American, East Asian, European and South Asian) available at the 1000 Genomes Project database (Materials section 1.6.4). We only considered SNVs within the Regulome-seq regions, and compared the frequencies of SNVs shared between these 2,504 individuals together with BrS and Wellderly individuals, to avoid the separation of populations for any other reason than the frequency of variants. This analysis was performed using **t-SNE**[257] as explained in Methods section 2.5. T-SNE is an algorithm for dimensionality reduction used to visualize high-dimensional data by giving each data point a location in a two or three-dimensional map. As expected, the results obtained with t-SNE analysis reveal that the frequency of variants analyzed is related to the ancestral origin of each individual (**Figure 92**). Importantly, both BrS and Wellderly cohorts overlap with individuals of European ancestry, demonstrating the European origin of both cohorts. However, a small group of European and Wellderly individuals are admixed with individuals from American ancestry. This admixture can be explained by the fact that the American population from 1000 Genomes includes Latin-american individuals that are descendants of the European colonizers that reached America after its discovery in 1,492.

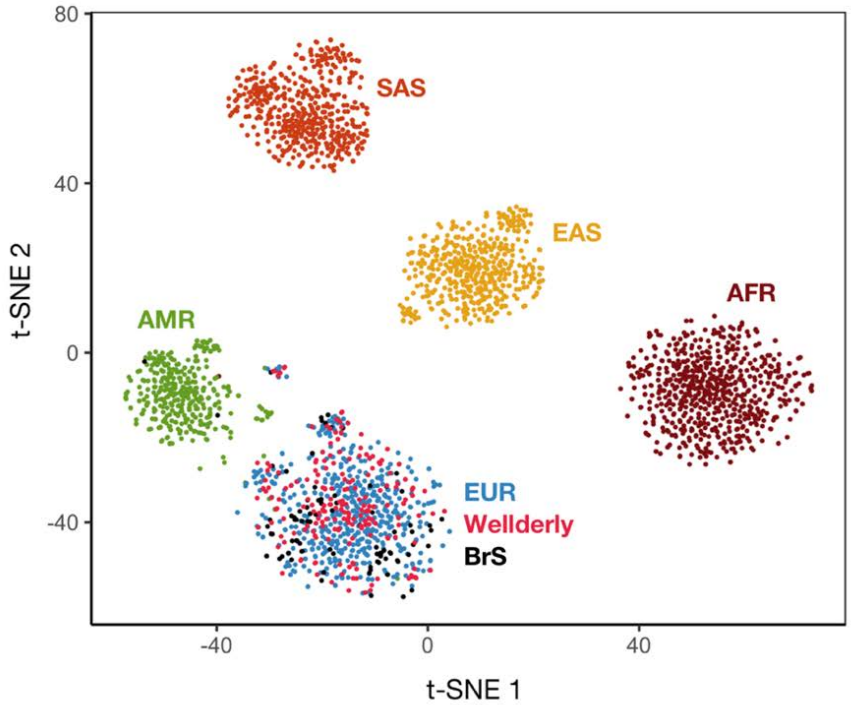


Figure 92. Ancestry admixture analysis. t-SNE analysis of 2,793 samples (dots): with 661 AFR (African, garnet), 347 AMR (American, green), 504 EAS (East-Asia, orange), 489 SAS (South-Asia, yellow), 503 EUR (European, blue), 200 Wellderly (pink), and 89 BrS (black).

From these analyses, we concluded that BrS and Wellderly individuals share the same European ancestry. Therefore, these cohorts could be compared in further analysis as we describe in sections below.

5.2.2. Curation of the Wellderly variant call set

Previous to the comparison of the frequency of genetic variants between BrS and Wellderly cohorts, we filtered the list of variants found in the 200 Wellderly individuals (Wellderly raw) as explained in Methods section 2.4.3. After each filter, the average number of variants per individual was used to determine if the BrS and Wellderly variant call sets were comparable (the average number of variants per individual should be similar in both cohorts as they share the same ancestral origin).

The variant discovery pipeline–followed by the group of Dr. Eric Topol–to identify genetic variants in the Wellderly cohortwas identical to ours. Therefore, **Filter 1** (removing 600 variants labelled as low quality after GATK variant recalibration) and **Filter 2** (removing 271 variants with non-resolved genotypes for some samples) were applied right away (**Table 30**). However, the coverage filter had to be redefined because the average coverage was not comparable between BrS and Wellderly cohorts (384X for BrS versus 30X for Welldelry). This difference in coverage

is explained by the fact that BrS genetic information was obtained through targeted sequencing, where all the sequencing power was destined to a small fraction of the genome (1.13 Mb). In contrast, the Wellderly data was obtained through WGS and therefore, the sequencing power had to be distributed along the whole genome (3.2 Gb). In consequence, we tested two different coverage filters consisting in the removal of all genomic positions that were not covered ≥10X or ≥20X in all samples (**Filters 3** and **4**, respectively).

After applying Filter 1 + Filter 2, the average variants per individual in the Wellderly cohort was 1,492 ± 63 (209 variants more than BrS; **Table 30** and **Figure 93A**). With Filter 3 (≥10X) we removed 772 variants, resulting in an average of 1,349 ± 58 variants per individual in the Wellderly cohort (66 variants more than BrS; **Table 30** and **Figure 93B**). With Filter 4 (≥20X) we removed 4,332 variants, resulting in an average of 652 ± 33.9 in the Wellderly cohort (631 variants less than BrS; **Table 30** and **Figure 93C**).

Table 30. Filters applied in Wellderly variant call set. Number of initial variants, filtered variants and remaining variants in the 200 Wellderly individuals after applying each filter. The average number of variants per individual for Wellderly and BrS cohorts are also shown. The final filter selected is highlighted in bold.

Filter	Initial variants	Filtered variants	Remaining variants	Var/ind Wellderly*	Var/ind BrS*
Filter 1 (Low quality)	9,200	600	8,600	-	1,283 ± 59.3
Filter 2 (non-resolved genotypes)	8,600	271	8,329	1,492 ± 63	1,283 ± 59.3
Filter 3 (≥10X in all samples)	**8,329**	**772**	**7,557**	**1,349 ± 58**	**1,283 ± 59.3**
Filter 4 (≥20X in all samples)	8,329	4,332	3,997	652 ± 33.9	1,283 ± 59.3

Var/ind (variants per individual). *Results presented as average ± SD.

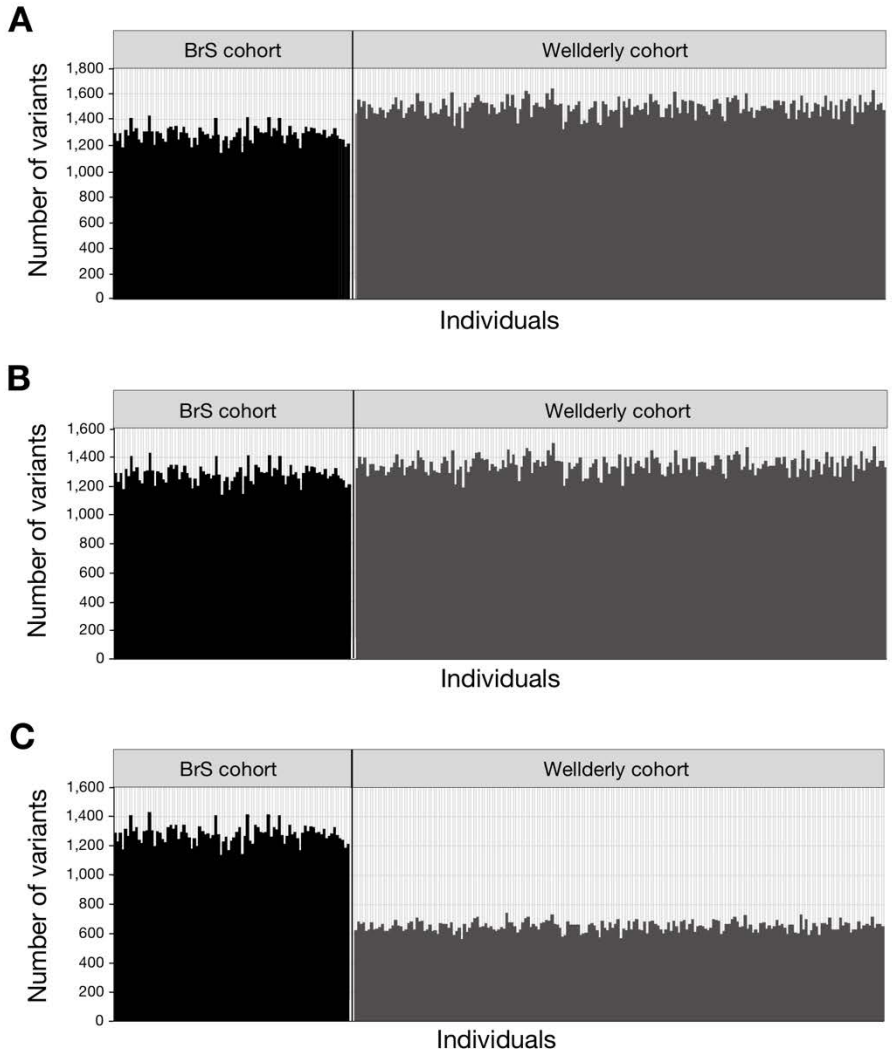


Figure 93. Curation of Wellderly variant call set. (A-C) Number of variants (y-axis) per individual (x-axis) for BrS and Wellderly cohorts after applying Filter 1+2 **(A)**, after filtering for ≥10X coverage in Filter 3 **(B)** and after filtering for ≥20X coverage in Filter 4 **(C)**.

In the light of these results, the filter chosen to curate our variant call set was **Filter1+2** followed by **Filter 3 (≥10X)**, given that it was showing an average number of variants per individual very similar for both cohorts, ensuring that the differences observed in the subsequent analysis are not due to technical differences.

5.2.3. Selection of variants significantly enriched in BrS individuals

In addition to the ancestral origin, another element to consider when comparing human cohorts is the number of individuals included in each group. The higher number of individuals in the Wellderly cohort might result in the identification of different variants than BrS cohort, but it might also result in shared variants displaying different AF. To be able to compare the

Wellderly cohort with the BrS cohort, we performed **100 permutations** in the Wellderly cohort using an in-house perl script as described in Methods section 2.12.3. In each permutation, we randomly selected 89 Wellderly individuals and calculated the AF of each variant. Then, we compared the variantants from Wellderly indiviuals with the variants found in the 89 BrS individuals. We found that 1,608 variants are BrS-specific, 3,567 are Wellderly-specific and 3,738 are shared by the two cohorts (**Figure 94A**).

For shared variants, we determined the number of times that the AF was different (higher or lower) in the Wellderly than BrS cohort, and we divided this value by the total number of permutations performed to compute a significance p-value for each variant. The remaining cohort-specific variants were differently processed as it will be explained later on.

We found that **336** variants are enriched in the BrS cohort (**BrS-enriched**; p-value < 0.05), **223** variants are enriched in Wellderly cohort (**Wellderly-enriched**; p-value < 0.05), and **3,179** variants are equally frequent in both cohorts (**No-difference**; p-value < 0.05). An example of each category is shown in **Figure 94B**.

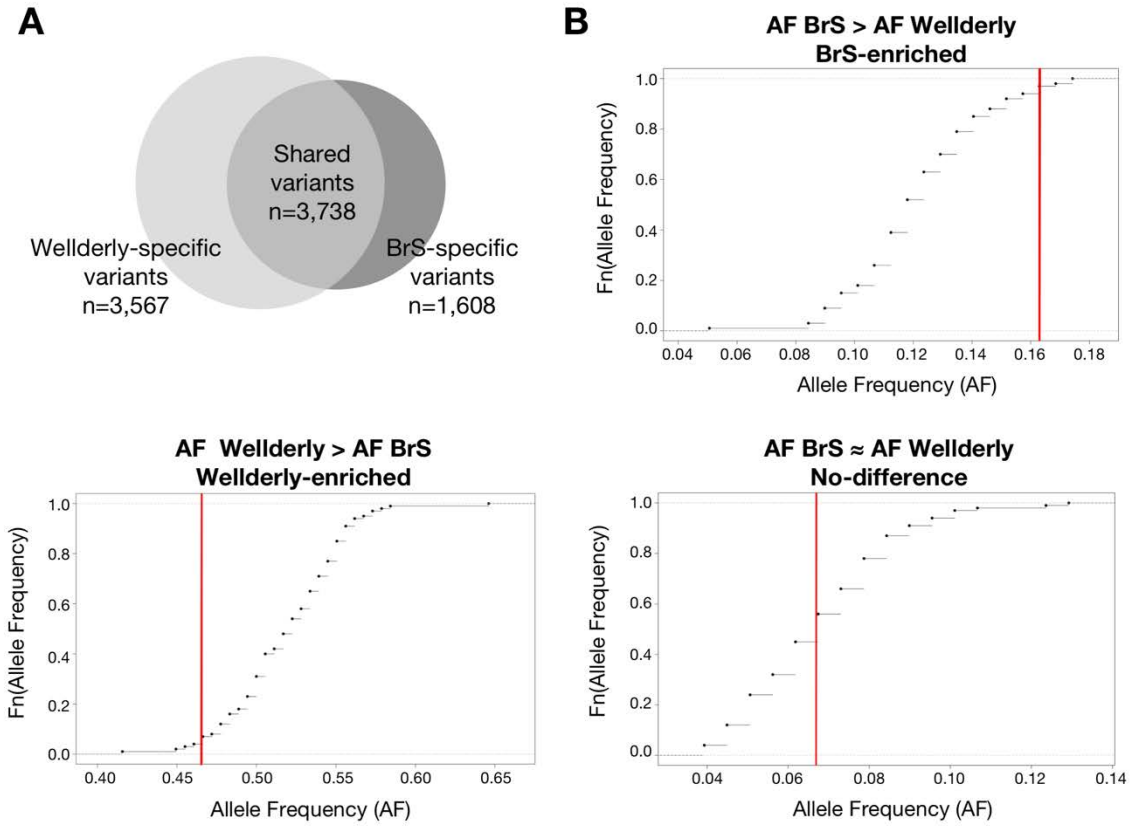


Figure 94. Comparison of BrS and Wellderly cohorts using permutations. (A) Venn diagram showing the number of variants that are only present in one of the cohorts (Wellderly-specific and BrS-specific) and the number of variants shared by the two cohorts. **(B)** Cumulative frequency distribution of three different variants classified as Wellderly-enriched, BrS-enriched and No-difference. The vertical red line corresponds to the frequency of the variant in the BrS cohort.

After analyzing the shared variants, we proceeded with the analysis of BrS-specific variants. These are especially interesting in this thesis because genetic variants found in BrS individuals and not in any healthy individual might explain BrS phenotype. Ideally, we should analyze thousands of individuals in each cohort to identify BrS-specific variants with a strong association to BrS phenotype. However, the fact that we are dealing with such a small sample size increases the probability of identifying a variant in one cohort and not the other by chance, which could lead to a miss-association of variants to BrS phenotype. To overcome this issue, we used **Pearson's chi-squared test** to interrogate whether BrS-specific variants are more likely to be related to BrS individuals than Wellderly individuals, as explained in Methods section 2.12.4. Chi-squared test is typically used to determine if the difference between the expected frequencies and the observed frequencies in one or more categories is significant. We used the Wellderly cohort as our expected category and BrS cohort as our observed category, and the comparative was performed between the number of reference and alternative alleles found for each variant in each cohort. It is important to note that, for the present study, we only analyzed BrS-specific variants, although we also aim to identify significant Wellderly-specific variants for future analysis.

After running the Pearson's chi-squared test on the 1,608 BrS-specific variants, we found that only **201 variants** are statistically significant (p-value < 0.05).

Together, the permutations and the chi-squared test allowed us to narrow down the initial 5,349 variants identified in the Regulome-seq regions of 89 BrS individuals, to a list of **537 variants** (336 BrS-enriched and 201 BrS-specific) that are significantly enriched in BrS individuals.

5.2.4. Scoring variants significantly enriched in BrS individuals

From the 537 variants significantly enriched in BrS individuals, we sought to identify those that are more likely to have stronger associated functional consequences. For this purpose, we used the genome-wide pre-computed scores of deleteriousness (**CADD scores**[150]) and pre-computed scores of tolerance to genetic variation (**CDTS**[158]), described in Materials section 1.6.5.2.

5.2.4.1. CADD scores

As reviewed in the Introduction section 2.5, the CADD framework integrates diverse functional annotations such as regulatory information into a score of deleteriousness, which

strongly correlates with molecular functionality and pathogenicity. CADD scores increase as the variants are predicted to be more deleterious, and a cutoff of 15 is suggested by the authors to identify potentially pathogenic variants.
We obtained CADD scores for the 537 variants significantly enriched in BrS individuals (201 BrS-specific and 336 BrS-enriched), as well as for the 3,179 No-difference variants.

Globally, we observed that there is a wide distribution of CADD scores in the three different groups of variants (ranging from 0 to 25). We also noticed that the median CADD score of each group is similar, although slightly higher for the BrS-specific variants (4.88 for BrS-specific, 3.99 for BrS-enriched and 4.61 for No-difference). This difference is even more pronounced for those variants found within the ***SCN5A*** locus, where BrS-specific variants show a 1.8-fold increase in the median CADD score relative to No-difference variants (**Figure 95A**). Similar observations can be extracted from those variants found in ***CACNA1C*** and ***CACNB2*** locus, where BrS-specific variants show a 1.1-fold increase and 1.6-fold increase in the median CADD score relative to No-difference variants, respectively (**Figure 95D** and **E**). In contrast, for those variants found within the ***SCN2B***, ***SCN3B*** and ***CACNA2D1*** loci, BrS-specific variants show a similar median CADD score than No-difference variants (**Figure 95B, C** and **F**). When analyzing BrS-enriched variants, we found that they only show higher CADD scores than the No-difference variants when they are present in the *SCN5A* locus (**Figure 95**).

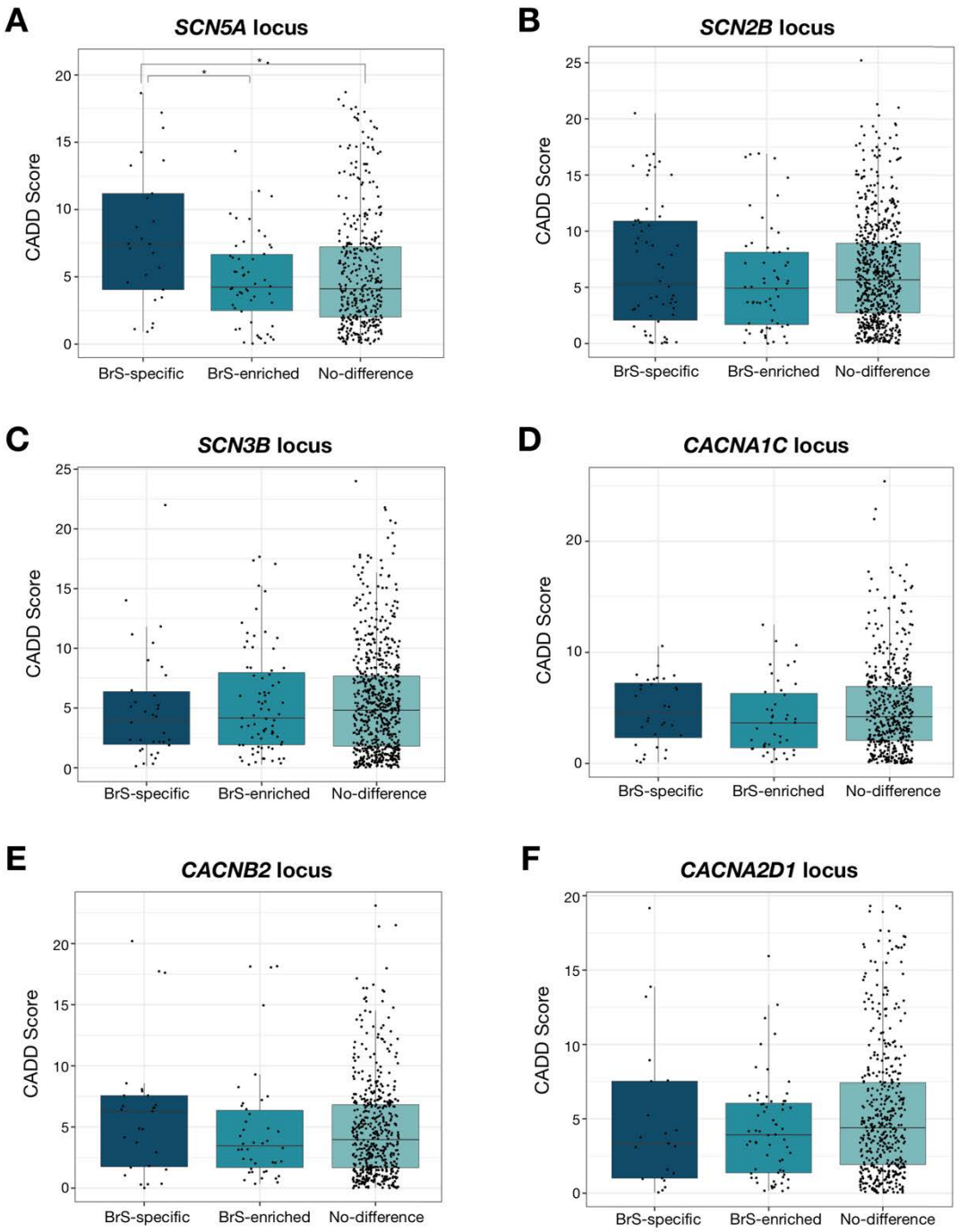


Figure 95. CADD score distribution for each category and locus. (A-F) Boxplots of the median CADD scores for BrS-specific variants, BrS-enriched variants and No-difference variants. Variants are separated based on the BrS-associated locus where they are found: *SCN5A* **(A)**, *CACNB2* **(B)**, *CACNA1C* **(C)**, *CACN2D1* **(D)**, *SCN2B* **(E)** and *SCN3B* **(F)**. Statistical differences between categories of variants were calculated using a pairwise t-test with Benjamini-Hochberg p-value correction for multiple comparisons. Only the significant differences are represented (*p≤0.05).

We next examined the proportion of potentially pathogenic variants (with a **CADD score ≥15**) in each of the three variant categories (BrS-specific, BrS-enriched and No-difference variants). Overall, we found that approximately 9% of BrS-specific variants are classified as potentially

pathogenic, while only 3.8% of BrS-enriched variants and 5.31% of No-difference variants are classified as such. This result indicates that BrS-specific variants are more likely to be pathogenic than the other groups.

We next performed this analysis for the variants surrounding each BrS-associated gene. We observed that, for variants found within ***SCN2B, SCN5A*** and ***CACNB2*** loci, the highest proportion of potentially pathogenic variants belongs to the category of BrS-specific variants (17%, 12% and 11%, respectively), compared to 5.3%, 4.0% and 2.79% of No-difference variants (**Table 31**). In contrast, for variants found within ***CACNA2D1*** and ***CACNA1C***, the highest proportion of potentially pathogenic variants belongs to the category of No-difference variants (5.3% and 3.25%, respectively). Only in the case of ***SCN3B*** locus, BrS-enriched variants are displaying a higher proportion of potentially pathogenic variants (5.33%) compared to the observed 4.52% of No-difference variants (**Table 31**).

Table 31. Proportion (%) of potentially pathogenic variants (CADD score ≥15) for each category and locus.

	SCN5A locus	*SCN2B* locus	*SCN3B* locus	*CACNA1C* locus	*CACNB2* locus	*CACNA2D1* locus
BrS-specific	12.00	16.95	2.94	0.00	11.11	4.55
BrS-enriched	1.85	6.90	5.33	0.00	7.14	1.56
No-difference	4.04	5.33	4.52	3.26	2.79	5.31

Together, using the criteria of CADD scores ≥15, we identified 31 variants significantly enriched in BrS individuals that were classified as potentially pathogenic. From these variants, 18 are BrS-specific while the remaining 13 are BrS-enriched.

The full list of the 31 variants identified in this section is not provided, as the final goal was to complement this annotation with CDTS criteria explained below. Of note, potentially pathogenic variants separately identified by CADD ≥15 will not be completely discarded in future studies, but they will not be discussed further in this thesis.

The variants resulting from the combination of both CADD scores and CDTS percentiles are shown in Results section 5.2.4.3.

5.2.4.2. CDTS percentiles

As reviewed in the Introduction section 2.5, CDTS measures how tolerant are genomic regions to genetic variation in the context of surrounding sequences. CDTS are organized in percentiles, being regions found in the 1st percentile the least tolerant to genetic variation. Those

variants found in regions less tolerant to variation are presumed to have the highest impacts in terms of disease.

We obtained the CDTS for the 537 variants significantly more frequent among BrS individuals (201 BrS-specific and 336 BrS-enriched), as well as for the 3,179 No-difference variants.

Globally, we found that in all categories there are variants spanning all CDTS percentiles. Notwithstanding, genetic variants significantly enriched in BrS individuals tend to be more often found at lower percentiles than No-difference variants. This divergence is more pronounced in the case of BrS-specific variants (median of 31.98 for BrS-specific, 42.72 for BrS-enriched and 44.64 for No-difference).

Next, we examined the proportion of potentially pathogenic variants (fount in the **1st CDTS percentile**) in each of the three variant categories (BrS-specific, BrS-enriched and No-difference variants). We observed that genetic variants significantly enriched in BrS individuals show a higher proportion of potentially pathogenic variants than No-difference variants, especially in the category of BrS-specific variants: 16.44% of BrS-specific variants, 7.29% of BrS-enriched variants and 6.50% of No-difference variants.

When analyzing the variants surrounding each BrS-associated gene, we clearly observed that in four loci (***SCN5A***, ***SCN2B***, ***CACNB2*** and ***CACNA2D1***), the category of BrS-specific variants shows the highest proportion of potentially pathogenic variants compared to No-difference variants (**Figure 96A**, **B**, **E** and **F**). The two exceptions are the ***SCN3B*** locus, in which the proportion of potentially pathogenic variants is similar among BrS-specific and BrS-enriched variants (**Figure 96C**), and ***CACNA1C*** locus, in which the proportion of potentially pathogenic variants is superior in BrS-enriched variants (**Figure 96D**).

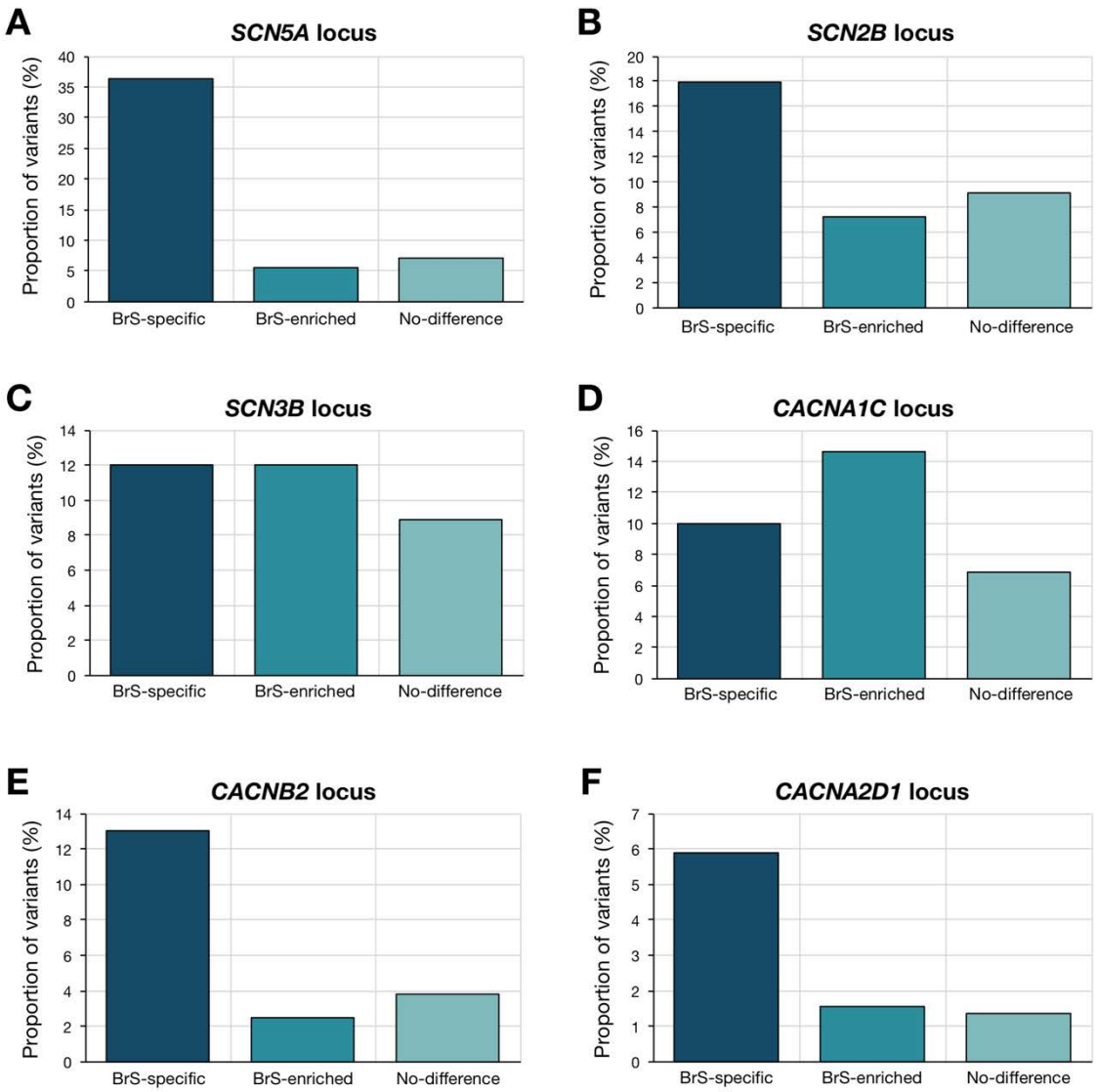


Figure 96. Proportion of variants at the 1st CDTS percentile for each category and locus. (A-F) Proportion (y-axis) of BrS-specific variants, BrS-enriched variants and No-difference variants (x-axis) found in the 1st CDTS percentile. Variants are separated for each BrS-associated locus: *SCN5A* **(A)**, *CACNB2* **(B)**, *CACNA1C* **(C)**, *CACN2D1* **(D)**, *SCN2B* **(E)** and *SCN3B* **(F)**.

The observation that variants significantly enriched in BrS individuals are more often found in regions less tolerant to variation and that, precisely, this group is the one showing a higher proportion of potentially pathogenic variants, suggests that these variants are more likely to be pathogenic than No-difference variants.

Together, using the criteria of 1st CDTS percentile, we identified 48 variants significantly enriched in BrS individuals that may be considered as potentially pathogenic. Furthermore, 50% of the variants are BrS-specific and the other 50% are BrS-enriched.

Similar than with the CADD criteria, the full list of 48 variants identified in this section is not provided, as the final goal was to complement this annotation with the CADD criteria explained above. Of note, potentially pathogenic variants separately identified by 1st CDTS percentile will not be completely discarded in future studies, but they will not be discussed further in this thesis.

The variants resulting from the combination of both CADD scores and CDTS percentiles are detailed in the section below.

5.2.4.3. Candidate variants based on CADD and CDTS combination

The combination of CADD scores (CADD score ≥15) with CDTS percentiles (1st percentile) led to a list of 10 possible candidate variants to BrS–7 BrS-specific and 3 BrS-enriched–(**Table 32**). These variants are located in 4 of the 6 locus studied: 2 variants in *SCN5A,* 2 variants in *CACNB2,* 4 variants in *SCN2B,* 2 variants in *SCN3B*. Interestingly, all these variants are related to regions enriched in H3K4me3, a histone mark often associated to active promoters.

The variants are shared by a variable number of individuals, ranging from private variants found in only 1 individual to variants shared by 65 individuals. All variants were already present in dbSNP150, gnomAD and 1000 Genomes databases.

Table 32. BrS-candidate variants according to the combination of CADD scores and CDTS percentiles.

Chr	Pos	Ref	Alt	Category	Locus
chr3	38180103	C	CGGCGG	BrS-specific	*SCN5A*
chr3	39194227	G	A	BrS-specific	*SCN5A*
chr10	17272622	C	A	BrS-specific	*CACNB2*
chr10	20105996	TGGC	T	BrS-specific	*CACNB2*
chr11	116658654	G	A	BrS-enriched	*SCN2B*
chr11	118305802	G	C	BrS-enriched	*SCN2B*
chr11	118308085	CCTCTCGGGCGT	C	BrS-specific	*SCN2B*
chr11	118927907	C	G	BrS-specific	*SCN2B*
chr11	123065730	C	A	BrS-specific	*SCN3B*
chr11	123525066	G	A	BrS-enriched	*SCN3B*

Chr (chromosome), Pos (position), Ref (reference allele), Alt (alternative allele).

In summary, this strategy, based on the comparison between BrS and Wellderly cohorts, led to the classification of the variants as BrS-specific, BrS-enriched and No-difference. In addition, the combination of CADD scores and CDTS percentiles allowed us to propose some candidate variants that may explain the molecular basis of BrS pathogenesis. However, further analyses need to be performed to finally associate these candidates to BrS.

5.3. CTCF candidates based on CADD and CDTS information

Here, we used two different approaches to identify variants that may be functionally relevant to BrS. The first strategy, focused on the identification of genetic variants affecting TF binding, led to a list of 21 possible candidates affecting CTCF binding (**Table 29**). The second strategy, focused on the comparison between BrS and Wellderly cohorts, followed by the scoring of variants based on CADD scores and CDTS percentiles, led to a list of 11 potentially pathogenic variants (**Table 32**).

It is important to acknowledge that both strategies were applied successively, therefore, once obtained all the new information from BrS-Wellderly comparison, we re-examined the 21 CTCF-overlapping variants possibly affecting CTCF binding. We found that 10 variants are equally frequent among BrS and Wellderly cohorts (No-difference variants); 1 variant is shared between both cohorts, but more frequent among Wellderly individuals (Wellderly-enriched); 9 variants are exclusive of the BrS cohort (BrS-specific)–although only 2 of them are significant based on the Pearson's chi-squared test–; and, 1 variant is shared between both cohorts, but more frequent among BrS individuals (BrS-enriched; **Table 33**).

Next, we explored the distribution of CADD scores and CDTS percentiles among BrS-specific and BrS-enriched variants. They showed a median CADD score of 6.92 and a median CDTS percentile of 53.99. When examining those variants falling in the CADD and CDTS pathogenicity criteria (CADD score ≥15 or 1st CDTS percentile), we observed that only 1 variant shows a CADD score ≥15, while 2 completely different variants are found in the 1st CDTS percentile (**Table 33**, highlighted in bold). Therefore, none of the BrS-specific and BrS-enriched CTCF variants meet the CADD and CDTS criteria in conjunction.

Table 33. CTCF candidate variants complemented with CADD and CDTS information.

Chr	Pos	Ref	Alt	CADD Score	CDTS Percentile	Category	Chi-sq	Locus
chr3	38045036	T	A	9.630	51.767	BrS-specific	ns	*SCN5A*
chr7	**80549633**	**C**	**G**	**16.520**	**72.673**	**BrS-specific**	**ns**	***CACNA2D1***
chr7	80625526	G	A	3.248	6.513	BrS-enriched	-	*CACNA2D1*
chr7	81076865	TC	T	5.757	75.800	BrS-specific	ns	*CACNA2D1*
chr7	81076870	A	T	8.077	75.800	BrS-specific	ns	*CACNA2D1*
chr7	81130725	C	T	6.032	25.259	Wellderly-enriched	-	*CACNA2D1*
chr10	18883940	C	T	2.451	49.871	No-difference	-	*CACNB2*
chr10	19457714	C	G	0.734	56.219	BrS-specific	ns	*CACNB2*
chr11	118549803	C	T	7.946	82.057	No-difference	-	*SCN2B*

Chr	Pos	Ref	Alt	CADD Score	CDTS Percentile	Category	Chi-sq	Locus
chr11	**118560953**	**G**	**GC**	**10.560**	**0.182**	**BrS-specific**	*******	***SCN2B***
chr11	118886117	C	T	15.270	37.208	No-difference	-	*SCN2B*
chr11	**118889378**	**G**	**A**	**4.166**	**0.521**	**BrS-specific**	*******	***SCN2B***
chr11	120053724	C	T	7.884	69.767	No-difference	-	*SCN2B*
chr11	120100704	T	G	15.050	74.325	No-difference	-	*SCN2B*
chr11	120173911	C	CTGAAG	15.360	8.362	No-difference	-	*SCN2B*
chr11	120173914	C	CGGG	17.180	8.362	No-difference	-	*SCN2B*
chr11	120173917	C	CT	18.310	8.362	No-difference	-	*SCN2B*
chr11	123036887	G	C	4.519	7.599	BrS-specific	ns	*SCN3B*
chr11	124648367	T	G	0.934	35.774	No-difference	-	*SCN3B*
chr12	3451418	C	T	10.550	67.454	BrS-specific	ns	*CACNA1C*
chr12	3913211	G	A	3.528	65.384	No-difference	-	*CACNA1C*

Chr (chromosome), Pos (position), Ref (reference allele), Alt (alternative allele), Chi-sq (Pearson's chi-squraed test). *p≤0.05. **p≤0.01 and ***p≤0.001.

In conclusion, the two strategies applied in this thesis allowed us to identify a few set of candidate variants possibly related to BrS phenotype. When comparing the information obtained from both approaches, we observe that some of the variants accepted as candidates by one strategy would be directly discarded by the other, which manifests the complexity of prioritizing non-coding variants and highlights the need of developing new approaches for this purpose. Finally, it is important to keep in mind that the validity of each strategy cannot be determined until experimentally demonstrated that these candidates are involved in BrS pathogenesis.

V. Discussion

Discussion

In the present thesis, we sought to characterize, for the first time, the genetic variation found within *cis*-regulatory elements of six BrS-associated genes (*SCN5A, SCN2B, SCN3B, CACNA1C, CACNB2* and *CACNA2D1*). For this purpose, we have developed a targeted strategy, referred to as **Regulome-seq**, to selectively capture and sequence these regions in a cohort of 89 *SCN5A*-negative BrS cases. With our strategy, we have identified a total of 5,349 variants within the *cis*-regulatory regions of these patients (4,837 SNVs and 512 indels: 219 insertions and 293 deletions). To identify those variants that lie within TF binding sites and predict their effect on TF binding, we have used CTCF ChIP-seq data together with a machine learning algorithm and luciferase reporter assays. We have also compared genetic variation within the Regulome-seq regions of BrS individuals with a healthy-aging cohort (Wellderly), leading to the identification of those variants that are significantly enriched in BrS individuals (either specific or shared with the Wellderly cohort). Finally, we have scored the variants based on the tolerance to variation and other parameters, which has allowed us to propose candidate variants that may explain the molecular basis of some BrS 'orphan' cases.
To better understand the complexity of our study, there are several interesting issues that will be discussed below.

1. Development of the Regulome-seq strategy

Cis-regulatory elements involved in the regulation of gene transcription are embedded in the non-coding fraction of the genome, which covers approximately 98% of the whole genome. Genetic profiling of non-coding regions, therefore, requires the sequencing of the entire genome. WGS is increasingly being promoted as a platform for investigating the full spectrum of genetic variation associated with human disease[265]. However, WGS still remains technically and financially prohibiting for most laboratories in the world. Perhaps more importantly, the analysis and interpretation of the amount of information obtained from WGS remain a challenge even for those multinational consortiums that can afford to perform these studies.

To overcome this issue, we took advantage of the already available targeted sequencing technology–broadly used in clinical settings when applied to coding sequences (exome sequencing)–to profile non-coding variants at *cis*-regulatory regions exclusively linked to BrS-associated genes. Together, our **Regulome-seq** approach allowed us to reduce the overall cost of the project and analyze dozens of human samples at the same time (**Figure 97**). Furthermore,

it generated less raw and more disease-directed sequence data than a whole human genome, which enormously facilitated the subsequent analysis and interpretation of the results obtained.

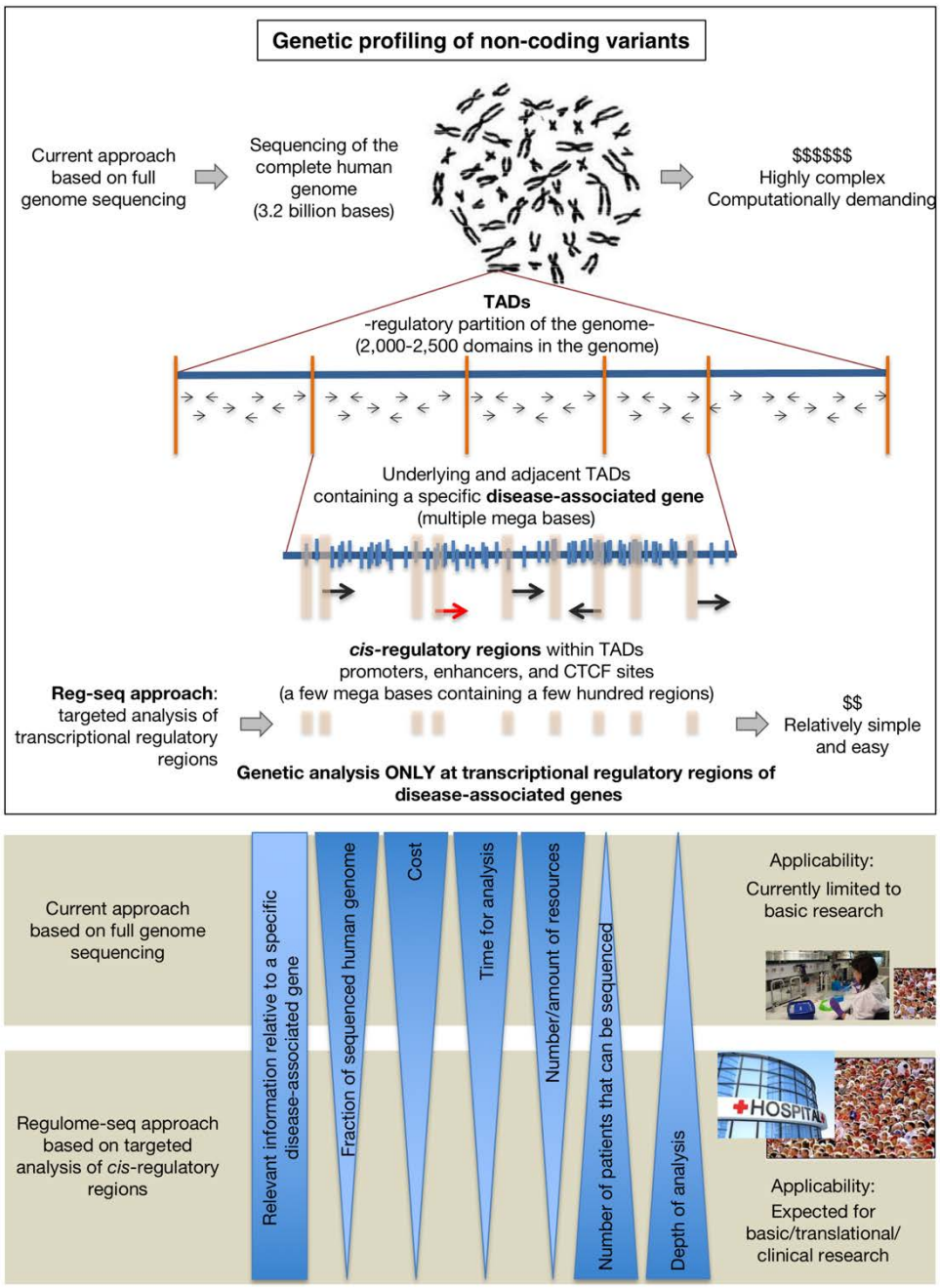


Figure 97. Advantages of the Regulome-seq approach compared WGS for the study of genetic variants at non-coding regions of the genome.

In addition to the advantages described above, the analysis of the sequencing results shows that all regions sequenced are highly and homogeneously covered (average coverage of 384X ± 149). Therefore, our approach clearly surpassed the limitations of capture by hybridization methods that tend to present an uneven coverage across the targeted regions–with some

regions being over-covered and some regions being missed or barely covered–. This high coverage allowed us to reach a call rate similar to that required for clinical diagnosis from massively parallel sequencing data[266], since 96% of the sequenced nucleotides were covered by at least 30X. Together, these results suggest that our Regulome-seq approach has the potential to be used as a diagnostic tool for genetic diseases in the future.

It is also important to highlight that, when this project was designed, our Regulome-seq approach was pioneer in the sense that no previous studies using targeted sequencing of non-coding regions were published. At present, Regulome-seq is no longer unique in this regard as a very recent paper published on Nature by Short *et al.,*[267] already uses targeted sequencing of non-coding regions to explore the role of *de novo* mutations in neurodevelopmental disorders. Their approach allowed the authors to observe a significant excess of *de novo* mutations only in highly evolutionary conserved elements that are active in fetal brain. However, they did not directly target *cis*-regulatory regions related to brain development. Instead, they indiscriminately sequenced more than 5,000 highly evolutionarily conserved non-coding elements and experimentally validated enhancers, some of them not even related to brain. Consequently, to be able to associate the observed *de novo* mutations to neurodevelopmental disorders, Short and colleagues had to filter *a posteriori* those regulatory elements not active in fetal brain, using DHS data from fetal brain and chromHMM, a software for learning and characterizing chromatin states[268].

Contrary to the aforementioned strategy, our Regulome-seq approach is more disease-directed because it only targets *cis*-regulatory elements active in the tissue of interest (human heart in our case), identified using available regulatory information (DHS, H3K4me3 and CTCF). Indeed, the observation that a polymorphism previously associated to cardiac conduction defects (rs6801957)[259] is embedded in one of our targeted regions indicates that our Regulome-seq approach is effective in selecting disease-associated regulatory regions. Therefore, although Regulome-seq is not the only approach that uses targeted sequencing of non-coding regions, to the best of our knowledge, it is the only strategy that targets regulatory regions directly linked to disease-associated genes.

2. Identification of Regulome-seq variants

The power and advantages of next-generation sequencing technologies are based on its massively parallel interrogation of nucleic acids. The ability to simultaneously evaluate millions of DNA base pairs allows clinicians and researchers to better understand the role of DNA

variation within our genome in terms of evolution, disease, drug-resistance, etc. In the field of clinical diagnosis, next-generation sequencing technologies have become frontline assays for a wide variety of inherited disorders, facilitating further advances in disease prediction and therapeutic decision-making for at-risk patients[266]. However, the accuracy of these tests can vary based on a number of variables including the sequencing technology, the coverage and the bioinformatics pipeline used for variant discovery. Some of the variants detected using next-generation sequencing technologies can have serious medical implications for the tested probands and their family members; therefore, the quality criteria for accepting the reported variants in the clinics are quite conservative (the variants have to be covered ≥30X) and sometimes they have to be complemented by Sanger sequencing[269] validation of low quality variants[266]. In contrast, when next-generation sequencing technologies are used for research purposes only, there are no guidelines established and the quality criteria to accept the variants are completely arbitrary.

We consider that our Regulome-seq project is found in a midpoint between research and clinical diagnosis, as it is a research project whose information intends to be translated into the clinical setting. For this reason, our main goal was to report a high-confident set of variants, even losing a few true positives. Moreover, since the coverage and sequencing quality reached by our Regulome-seq approach are high, we considered that it was not necessary to perform Sanger sequencing validation of our low quality variants. Instead, we optimized the filtering conditions to be applied after variant discovery using the Coriell NA12249. Application of these filters allowed us to reach a significant PPV of 0.99, pointing out that only 1 out of 100 variants reported will not be true (Results **Table 24**). Our decision of not validating low quality variants using Sanger sequencing was further supported by several studies showing that about 98-99% of the variants identified through next-generation sequencing panels are consistent with variants identified through Sanger sequencing[270,271].

It is important to remark that our filtering conditions as well as the analysis of sensitivity and PPV after each filter using an internal control (the Coriell NA12249 in our case), is not part of the standard variant discovery pipeline of research projects aimed to study genetic variation in human individuals. These projects usually use GATK for variant discovery, which is highly validated and it already applies powerful internal filters that result in a notable accuracy of the variants identified (as observed in our GATK output, Results **Table 23**). Therefore, researchers usually establish a subjective coverage threshold for the sequenced regions to be included in the variant call and then, they accept all variants reported by GATK in these accepted regions. For example, in the previously mentioned targeted sequencing analysis of non-coding regions,

the authors removed all targeted sequences with less than 10X coverage and did not apply any subsequent filter to the variants reported by GATK[267].

3. Prioritization of Regulome-seq variants

The tremendous progress recently achieved in massively parallel sequencing technologies enables investigators to efficiently obtain huge amounts of genetic information, whose functional significance is often difficult to interpret. Moreover, most disease-associated variants are found in non-coding regions of the genome[272,273], especially in *cis*-regulatory elements such as promoters, enhancers and insulators, making the final determination of causative non-coding variants a great challenge compared to coding variants. Recent efforts to characterize non-coding sequences have generated a huge amount of data, identified many regulatory elements, and clarified general aspects of gene regulation[8]. Nevertheless, a substantial gap remains between the outcomes of these experiments and a detailed understanding of non-coding function. As a consequence, the development of machine-learning methods that attempt to more precisely predict regulatory function by jointly considering several functional annotations has emerged as an active, fast-moving area of research. These computational algorithms allow the analysis of thousands of variants at the same time, which would be impossible to systematically evaluate in the laboratory. However, as shown by di Iulio *et al.,*[158] the information integrated by each computational method is very diverse, resulting in a proportion of variants being uniquely captured by each method. This observation indicate that the information provided by every single computational method is complementary and needs to be combined to increase the detection strength of non-coding variants with functional effects.

In this thesis, we used two different approaches to identify non-coding variants with potential functional effects in BrS: (i) an approach focused on the identification of genetic variants that might be affecting the binding of TFs; and (ii), an approach focused on the identification of genetic variants significantly enriched in BrS individuals compared to healthy-aging individuals (Wellderly). In both approaches we took advantage of recently published computational methods to predict the functional impact of non-coding variants.

3.1. Effects of BrS variants in transcription factor binding

Here, we sought to identify non-coding variants from our cohort of BrS individuals that are affecting the binding of cardiac TFs, focusing on GATA4, GATA6 and NKX2.5. To achieve this goal, we performed ChIP-seq experiments for these three TFs in an iPS-derived cardiomyocyte

cell model developed in Farah Sheik's laboratory at UCSD. However, our ChIP-seqs for these TFs were of low quality and could not be used for this purpose. There are several reasons that could have influenced our ChIP-seq performance. First of all, it should be taken into account that, as it is the case for other cardiac TFs, the antibodies used in these experiments had not been validated for genome-wide ChIP-seq and there were no other ChIP-grade antibodies available for these TFs at the time we performed the experiments. Second, the antibodies were shipped from Girona (Catalonia, Spain) to San Diego (California, US), where we performed the ChIP-seq experiments. Antibodies have to be stored at 4ºC, but due to the shipping they were exposed to changes in the environmental temperature previous to their arrival to San Diego, which might have compromised the efficiency of the antibodies in immunoprecipitating their targets. Finally, during the ChIP-seq protocol we experienced some troubles when shearing the chromatin with the Bioruptor: after resuspension of the cell pellets in lysis buffer, pellets became very glutinous, and we had to apply a remarkably high number of sonication cycles to disaggregate them. These high number of sonication cycles (40 – 50) might have resulted in small DNA fragments, which are removed during ChIP-seq library preparation.

Since our ChIP-seq data for cardiac TFs could not be used, we took advantage of the already available information of CTCF binding from HCMs. We postulated that genetic variants affecting the binding of CTCF at boundary elements could disrupt the insulation of BrS-associated TADs and result in an aberrant expression of BrS-associated genes. We overlapped the 5,349 non-coding variants found in the Regulome-seq regions 89 BrS individuals with CTCF binding sites and identified 59 CTCF-overlapping variants. Then, we obtained binding predictions for these CTCF-overlapping variants using **DeepBind**[159], which is a machine learning-based algorithm that analyses sequence specificities of TFs genome-wide to predict the effects of non-coding variants in TF binding. The utilization of DeepBind to predict CTCF binding effects was possible because DeepBind CTCF model (generated from more than 30 different cell lines) was already available.

We also performed luciferase reporter assays to experimentally test the effect of the 59 CTCF-overlapping variants on CTCF binding. The results obtained in these assays are in agreement with DeepBind results when analyzing the binding effects depending on the location of the variant within the CTCF motif. These observations demonstrate that DeepBind is a powerful tool to predict, in a high-throughput manner, the effects of non-coding variants in CTCF binding. Moreover, the results also show that DeepBind is able to prioritize non-coding variants affecting the core motif from distant variants less likely to affect CTCF binding (Results **Figures 88** and **91**). The fact that DeepBind predicts increased effects in CTCF binding when

variants are affecting core nucleotides goes in accordance with these variants showing higher CADD scores, being ≥15 in some cases (median CADD score of 9.63 for core nucleotides versus a median CADD score of 5.21 for more distant variants).

Together, DeepBind predictions and luciferase assays allowed us to identify 21 CTCF-overlapping variants that are potentially affecting CTCF binding (Results **Table 29**). However, it is important to acknowledge that, even non-coding variants affecting TF binding have been associated to a broad range of diseases such as cancer, neurodevelopmental disorders and cardiac diseases[26], DeepBind predictions of TF binding alterations are not necessarily implying pathogenicity of the variants. In fact, Maurano *et al.,*[274] analyzed the genetic variation present in CTCF binding sites across a multi-generational pedigree and observed that individual TF binding sites are surprisingly robust to genetic variation and that the effects of genetic variants are buffered by striking context dependencies. Similarly, when we combined DeepBind predictions with CADD and CDTS, we also detected this tolerance to genetic variation of CTCF binding sites as they were found in high CDTS percentiles.

The premise that genetic variants affecting TF binding are not necessarily translated into phenotypic effects is also manifested in our results, since we found that 10 of the 21 CTCF-overlapping variants predicted to affect CTCF binding are equally frequent among BrS and Wellderly individuals (No-difference) and 1 variant is shared by both cohorts but more frequent among Wellderly individuals (Wellderly-enriched). From the remaining 10 CTCF-overlapping variants, 9 are BrS-specific–although only 2 are significant after applying the Pearsons's chi-squared test–and 1 is shared between BrS and Wellderly cohorts but is more frequent among BrS individuals (BrS-enriched; Results **Table 33**). The observation of such a small fraction of CTCF-overlapping variants being possibly related to BrS phenotype (2 BrS-specific significant variants and 1 BrS-enriched) could be explained by the small number of individuals sequenced. As mentioned earlier in the results' section, sequencing thousands of individuals could result in some of the low-frequency variants turning into highly-shared variants and, hence, being significant under permutations or Pearson's chi-squared test.

Another hypothesis in regard to these CTCF variants could be that alterations in CTCF binding at boundary elements are not the most recurrent cause of BrS pathogenesis–opposite to observations in other pathologies such as cancer[138,139]–, and is only applicable to a reduced number of BrS individuals. As a matter of fact, we identified two interesting cases with low-frequency CTCF-overlapping variants in BrS individuals (**Figures 98** and **99**). The first case corresponds to a CTCF binding site that contains two different heterozygous variants (a 1 bp deletion and a SNV) that are found together in the same BrS individual (**Figure 98**). Both variants

(rs764351689 and rs772963535, respectively) are individually predicted to decrease CTCF binding by DeepBind and luciferase assays. In addition, preliminary data from our lab suggests that when these variants are found together within the CTCF motif they further reduce CTCF binding affinity. This affected CTCF binding site is close to *CACNA2D* gene, in a region enriched in CTCF peaks, a common feature of boundary regions. *CACNA2D1* encodes the α2/δ-subunit 1 of the voltage-dependent L-type calcium channel (Ca_Vα2/δ), which regulates the current density and activation/inactivation kinetics of the Ca^{2+} channel. Loss-of-function genetic variants in the coding regions of this gene have been associated to BrS susceptibility. Indeed, Burashnikov *et al.,*[238] identified several mutations in the coding regions of *CACNA2D1* in different BrS individuals. Based on preliminary functional expression studies, they observed that double mutation in *CACNA2D1* reduces calcium currents, which has been already associated to electrical conduction defects of the heart. Similarly, we postulate that these two variants could be disrupting the boundary where they are embedded, decreasing *CACNA2D1* gene expression and causing similar effects to those observed by Burashnikov and colleagues. However, further analyses are required in order to determine the exact molecular mechanism by which the two variants could be affecting *CACNA2D1* expression.

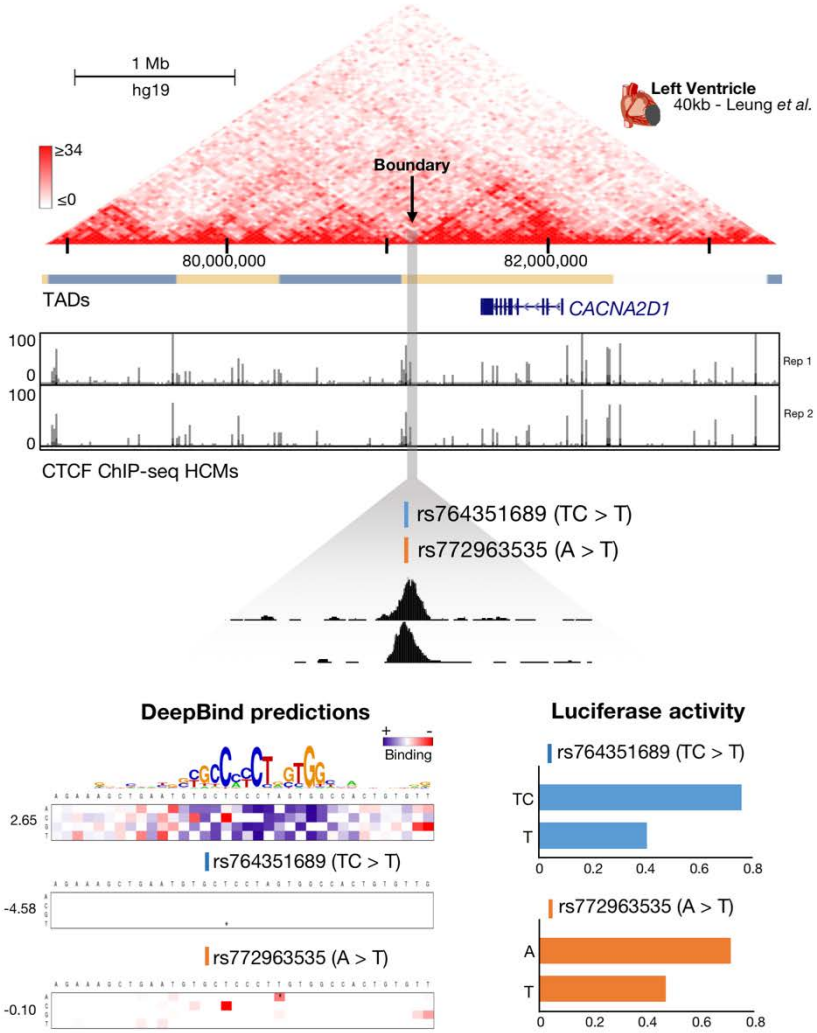


Figure 98. Two CTCF-overlapping variants nearby *CACNA2D1* identified in a single BrS patient predicted to reduce CTCF binding. Top: Hi-C matrix representing the chromatin contacts detected for *CACNA2D1* locus in left ventricle cells[245]. **Middle:** the two CTCF-overlapping variants (rs764351689 and rs772963535) with the reference and alternative alleles. The CTCF peak for the two HCM replicas is also shown. **Bottom:** DeepBind predictions of both variants as variation maps (left) and results from luciferase assays (right).

The second interesting case corresponds to a SNV found in two distinct BrS individuals that overlaps a single CTCF binding site (**Figure 99**). This SNV (rs6781889) is predicted to decrease CTCF binding both by DeepBind and luciferase assays. The affected CTCF binding site is close to the *SCN5A* gene, in a boundary region enriched in CTCF peaks. *SCN5A* encodes the α-subunit o the cardiac sodium channel ($Na_V1.5$), which plays a key role during the depolarization phase of the cardiac action potential. Loss-of-function variants at the coding regions of this gene are the most common cause of BrS[234]. However, recent findings also suggest that aberrant *SCN5A* gene expression may increase susceptibility to arrhythmogenic diseases. For example, Leoni *et al.,*[275] observed that low $Na_V1.5$ levels in heterozygous *Scn5a* +/– knockout mice

recapitulate cardiac conduction defects found in human individuals carrying disease-associated *SCN5A* mutations, and that the severity of these defects is directly correlated with levels of $Na_V1.5$ expression. Therefore, we hypothesize that this CTCF-overlapping SNV could disrupt *SCN5A*-TAD regulation, resulting in a reduced *SCN5A* gene expression that may be linked to sodium channel-related cardiac diseases such as BrS. Still, the molecular mechanism by which this SNV could be affecting *SCN5A* expression needs to be further inspected.

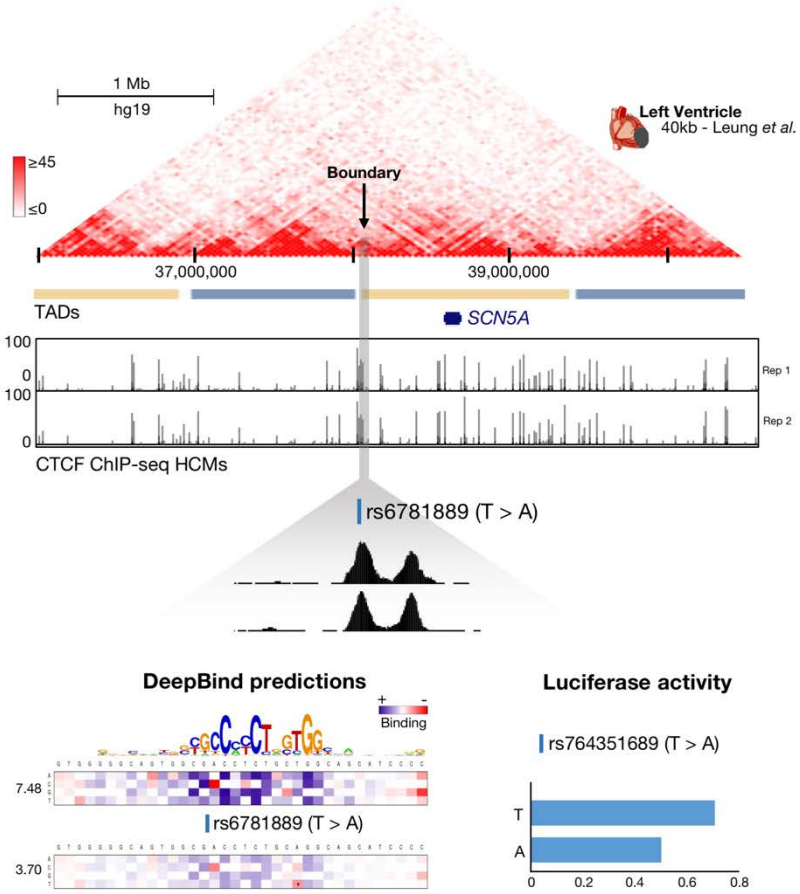


Figure 99. CTCF-overlapping SNV nearby *SCN5A* gene found in two BrS patients predicted to diminish CTCF binding. Top: Hi-C matrix representing the chromatin contacts detected for *SCN5A* locus in left ventricle cells[245]. **Middle:** the reported SNV (rs6781889) with the reference and alternative alleles. The CTCF peak for the two HCM replicas is also shown. **Bottom:** represented DeepBind predictions of as variation maps (left) and results from luciferase assays (right).

It is important to highlight that these 3 proposed CTCF-overlapping candidates have been identified through computational predictions and *in vitro* luciferase assays, and that the hypothesized mechanisms by which they can cause BrS need to be experimentally validated. The most comprehensive approach would be the introduction of these variants in the genome of iPS-derived cardiomyocytes using CRISPR/Cas9 technology, followed by RNA quantification of the corresponding gene (*SCN5A* or *CACNA2D1*) in wild-type and engineered

cardiomyocytes. RNA quantification of surrounding genes could also be performed to detect other genes with altered expression due to the presence of these variants. The same approach should be applied to iPS-derived cardiomyocytes from the patients harboring the variant, but in this case CRISPR/Cas9 technology should be used to reverse the variant and restore the wild-type sequence. Finally, electrophysiological recordings should be performed to ensure that the altered expression patterns are translated into alterations in the electrical activity of the cardiomyocytes.

In summary, our Regulome-seq approach coupled with the DeepBind strategy has allowed us to identify 3 CTCF-overlapping variants that could be involved in BrS pathogenesis. Still, CTCF binding alterations do not appear to be a common mechanism of BrS. The remaining 7 CTCF-overlapping variants are located in CTCF binding sites that seem not related to TAD insulation. Therefore, their potential association to BrS may be related to a different mechanism.

Our results also show that DeepBind is a powerful algorithm to predict, in a high-throughput manner, the effects of variants in TF binding. We suspect that if variant prioritization based on TF binding alterations could have been applied to cardiac TFs such as GATA4, GATA6 and NKX2.5, we would have identified more powerful candidates to BrS than using a general TF as it is CTCF. However, DeepBind binding models for these cardiac TFs are not created. As a consequence, we would have to either: (i) train DeepBind for these TFs; (ii) find another machine-learning algorithm suitable for these TFs or, (iii) use a different computational approach such as Position Weight Matrixes, whose predictions are based in the frequency in which a variant occurs in the position assessed when performing multiple alignments of TF binding sequences obtained from ChIP-seq data.

Based on our results, we predict that in the near future algorithms similar to DeepBind will be improved and will facilitate the analysis of thousands of variants for their potential effect on TF binding.

3.2. BrS and Wellderly comparison

In this approach, we sought to identify non-coding variants significantly enriched in BrS individuals compared to a healthy cohort. The healthy cohort used in this study was the so-called Wellderly cohort, consisting of 1,354 individuals that were >80 years old with no chronic diseases and not taking chronic medications at the moment of recruitment. From these 1,354 individuals, we could only use the data from those individuals that had been sequenced using an Illumina platform. Thanks to our Material Transfer Agreement-collaboration with Dr. Eric

Topol's group, we obtained the genetic information found in the Regulome-seq regions of 200 Wellderly individuals.

The comparison with the Wellderly cohort, together with the statistical analysis applied (**100 permutations** and **Pearson's chi-squared test**), led us to classify the variants as BrS-specific, BrS-enriched and No-difference. At the same time, the selection of statistically significant variants allowed us to increase our predictive power, since some of the differing allelic frequencies observed between both cohorts–including finding cohort-specific variants–could be found by chance due to the small sample size in both cohorts.

Considering that the BrS-specific and BrS-enriched variants will be the most likely candidates to be involved in the BrS phenotype, this strategy allowed us to reduce the 5,349 non-coding variants to a few candidates that could be related to the BrS phenotype (**537 variants**).

After the detection of BrS-specific and BrS-enriched variants, we combined the predictions obtained from two different computational algorithms to identify potentially pathogenic variants: CADD[150] and CDTS[158]. **CADD** integrates genomic and epigenomic annotations in a single score of deleteriousness, which strongly correlates with both molecular functionality and pathogenicity. Variants are considered to be deleterious when their CADD score is ≥ 15. **CDTS** computes the tolerance to genetic variation of all genomic regions in the context of surrounding sequences and divides them in percentiles of tolerance. Variants within the 1st CDTS percentile are considered to be more likely to be pathogenic.

When considering all variants together (BrS-specific, BrS-enriched and No-difference) we observed that the median CADD score is around 4.5, although a few variants with CADD scores greater than 15 are also found in all categories. This median score is in agreement with the results reported by the authors of CADD. In particular, they observed that variants within regulatory regions had a median score of 5, while variants in coding regions had a median CADD score exceeding 15, reaching 37 for stop gain variants. The lower CADD scores observed in non-coding regions could be explained by the fact that these regions are less constrained and more tolerant to genetic variation than coding regions. This assumption is supported by the results of di Iulio and colleagues[158] . They represented the cumulative territory fraction covered by non-coding regions in all CDTS percentiles and observed that, except for promoters and regulatory regions related to essential genes, non-coding regions are found at higher CDTS percentiles (regions more tolerant to genetic variation). Following the same line of evidence, our results show that most of the non-coding variants in our study tend to be found at high CDTS percentiles (median of 44).

CADD and CDTS, independently, revealed that BrS-specific variants are closer to the pathogenicity thresholds (CADD ≥ 15 and 1st CDTS percentile), followed by BrS-enriched variants and to a less extent by No-difference variants (Results **Figures 95** and **96**). Interestingly, this difference is even more pronounced when analyzing the distribution of scores for each locus sequenced. Notably, we observe that, when variants significantly enriched in our BrS cohort (BrS-specific and BrS-enriched) are found within the *SCN5A* locus, they are classified as more pathogenic than for other loci. These results are in agreement with the fact that *SCN5A* is the major gene contributing to BrS pathogenesis[234], and the only gene presenting sufficient causality evidence to be used for clinical diagnosis of BrS[282]. The results also support the hypothesis that not only coding variants at *SCN5A* would be involved in BrS but also non-coding variants affecting *SCN5A* expression might have a role as well.

3.3. Combination of CADD and CDTS pathogenicity thresholds

The combination of the pathogenicity thresholds (CADD ≥ 15 and 1st CDTS percentile) applied to BrS-specific and BrS-enriched variants led to the identification of 10 candidate variants to BrS pathogenesis (Results **Table 32**).

Among the 10 candidates identified, an interesting example is a SNV found in the *SCN3B* promoter region (**Figure 100**). This SNV (rs72552195) was detected in only one BrS individual and it was also found in the Wellderly cohort, although with less frequency (BrS-enriched). *SCN3B* encodes the regulatory β3-subunit of the cardiac sodium channel ($Na_V\beta3$). Loss-of-function variants in the coding regions of this gene have been related to BrS. In particular, Hu *et al.,*[236] identified a mutation in the exon 1 of *SCN3B* in a BrS individual that results in a defective transport of the $Na_V1.5$ protein to the cell membrane, leading to a reduction in sodium channel current and clinical manifestation of a BrS phenotype. Similarly, we propose that our reported SNV might be affecting the *SCN3B* promoter activity, decreasing *SCN3B* gene expression levels and causing similar effects to those observed by Hu and colleagues. However, further investigations are required to the exact molecular mechanism by which the SNV could be affecting *SCN3B* expression. It is our hypothesis that this SNV is possibly disrupting the binding of a TF involved in *SCN3B* transcriptional regulation. Once unraveled the molecular mechanism, the effects of the SNV and their association to BrS should be demonstrated in functional studies such as the genetic engineering of iPS-derived cardiomyocytes as described for the two CTCF candidate variants.

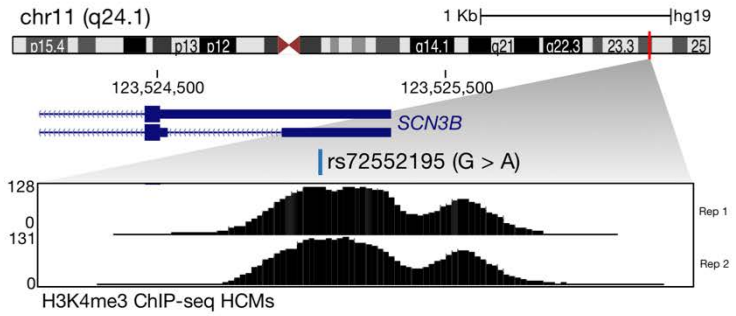


Figure 100. SNV at *SCN3B* promoter identified in one patient proposed as possible candidate for BrS pathogenesis. UCSC genome browser track for the *SCN3B* gene showing the genomic position of the reported SNV (rs72552195) on top of the active promoter-associated H3K4me3 from HCMs. Reference and alternative alleles are also shown.

A third interesting example is a 6 bp insertion in the *MYD88* promoter region, embedded in the *SCN5A* locus (**Figure 101**). This insertion (rs545602132) is shared by four BrS individuals and is BrS-specific (not found in Welldelry individuals). *MYD88* encodes a cytosolic adapter protein that plays an essential role as signaling transducer in the interleukin-1 and Toll-like receptor signaling pathways[276]. In addition to its importance for the innate immune response, *MYD88* has also been related to the maintenance of physiological function in the adult heart[277–279]. In a recent study, Chen and colleagues[277], generated a mouse strain with cardiac overexpression of MYD88 protein. Interestingly, their transgenic mice presented structural normal hearts but showed decreased contractility of the left ventricle, a phenotype also observed for some electrical diseases such as BrS and LQTS[280,281]. We therefore propose that our reported insertion might be affecting *MYD88* promoter activity, increasing *MYD88* gene expression and causing similar effects than those observed by Chen and colleagues. However, we need to perform further analysis to define the molecular mechanism by which the insertion could be affecting *MYD88* expression. We suggest that this insertion creating a new binding site for a TF involved in its transcriptional regulation. The effects of the insertion and their association to BrS could be demonstrated in functional studies such as the genetic engineering of iPS-derived cardiomyocytes as described previously.

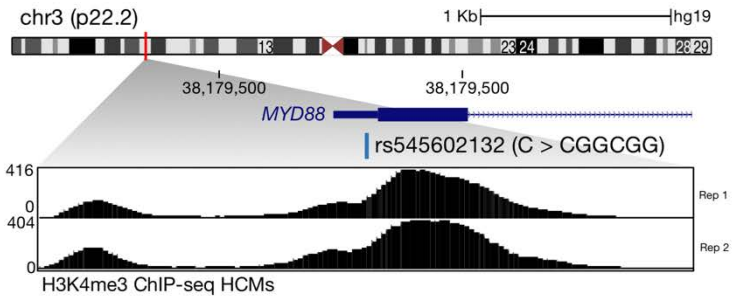


Figure 101. Insertion at *MYD88* promoter identified in four patients proposed as possible candidate for BrS pathogenesis. UCSC genome browser track for the *MYD88* gene showing the genomic position of the reported insertion (rs545602132) on top of the active promoter-associated H3K4me3 from HCMs. Reference and alternative alleles are also shown.

The remaining variants are located in promoters of genes *a priori* not related to cardiovascular diseases. Additional inspection of these variants needs to be done to establish the molecular link between them and alterations in the cardiac electrical activity. Of note, some of these variants might appear as enriched in BrS individuals due to the exceptionality of the super-healthy Wellderly cohort, making their absence a characteristic feature of Wellderly phenotype instead of their presence a characteristic feature of BrS phenotype.

In summary, the comparison with the Wellderly cohort highlights the importance of comparing the genetic variation found in a diseased cohort to that found in a healthy cohort to facilitate the association of genetic variants to diseased phenotypes. Genetic profiling from Wellderly and BrS individuals was performed with two different techniques (WGS for Wellderly and targeted sequencing for BrS). Therefore, we had to redefine the coverage filters to be applied to the Wellderly variant call set to make it comparable with our BrS call set. Application of the Regulome-seq approach to 89 healthy individuals in parallel to the BrS cohort would have facilitated data analysis. However, most current studies also do not include healthy patients and rely on data from public repositories such as the 1000 Genomes Project. Even this limitation, scoring of BrS-specific and BrS-enriched variants with CADD and CDTS shows that non-coding variants at *SCN5A* regulatory regions are more related to BrS pathogenesis that non-coding variants at regulatory regions of other BrS-associated genes. We propose 10 candidate variants, some of which appeared more easily related to BrS pathogenesis. Still, we are aware that we require further functional studies to finally associate them to BrS.

4. Further considerations

In this thesis, we have discussed two different strategies that led to the identification of non-coding variants that could be related to the BrS phenotype. However, it is also important to acknowledge the growing evidence suggesting that BrS may not be always caused by a single pathogenic variant but rather by the presence of multiple susceptibility variants acting synergistically through one or more mechanistic pathways[283]. In fact, a recent GWAS published by Bezzina *et al.,*[284] identified three loci to be associated to BrS (rs9388451 in proximity to *HEY2* gene, rs10428132 close to the *SCN10A* gene, and rs11708996 close to *SCN5A* gene). Interestingly, when assessing the cumulative effect of the three loci on susceptibility to BrS, the authors found that disease risk consistently increased with the number of risk alleles, with patients carrying more than four of the risk alleles presenting a stronger association to BrS compared with patients carrying less than two risk alleles[284]. Bezzina's results also fit with the findings of Probst *et al.,*[285] who studied 13 large families with a putative *SCN5A* pathogenic variant, and found that in some patients the presence of the variant did not co-segregate with their BrS phenotype. Together, these results indicate that the genetic background of each individual (i.e. the presence of different genetic variants) may confer disparate susceptibilities to BrS. Based on these observations, we cannot discard the possibility that combinations of non-coding variants disrupting the expression of BrS-associated genes as another underlying mechanism to explain some BrS cases. To address this issue, we aim to apply the same analysis used to identify the proposed BrS-candidate variants but this time considering combinations of variants (or haplotypes) rather than single variants.

VI. Conclusions

1. In the present thesis, we have developed a highly cost-effective approach, referred to as Regulome-seq. We used this strategy to identify 1,293 *cis*-regulatory regions related to six BrS-associated genes and sequence them in a cohort of 89 BrS individuals.

2. The presence of the rs6801975 polymorphism (previously associated to cardiac conduction defects) in one of the Regulome-seq regions indicate that our strategy to identify *cis*-regulatory regions linked to BrS-associated genes is effective in selecting disease-associated regulatory regions.

3. The good quality of the sequencing together with the high and homogeneously distributed coverage presented by all targeted regions indicates that both the design of the capturing probes using DesignStudio™ (Illumina®) and the sequencing conditions were efficiently optimized.

4. Our variant discovery pipeline, coupled with the filtering conditions established using the Coriell NA12249, led us to report a high-confident subset of 5,349 variants (4,837 SNVs and 512 indels) present in the Regulome-seq regions of the 89 BrS individuals with a 99% precision.

5. Approximately 33% of the variants identified are private to each BrS individual, while the remaining 67% are shared by 2 or more individuals. The prevalence of SNVs is higher among private variants, while the prevalence of indels is higher among the most shared variants.

6. Using the available information of CTCF binding on HCMs, we identified 59 variants overlapping CTCF binding sites. According to the machine-learning algorithm DeepBind, 43 were predicted to increase CTCF binding, 14 were predicted to decrease CTCF binding and 2 were predicted to have no effects on CTCF binding.

7. We have developed a luciferase reporter assay to evaluate the effect of genetic variants on CTCF binding. Results obtained with this assay resemble DeepBind predictions and suggest that DeepBind is a valid approach to prioritize non-coding variants affecting TF binding in a high-throughput manner. We identified 21 CTCF-overlapping variants that were predicted to have the same binding effects by both DeepBind predictions and luciferase activity.

8. Comparison of the sequencing data from the BrS cohort with the healthy-aging cohort (Wellderly), in combination with the statistical analysis applied, led to the identification of 537 variants significantly enriched in BrS individuals (201 BrS-specific and 336 BrS-enriched). This significant reduction in the number of possible candidate variants to BrS highlights the importance of comparing a diseased cohort to a healthy cohort.

9. Scoring of BrS-specific, BrS-enriched and No-difference variants with CADD and CDTS show that variants significantly enriched in BrS individuals are more likely to be pathogenic. This observation is even more pronounced when analyzing variants found in the *SCN5A* locus, suggesting that non-coding variants affecting *SCN5A* expression might be also involved in BrS pathogenesis.

10. After applying our pathogenicity threshold (combination of CADD scores ≥15 and 1st CDTS percentile), we propose a total of 10 non-coding variants as possible candidates to be associated to BrS pathogenesis.

11. All candidate variants proposed in this thesis, either with DeepBind or CADD/CDTS, are based on computational predictions. Therefore, further functional studies will be required to associate them to BrS pathogenesis.

VII. Bibliography

1. Miescher, F. Die Histochemischen und physiologischen. (Vogel, 1897).
2. Watson, J. D. & Crick, F. H. C. Molecular structure of nucleic acids: a structure for deoxyribose nucleic Acid. *Nature* **171,** 737–738 (1953).
3. Alberts, B. *et al.* Molecular biology of the cell. (Garland Science, 2008).
4. Ecker, J. R. *et al.* Genomics: ENCODE explained. *Nature* **489,** 52–55 (2012).
5. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
6. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
7. Parker, J. in Encyclopedia of Genetics (eds. Brenner, S. & Miller, J. H.) 401–402 (Academic Press, 2001).
8. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
9. Mattick, J. S. & Makunin, I. V. Non-coding RNA. *Hum. Mol. Genet.* **15,** R17–R29 (2006).
10. Natoli, G. & Andrau, J-C. Noncoding transcription at enhancers: general principles and functional models. *Annu. Rev. Genet.* **46,** 1–19 (2012).
11. Vanin, E. F. Processed pseudogenes: characteristics and evolution. *Annu. Rev. Genet.* **19,** 253–272 (1985).
12. Sasidharan, R. & Gerstein, M. Genomics: Protein fossils live on as RNA. *Nature* **453,** 729–731 (2008).
13. Tutar, Y. Pseudogenes. *Comp. Funct. Genomics* **2012,** 1–4 (2012).
14. Stoltzfus, A. in Encyclopedia of Genetics (eds. Brenner, S. & Miller, J. H.) 1052–1053 (Academic Press, 2001).
15. Jo, B-S. & Choi, S. S. Introns: The Functional Benefits of Introns in Genomes. *Genomics Inform.* **13,** 112–118 (2015).
16. Chabot, B. & Shkreta, L. Defective control of pre-messenger RNA splicing in human disease. *J. Cell Biol.* **212,** 13–27 (2016).
17. Anna, A. & Monika, G. Splicing mutations in human genetic disorders: examples, detection, and confirmation. *J. Appl. Genet.* 1–16 (2018).
18. Biscotti, M. A. *et al.* Repetitive DNA in eukaryotic genomes. **23,** 415–420 (2015).
19. Griffiths, A. J. F. *et al*. in An Introduction to genetic analysis - Molecular nature of transposable elements in eukaryotes (W. H. Freeman, 2003).
20. Shapiro, J. A. & Sternberg, R. von. Why repetitive DNA is essential to genome function. *Biol. Rev.* **80,** 227–250 (2005).
21. Jones, D. L. *et al.* Promoter architecture dictates cell-to-cell variability in gene expression. *Science* **346,** 1533–1536 (2014).
22. Nord, A. S. *et al.* Rapid and pervasive changes in genome-wide enhancer usage during mammalian development. *Cell* **155,** 1521–1531 (2013).
23. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507,** 455–461 (2014).
24. Wittkopp, P. J. & Kalay, G. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat. Rev. Genet.* **13,** 59–69 (2012).
25. Cookson, W. *et al.* Mapping complex disease traits with global gene expression. *Nat. Rev. Genet.* **10,** 184–194 (2009).
26. Lee, T. I. & Young, R. A. Transcriptional Regulation and Its Misregulation in Disease. *Cell* **152,** 1237–1251 (2013).
27. Keung, A. J. *et al*. Chromatin regulation at the frontier of synthetic biology. *Nat. Rev. Genet.* **16,** 159–171 (2015).

28. Dixon, J. R. *et al.* Chromatin Domains: The Unit of Chromosome Organization. *Mol. Cell* **62,** 668–680 (2016).
29. O'Connor, T. *et al.* CisMapper: predicting regulatory interactions from transcription factor ChIP-seq data. *Nucleic Acids Res.* **45,** e19–e19 (2017).
30. Smale, S. T. & Kadonaga, J. T. The RNA polymerase II core promoter. *Annu. Rev. Biochem.* **72,** 449–479 (2003).
31. Gershenzon, N. I. & Ioshikhes, I. P. Synergy of human Pol II core promoter elements revealed by statistical sequence analysis. *Bioinformatics* **21,** 1295–1300 (2005).
32. Morris, J. R. *et al.* Enhancer choice in cis and in trans in Drosophila melanogaster: role of the promoter. *Genetics* **167,** 1739–1747 (2004).
33. Maston, G. A. *et al.* Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.* **7,** 29–59 (2006).
34. Bulger, M. & Groudine, M. Enhancers: the abundance and function of regulatory sequences beyond promoters. *Dev. Biol.* **339,** 250–257 (2010).
35. Atchison, M. L. Enhancers: mechanisms of action and cell specificity. *Annu. Rev. Cell Biol.* **4,** 127–153 (1988).
36. Mora, A. *et al.* In the loop: promoter–enhancer interactions and bioinformatics. *Brief. Bioinform.* **17,** 980–995 (2016).
37. Sagai, T. *et al*. Elimination of a long-range cis-regulatory module causes complete loss of limb-specific Shh expression and truncation of the mouse limb. *Development* **132,** 797–803 (2005).
38. Shim, S. *et al*. Cis-regulatory control of corticospinal system development and evolution. *Nature* **486,** 74–79 (2012).
39. Frankel, N. *et al.* Phenotypic robustness conferred by apparently redundant transcriptional enhancers. *Nature* **466,** 490–493 (2010).
40. Perry, M. W. *et al.* Shadow enhancers foster robustness of Drosophila gastrulation. *Curr. Biol. CB* **20,** 1562–1567 (2010).
41. Osterwalder, M. *et al.* Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature* **554**, 239–243 (2018).
42. Dickel, D. E. *et al.* Ultraconserved enhancers are required for normal development. *Cell* **172,** 491–499.e15 (2018).
43. Oldridge, D. A. *et al.* Genetic predisposition to neuroblastoma mediated by a *LMO1* super-enhancer polymorphism. *Nature* **528,** 418–421 (2015).
44. van den Boogaard, M. *et al.* Genetic variation in T-box binding element functionally affects *SCN5A/SCN10A* enhancer. *J. Clin. Invest.* **122,** 2519–2530 (2012).
45. Pott, S. & Lieb, J. D. What are super-enhancers? *Nat. Publ. Gr.* **47,** 8–12 (2015).
46. Whyte, W. A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153,** 307–319 (2013).
47. Latchman, D. S. Transcription factors: An overview. *Int. J. Biochem. Cell Biol.* **29,** 1305–1312 (1997).
48. Orphanides, G. *et al.* The general transcription factors of RNA polymerase II. *Genes Dev.* **10,** 2657–2683 (1996).
49. Levine, M. Paused RNA polymerase II as a developmental checkpoint. *Cell* **145,** 502–511 (2011).
50. Knuesel, M. T. *et al.* The human CDK8 subcomplex is a molecular switch that controls Mediator coactivator function. *Genes Dev.* **23,** 439–451 (2009).
51. Reynolds, N. *et al.* Transcriptional repressors: multifaceted regulators of gene expression. *Development* **140,** 505–512 (2013).

52. Deckert, J. & Struhl, K. Histone acetylation at promoters is differentially affected by specific activators and repressors. *Mol. Cell. Biol.* **21,** 2726–2735 (2001).
53. Kummerfeld, S. K. & Teichmann, S. A. DBD: a transcription factor prediction database. *Nucleic Acids Res.* **34,** D74–D81 (2006).
54. Stegmaier, P. *et al.* Systematic DNA-binding domain classification of transcription factors. *Genome Inform.* **15,** 276–286 (2004).
55. Gonzalez, D. H. in Plant Transcription Factors 3–11 (Academic Press, 2016).
56. Wingender, E. Criteria for an updated classification of human transcription factor DNA-binding domains. *J. Bioinform. Comput. Biol.* **11,** 1340007 (2013).
57. Sandelin, A. *et al.* JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* **32,** D91–D94 (2004).
58. Machanick, P. & Bailey, T. L. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* **27,** 1696–1697 (2011).
59. Carrol, S.B. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* **134,** 25–36 (2008).
60. Wray, G. A. The evolutionary significance of cis-regulatory mutations. *Nat. Rev. Genet.* **8,** 206–216 (2007).
61. Inukai, S. *et al.* Transcription factor–DNA binding: beyond binding site motifs. *Curr. Opin. Genet. Dev.* **43,** 110–119 (2017).
62. Morales, V. *et al.* Chromatin structure and dynamics: functional implications. *Biochimie* **83,** 1029–1039 (2001).
63. Jeon, K. W. International review of cytology: a survey of cell biology. (Elsevier, 2005).
64. Kouzarides, T. Chromatin modifications and their function. *Cell* **128,** 693–705 (2007).
65. Hayes, J. J. & Wolffe, A. P. The interaction of transcription factors with nucleosomal DNA. *BioEssays News Rev. Mol. Cell. Dev. Biol.* **14,** 597–603 (1992).
66. Barski, A. *et al.* High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell* **129,** 823–837 (2007).
67. Bassett, S. A. & Barnett, M. P. G. The role of dietary histone deacetylases (HDACs) inhibitors in health and disease. *Nutrients* **6,** 4273–4301 (2014).
68. Musselman, C. A. *et al.* Perceiving the epigenetic landscape through histone readers. *Nat. Struct. Mol. Biol.* **19,** 1218–1227 (2012).
69. Strahl, B. D. & Allis, C. D. The language of covalent histone modifications. *Nature* **403,** 41–45 (2000).
70. Petty, E. & Pillus, L. Balancing chromatin remodeling and histone modifications in transcription. *Trends Genet.* **29,** 621–629 (2013).
71. Wu, H. & Zhang, Y. Mechanisms and functions of Tet protein-mediated 5-methylcytosine oxidation. *Genes Dev.* **25,** 2436–2452 (2011).
72. Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. *Genes Dev.* **25,** 1010–1022 (2011).
73. Geahigan, K. B. *et al.* The dynamic impact of CpG methylation in DNA. *Biochemistry* **39,** 4939–4946 (2000).
74. Derreumaux, S. *et al.* Impact of CpG methylation on structure, dynamics and solvation of cAMP DNA responsive element. *Nucleic Acids Res.* **29,** 2314–2326 (2001).
75. Yu, M. & Ren, B. The Three-Dimensional Organization of Mammalian Genomes. *Annu. Rev. Cell Dev. Biol.* **33,** 265–289 (2017).
76. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518,** 317–30 (2015).

77. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326,** 289–293 (2009).
78. Ji, X. *et al.* 3D chromosome regulatory landscape of human pluripotent cells. *Cell Stem Cell* **18,** 262–275 (2016).
79. Phillips-Cremins, J. E. *et al.* Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell* **153,** 1281–1295 (2013).
80. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485,** 376–380 (2012).
81. Nakahashi, H. *et al.* A genome-wide map of CTCF multivalency redefines the CTCF code. *Cell Rep.* **3,** 1678–1689 (2013).
82. Vietri Rudan, M. *et al.* Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell Rep.* **10,** 1297–1309 (2015).
83. Peters, J-M., *et al.* The cohesin complex and its roles in chromosome biology. *Genes Dev.* **22,** 3089–3114 (2008).
84. Sanborn, A. L. *et al.* Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl. Acad. Sci.* **112,** E6456–E6465 (2015).
85. Dixon, J. R. *et al.* Chromatin architecture reorganization during stem cell differentiation. *Nature* **518,** 331–336 (2015).
86. Schmitt, A. D. *et al.* A compendium of chromatin contact maps reveals spatially active regions in the human genome. *Cell Rep* **17,** 2042–2059 (2016).
87. Ryba, T. *et al.* Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Res.* **20,** 761–770 (2010).
88. Guelen, L. *et al.* Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature*. **453 (7179),** 948–51 (2008).
89. Cremer, T. & Cremer, C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat. Rev. Genet.* **2,** 292–301 (2001).
90. Takizawa, T. *et al.* The Meaning of Gene Positioning. *Cell* **135,** 9–13 (2008).
91. Grasser, F. *et al.* Replication-timing-correlated spatial chromatin arrangements in cancer and in primate interphase nuclei. *J. Cell Sci.* **121,** 1876–1886 (2008).
92. Natarajan, A. *et al.* Predicting cell-type–specific gene expression from regions of open chromatin. *Genome Res.* **22,** 1711–1722 (2012).
93. Qiu, Z. *et al.* Identification of regulatory DNA elements using genome-wide mapping of DNase I hypersensitive sites during tomato fruit development. *Mol. Plant* **9,** 1168–1182 (2016).
94. Wilken, M. S. *et al.* DNase I hypersensitivity analysis of the mouse brain and retina identifies region-specific regulatory elements. *Epigenetics Chromatin* **8,** 8 (2015).
95. Boyle, A. P. *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132,** 311–322 (2008).
96. Song, L. & Crawford, G. E. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb. Protoc.***121,** 1876–1886 (2010).
97. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38,** 576–589 (2010).
98. Gilmour, D. S. & Lis, J. T. Detecting protein-DNA interactions in vivo: distribution of RNA polymerase on specific bacterial genes. *Proc. Natl. Acad. Sci. U. S. A.* **81,** 4275–4279 (1984).
99. Hon, G. C. *et al.* Predictive chromatin signatures in the mammalian genome. *Hum. Mol. Genet.* **18,** R195-201 (2009).

100. Pan, G. *et al.* Whole-genome analysis of histone H3 lysine 4 and lysine 27 methylation in human embryonic stem cells. *Cell Stem Cell* **1,** 299–312 (2007).
101. Zentner, G. E. *et al.* Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. *Genome Res.* **21,** 1273–83 (2011).
102. Schmitt, A. D. *et al.* Genome-wide mapping and analysis of chromosome architecture. *Nat. Rev. Mol. Cell Biol.* **17,** 743–755 (2016).
103. Zhao, Z. *et al.* Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat. Genet.* **38,** 1341–1347 (2006).
104. Spencer, D. H. *et al.* in Clinical Genomics (eds. Kulkarni, S. & Pfeifer, J.) 109–127 (Academic Press, 2015).
105. Mills, R. E. *et al.* An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res.* **16,** 1182–1190 (2006).
106. Girirajan, S. *et al.* Human copy number variation and complex genetic disease. *Annu. Rev. Genet.* **45,** 203–226 (2011).
107. The 1000 Genomes projec Consortium. A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
108. Lin, M. *et al.* Effects of short indels on protein structure and function in human genomes. *Sci. Rep.* **7,** 9313 (2017).
109. Sehn, J. K. in Clinical Genomics (eds. Kulkarni, S. & Pfeifer, J.) 129–150 (Academic Press, 2015).
110. Escaramís, G. *et al.* A decade of structural variants: description, history and methods to detect structural variation. *Brief. Funct. Genomics* **14,** 305–314 (2015).
111. De Bont, R. & van Larebeke, N. Endogenous DNA damage in humans: a review of quantitative data. *Mutagenesis* **19,** 169–185 (2004).
112. Griffiths, A. J. F. *et al.* in An introduction to genetic analysis - Induced mutations (W. H. Freeman, 2000).
113. Hang, B. Formation and Repair of Tobacco Carcinogen-Derived Bulky DNA Adducts. *Journal of Nucleic Acids.* **2010,** 1–29 (2010).
114. Kuefner, M. A. *et al.* Radiation Induced DNA Double-Strand Breaks in Radiology. *RoFo Fortschritte Auf Dem Gebiete Der Rontgenstrahlen Und Der Nukl.* **187,** 872–878 (2015).
115. Dizdaroglu, M. *et al.* Letter: Strand breaks and sugar release by gamma-irradiation of DNA in aqueous solution. *J. Am. Chem. Soc.* **97,** 2277–2278 (1975).
116. Khan Academy Courses. Available at: https://www.khanacademy.org/science/biology/dna-as-the-genetic-material/.
117. Berg, J. M. *et al.* in Biochemistry - DNA Polymerases Require a Template and a Primer (W. H. Freeman, 2002).
118. Mathews, L. A. *et al.* DNA Repair of Cancer Stem Cells. (Springer Netherlands, 2013).
119. Cooper, G. M. in The Cell - DNA Repair (Sinauer Associates, 2000).
120. Lupski, J. R. *et al.* Clan genomics and the complex architecture of human disease. *Cell* **147,** 32–43 (2011).
121. Schork, N. J. e*t al.* Common vs. rare allele hypotheses for complex diseases. *Curr. Opin. Genet. Dev.* **19,** 212–219 (2009).
122. Zuk, O., *et al.* The mystery of missing heritability: genetic interactions create phantom heritability. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 1193–1198 (2012).
123. The International HapMap Consortium *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449,** 851–861 (2007).

124. The International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* **467,** 52–58 (2010).
125. The International HapMap Consortium *et al*. A haplotype map of the human genome. *Nature* **437,** 1299–1320 (2005).
126. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).
127. Ahn, S-M. *et al.* The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Res.* **19,** 1622–1629 (2009).
128. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456,** 53–59 (2008).
129. Kim, J-I. *et al.* A highly annotated whole-genome sequence of a Korean individual. *Nature* **460,** 1011–1015 (2009).
130. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5,** e254 (2007).
131. Lupski, J. R. *et al.* Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *N. Engl. J. Med.* **362,** 1181–1191 (2010).
132. Schuster, S. C. *et al.* Complete Khoisan and Bantu genomes from southern Africa. *Nature* **463,** 943–947 (2010).
133. Wang, J. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456,** 60–65 (2008).
134. Wheeler, D. A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452,** 872–876 (2008).
135. Coventry, A. *et al.* Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nat. Commun.* **1,** 131 (2010).
136. Turner, D. J. *et al.* Germline rates of de novo meiotic deletions and duplications causing several genomic disorders. *Nat. Genet.* **40,** 90–95 (2008).
137. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461,** 747–753 (2009).
138. Katainen, R. *et al.* CTCF/cohesin-binding sites are frequently mutated in cancer. *Nat. Genet.* **47,** 818–821 (2015).
139. Hnisz, D. *et al.* Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science.* **351,** 1454–1458 (2016).
140. Wang, X. *et al.* A Polymorphic Antioxidant Response Element Links NRF2/sMAF Binding to Enhanced MAPT Expression and Reduced Risk of Parkinsonian Disorders. *Cell Rep.* **15,** 830–842 (2016).
141. Editorial. E pluribus unum. *Nat. Methods* **7,** 331 (2010).
142. Mardis, E. R. The impact of next-generation sequencing technology on genetics. *Trends Genet.* **24,** 133–141 (2008).
143. Metzker, M. L. Sequencing technologies - the next generation. *Nat. Rev. Genet.* **11,** 31–46 (2010).
144. Zhang, J. *et al.* The impact of next-generation sequencing on genomics. *J. Genet. Genomics* **38,** 95–109 (2011).
145. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (2016).
146. Zhang, F. & Lupski, J. R. Non-coding genetic variants in human disease. *Hum. Mol. Genet.* **24,** R102-110 (2015).
147. Knight, J. C. Approaches for establishing the function of regulatory genetic variants involved in disease. *Genome Med.* **6,** 92 (2014).

148. GTEx Consortium *et al*. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348,** 648–660 (2015).
149. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48,** 481–487 (2016).
150. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46,** 310–315 (2014).
151. Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15,** 901–913 (2005).
152. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15,** 1034–1050 (2005).
153. Pollard, K. S. *et al.* Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20,** 110–121 (2010).
154. Johnson, D. S. *et al.* Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316,** 1497–1502 (2007).
155. Grantham, R. Amino acid difference formula to help explain protein evolution. *Science* **185,** 862–864 (1974).
156. Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31,** 3812–3814 (2003).
157. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7,** 248–249 (2010).
158. di Iulio, J. *et al.* The human noncoding genome defined by genetic diversity. *Nat. Genet.* **50,** 333–337 (2018).
159. Alipanahi, B. *et al.* Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol* **33,** 831–838 (2015).
160. Hall, E. J. Guyton and Hall Textbook of Medical Physiology. (Elsevier, 2006).
161. OpenStax CNX. *Anatomy & Physiology.*Available at: https://cnx.org/contents/FPtK1zmh@9.1:Y5T_wVSC@4/Heart-Anatomy.
162. Science Learning Hub. *Label the heart.* Available at: https://www.sciencelearn.org.nz/labelling_interactives/1-label-the-heart.
163. Studyblue. *Heart Wall diagram at University of Cincinnati*. Available at: https://www.studyblue.com/notes/note/n/heart-wall-diagram/deck/9767307.
164. van Weerd, J. H. & Christoffels, V. M. The formation and function of the cardiac conduction system. *Development* **143,** 197–210 (2016).
165. Kloesel, B. *et al.* Cardiac Embryology and Molecular Mechanisms of Congenital Heart Disease: A Primer for Anesthesiologists. *Anesth. Analg.* **123,** 551 (2016).
166. Ostadal, B. & Dhalla, N. S. Cardiac Adaptations: Molecular Mechanisms. Advances in Biochemistry in Health and Disease. (Springer-Verlag, 2013).
167. Lindsey, S. E. *et al.* Mechanical regulation of cardiac development. *Front. Physiol.* **5,** (2014).
168. He, A. *et al.* Co-occupancy by multiple cardiac transcription factors identifies transcriptional enhancers active in heart. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 5632–5637 (2011).
169. Stanfel, M. N., *et al.* Regulation of organ development by the NKX-homeodomain factors: an NKX code. *Cell. Mol. Biol. (Noisy-le-grand).* **Suppl 51,** OL785-799 (2005).
170. Lyons, I. *et al.* Myogenic and morphogenetic defects in the heart tubes of murine embryos lacking the homeo box gene Nkx2-5. *Genes Dev.* **9,** 1654–1666 (1995).
171. Tanaka, M. *et al.* The cardiac homeobox gene Csx/Nkx2.5 lies genetically upstream of multiple genes essential for heart development. *Development* **126,** 1269–1280 (1999).

172. Schott, J. J. *et al.* Congenital heart disease caused by mutations in the transcription factor NKX2-5. *Science* **281,** 108–111 (1998).
173. Lin, Q. *et al.* Control of mouse cardiac morphogenesis and myogenesis by transcription factor MEF2C. *Science* **276,** 1404–1407 (1997).
174. Edmondson, D. G. *et al.* Mef2 gene expression marks the cardiac and skeletal muscle lineages during mouse embryogenesis. *Development* **120,** 1251–1263 (1994).
175. Biben, C. & Harvey, R. P. Homeodomain factor Nkx2-5 controls left/right asymmetric expression of bHLH gene eHand during murine heart development. *Genes Dev.* **11,** 1357–1369 (1997).
176. Srivastava, D. *et al.* Regulation of cardiac mesodermal and neural crest development by the bHLH transcription factor, dHAND. *Nat. Genet.* **16,** 154–160 (1997).
177. Bruneau, B. G. *et al.* A murine model of Holt-Oram syndrome defines roles of the T-box transcription factor Tbx5 in cardiogenesis and disease. *Cell* **106,** 709–721 (2001).
178. Zhou, P. *et al.* Regulation of GATA4 transcriptional activity in cardiovascular development and disease. *Curr. Top. Dev. Biol.* **100,** 143–169 (2012).
179. Liang, Q. & Molkentin, J. D. Divergent signaling pathways converge on GATA4 to regulate cardiac hypertrophic gene expression. *J. Mol. Cell. Cardiol.* **34,** 611–616 (2002).
180. van Berlo, J. H. *et al.* The transcription factor GATA-6 regulates pathological cardiac hypertrophy. *Circ. Res.* **107,** 1032–1040 (2010).
181. Arceci, R. J. *et al.* Mouse GATA-4: a retinoic acid-inducible GATA-binding transcription factor expressed in endodermally derived tissues and heart. *Mol. Cell. Biol.* **13,** 2235–2246 (1993).
182. Heikinheimo, M. *et al.* Localization of transcription factor GATA-4 to regions of the mouse embryo involved in cardiac development. *Dev. Biol.* **164,** 361–373 (1994).
183. Kuo, C. T. *et al.* GATA4 transcription factor is required for ventral morphogenesis and heart tube formation. *Genes Dev.* **11,** 1048–1060 (1997).
184. Molkentin, J. D. *et al.* Requirement of the transcription factor GATA4 for heart tube formation and ventral morphogenesis. *Genes Dev.* **11,** 1061–1072 (1997).
185. Belaguli, N. S. *et al.* Cardiac Tissue Enriched Factors Serum Response Factor and GATA-4 Are Mutual Coregulators. *Mol. Cell. Biol.* **20,** 7550–7558 (2000).
186. Sepulveda, J. L. *et al.* GATA-4 and Nkx-2.5 coactivate Nkx-2 DNA binding targets: role for regulating early cardiac gene expression. *Mol. Cell. Biol.* **18,** 3405–3415 (1998).
187. Rajagopal, S. K. *et al.* Spectrum of heart disease associated with murine and human GATA4 mutation. *J. Mol. Cell. Cardiol.* **43,** 677–685 (2007).
188. Tarradas, A. *et al.* Transcriptional regulation of the sodium channel gene (SCN5A) by GATA4 in human heart. *J. Mol. Cell. Cardiol.* **102,** 74–82 (2017).
189. Betts, J. G. *et al.* Anatomy & Physiology. (OpenStax, 2013).
190. Campbell, N. A. & Reece, J. B. Biology. (Editorial Medica Panamericana)
191. Munshi, N. V. Gene regulatory networks in cardiac conduction system development. *Circ. Res.* **110,** 1525–1537 (2012).
192. Sands, Z. *et al.* Voltage-gated ion channels. *Curr. Biol.* **15,** R44–R47 (2005).
193. Bezanilla, F. Voltage-gated ion channels. *IEEE Trans. Nanobioscience* **4,** 34–48 (2005).
194. Amin, A. S. *et al.* Cardiac ion channels in health and disease. *Hear. Rhythm* **7,** 117–126 (2010).
195. Theile, J. W. & Cummins, T. R. Recent developments regarding voltage-gated sodium channel blockers for the treatment of inherited and acquired neuropathic pain syndromes. *Front. Pharmacol.* **2,** (2011).
196. Nau, C. in Modern Anesthetics - Voltage-gated ion channels. 85–92 (Springer, 2008).

197. Catterall, W. A. *et al.* International Union of Pharmacology. XLVII. Nomenclature and structure-function relationships of voltage-gated sodium channels. *Pharmacol. Rev.* **57,** 397–409 (2005).
198. Cusdin, F. S. *et al.* Trafficking and cellular distribution of voltage-gated sodium channels. *Traffic* **9,** 17–26 (2008).
199. Brackenbury, W. J. & Isom, L. L. Na Channel β Subunits: Overachievers of the Ion Channel Family. *Front. Pharmacol.* **2,** 53 (2011).
200. Qin, N. *et al.* Molecular cloning and functional expression of the human sodium channel β1B subunit, a novel splicing variant of the β1 subunit. *Eur. J. Biochem.* **270,** 4762–4770
201. Waxman, S. G. Sodium channels, the electrogenisome and the electrogenistat: lessons and questions from the clinic. *J. Physiol.* **590,** 2601–2612 (2012).
202. Ruan, Y. *et al.* Sodium channel mutations and arrhythmias. *Nat. Rev. Cardiol.* **6,** 337–348 (2009).
203. Rook, M. B. *et al.* Biology of cardiac sodium channel Nav1.5 expression. *Cardiovasc. Res.* **93,** 12–23 (2012).
204. Catterall, W. A. Structure and function of voltage-gated ion channels. *Annu. Rev. Biochem.* **64,** 493–531 (1995).
205. Vacher, H. *et al.* Localization and targeting of voltage-dependent ion channels in mammalian central neurons. *Physiol. Rev.* **88,** 1407–1447 (2008).
206. Amin, A. S. *et al.* Cardiac sodium channelopathies. *Pflugers Arch. Eur. J. Physiol.* **460,** 223–237 (2010).
207. Catterall, W. A. Voltage-gated calcium channels. *Cold Spring Harb. Perspect. Biol.* **3,** (2011).
208. Grant, A. O. Cardiac Ion Channels. *Circ. Arrhythmia Electrophysiol.* **2,** 185–194 (2009).
209. Hille, B. Ion Channels of Excitable Membranes, Third Edition. (Sinauer Associates, 2001).
210. Pongs, O. Ins and outs of cardiac voltage-gated potassium channels. *Curr. Opin. Pharmacol.* **9,** 311–315 (2009).
211. González, C. *et al.* K(+) channels: function-structural overview. *Compr. Physiol.* **2,** 2087–2149 (2012).
212. Mangoni, M. E. & Nargeot, J. Genesis and Regulation of the Heart Automaticity. *Physiol. Rev.* **88,** 919–982 (2008).
213. Behere, S. P. & Weindling, S. N. Inherited arrhythmias: the cardiac channelopathies. *Ann. Pediatr. Cardiol.* **8,** 210–220 (2015).
214. Basso, C. *et al.* Sudden cardiac death with normal heart: molecular autopsy. *Cardiovasc. Pathol.* **19,** 321–325 (2010).
215. Wu, Q. *et al.* Forensic pathological study of 1656 cases of sudden cardiac death in southern china. *Medicine.* **95,** (2016).
216. Fishman, G. I. *et al.* Sudden cardiac death prediction and prevention report from a National Heart, Lung, and Blood Institute and Heart Rhythm Society Workshop. *Circulation* **122,** 2335–2348 (2010).
217. Zipes, D. P. & Wellens, H. J. J. Sudden cardiac death. *Circulation* **98,** 2334–2351 (1998).
218. Josephson, M. E. Sudden cardiac arrest. *Indian Heart J.* **66 Suppl 1,** S2-3 (2014).
219. Burke, M. A. *et al.* Clinical and mechanistic insights into the genetics of cardiomyopathy. *J. Am. Coll. Cardiol.* **68,** 2871–2886 (2016).
220. Wexler, R. K. *et al.* Cardiomyopathy: an overview. *Am. Fam. Physician* **79,** 778–784 (2009).
221. Magi, S. *et al.* Sudden cardiac death: focus on the genetics of channelopathies and cardiomyopathies. *J. Biomed. Sci.* **24,** 56 (2017).
222. Fernández-Falgeras, A. *et al.* Cardiac Channelopathies and Sudden Death: Recent Clinical and Genetic Advances. *Biology (Basel).* **6,** 7 (2017).

223. Campuzano, O. *et al.* Genetics and cardiac channelopathies. *Genet Med.* **12,** 260–7 (2010).
224. Nerbonne, J. M. & Kass, R. S. Molecular physiology of cardiac repolarization. *Physiol. Rev.* **85,** 1205–1253 (2005).
225. van den Boogaard, M. *et al.* A common genetic variant within *SCN10A* modulates cardiac *SCN5A* expression. *J. Clin. Invest.* **124,** 1844–1852 (2014).
226. Kapplinger, J. D. *et al.* An international compendium of mutations in the SCN5A-encoded cardiac sodium channel in patients referred for Brugada syndrome genetic testing. *Hear. Rhythm* **7,** 33–46 (2010).
227. Brugada, P. & Brugada, J. Right bundle branch block, persistent ST segment elevation and sudden cardiac death: a distinct clinical and electrocardiographic syndrome. A multicenter report. *J. Am. Coll. Cardiol.* **20,** 1391–1396 (1992).
228. Antzelevitch, C. *et al.* Brugada syndrome: report of the second consensus conference. *Circulation* **111,** 659–670 (2005).
229. Brugada, R. *et al.* Sodium channel blockers identify risk for sudden death in patients with ST-segment elevation and right bundle branch block but structurally normal hearts. *Circulation* **101,** 510–515 (2000).
230. Morita, H. *et al.* Atrial fibrillation and atrial vulnerability in patients with Brugada syndrome. *J. Am. Coll. Cardiol.* **40,** 1437–1444 (2002).
231. Meregalli, P. G. *et al.* Pathophysiological mechanisms of Brugada syndrome: depolarization disorder, repolarization disorder, or more? *Cardiovasc. Res.* **67,** 367–378 (2005).
232. Hoogendijk, M. G. *et al.* The Brugada ECG pattern: a marker of channelopathy, structural heart disease, or neither? Toward a unifying mechanism of the Brugada syndrome. *Circ. Arrhythm. Electrophysiol.* **3,** 283–290 (2010).
233. Tse, G. *et al.* Electrophysiological Mechanisms of Brugada Syndrome: Insights from Pre-clinical and Clinical Studies. *Front. Physiol.* **7,** (2016).
234. Schott, J. J. *et al.* Cardiac conduction defects associate with mutations in SCN5A. *Nat. Genet.* **23,** 20–21 (1999).
235. Riuró, H. *et al.* A missense mutation in the sodium channel β2 subunit reveals SCN2B as a new candidate gene for Brugada syndrome. *Hum. Mutat.* **34,** 961–966 (2013).
236. Hu, D. *et al.* A mutation in the β3 subunit of the cardiac sodium channel associated with brugada ECG phenotype. *Circ. Cardiovasc. Genet.* **2,** 270–278 (2009).
237. Antzelevitch, C. *et al.* Loss-of-function mutations in the cardiac calcium channel underlie a new clinical entity characterized by ST-segment elevation, short QT intervals, and sudden cardiac death. *Circulation* **115,** 442–449 (2007).
238. Burashnikov, E. *et al.* Mutations in the cardiac L-type calcium channel associated with inherited J-wave syndromes and sudden cardiac death. *Hear. Rhythm* **7,** 1872–1882 (2010).
239. Nielsen, M. W. *et al.* The genetic component of brugada syndrome. *Front. Physiol.* **4 JUL,** 1–11 (2013).
240. Bezzina, C. R. *et al.* Common sodium channel promoter haplotype in asian subjects underlies variability in cardiac conduction. *Circulation* **113,** 338–344 (2006).
241. Erikson, G. A. *et al.* Whole-genome sequencing of a healthy aging cohort. *Cell* **165,** 1002–1011 (2016).
242. Kimes, B. W. & Brandt, B. L. Properties of a clonal muscle cell line from rat heart. *Exp. Cell Res.* **98,** 367–381 (1976).
243. Zanella, F. & Sheikh, F. Patient-Specific Induced Pluripotent Stem Cell Models: generation and characterization of cardiac cells*. in Methods Mol Biol.* **1353,** 147–162 (2016).

244. Pagans, S. *et al.* The cellular lysine methyltransferase Set7/9-KMT7 binds HIV-1 TAR RNA, monomethylates the viral transactivator Tat, and enhances HIV transcription. *Cell Host Microbe* **7,** 234–244 (2010).
245. Leung, D. *et al.* Integrative analysis of haplotype-resolved epigenomes across human tissues. *Nature* **518,** 350–354 (2015).
246. Schmitt, A. D. *et al.* A compendium of chromatin contact maps reveals spatially active regions in the human genome. *Cell Rep.* **17,** 2042–2059 (2016).
247. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science.* **337,** 1190–1195 (2012).
248. Wang, H. *et al.* Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Res.* **22,** 1680–1688 (2012).
249. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489,** 75–82 (2012).
250. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526,** 75–81 (2015).
251. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29,** 308–311 (2001).
252. Jiang, H. *et al.* Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics* **15,** 182 (2014).
253. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
254. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
255. Picard Tools - By Broad Institute. Available at: https://broadinstitute.github.io/picard/.
256. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).
257. van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9,** 2579–2605 (2008).
258. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (2012).
259. Sotoodehnia, N. *et al.* Common variants in 22 loci are associated with QRS duration and cardiac ventricular conduction. *Nat. Genet.* **42,** 1068–1076 (2010).
260. Hwang, S. *et al.* Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Sci. Rep.* **5,** 17875 (2015).
261. Mills, R. E. *et al.* Natural genetic variation caused by small insertions and deletions in the human genome. *Genome Res.* **21,** 830–839 (2011).
262. Telenti, A. *et al.* Deep sequencing of 10,000 human genomes. *Proc. Natl. Acad. Sci. U. S. A.* **113,** 11901–11906 (2016).
263. Christoffels, V. M. *et al.* Development of the pacemaker tissues of the heart. *Circ. Res.* **106,** 240–254 (2010).
264. Boogerd, C. J. *et al.* Protein interactions at the heart of cardiac chamber formation. *Ann. Anat.* **191,** 505–517 (2009).
265. Khromykh, A. & Solomon, B. D. The benefits of whole-genome sequencing now and in the future. *Mol. Syndromol.* **6,** 108–109 (2015).
266. Rehm, H. L. *et al.* ACMG clinical laboratory standards for next-generation sequencing. *Genet. Med.* **15,** 733–747 (2013).

267. Short, P. J. *et al.* De novo mutations in regulatory elements in neurodevelopmental disorders. *Nature* **555,** 611–616 (2018).
268. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9,** 215–216 (2012).
269. Sanger, F. *et al.* DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* **74,** 5463–5467 (1977).
270. Beck, T. F. *et al.* Systematic evaluation of sanger validation of next-generation sequencing variants. *Clin. Chem.* **62,** 647–654 (2016).
271. Mu, W. *et al.* Sanger confirmation is required to achieve optimal sensitivity and specificity in next-generation sequencing panel testing. *J. Mol. Diagnostics* **18,** 923–932 (2016).
272. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. **106,** 9362–9367 (2009).
273. Gulko, B. *et al.* A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nat. Genet.* **47,** 276–283 (2015).
274. Maurano, M. T. *et al.* Widespread site-dependent buffering of human regulatory polymorphism. **8,** e1002599 (2012).
275. Leoni, A.L. *et al.* Variable Na(v)1.5 protein expression from the wild-type allele correlates with the penetrance of cardiac conduction disease in the Scn5a(+/-) mouse model. *PLoS One* **5,** e9298 (2010).
276. Janssens, S. & Beyaert, R. A universal role for MyD88 in TLR/IL-1R-mediated signaling. *Trends Biochem. Sci.* **27,** 474–482 (2002).
277. Chen, W. *et al.* Overexpression of myeloid differentiation protein 88 in mice induces mild cardiac dysfunction, but no deficit in heart morphology. *Brazilian J. Med. Biol. Res.* **49,** (2015).
278. Feng, Y. *et al.* MyD88 and Trif signaling play distinct roles in cardiac dysfunction and mortality during endotoxin shock and polymicrobial sepsis. *Anesthesiology* **115,** 555–567 (2011).
279. Singh, M. V *et al.* MyD88 mediated inflammatory signaling leads to CaMKII oxidation, cardiac hypertrophy and death after myocardial infarction. *J. Mol. Cell. Cardiol.* **52,** 1135–1144 (2012).
280. van Hoorn, F. *et al.* SCN5A mutations in Brugada syndrome are associated with increased cardiac dimensions and reduced contractility. *PLoS One* **7,** e42037 (2012).
281. Hummel, Y. M. *et al.* Ventricular dysfunction in a family with long QT syndrome type 3. *EP Eur.* **15,** 1516–1521 (2013).
282. Hosseini, M. *et al.* Reappraisal of reported genes for sudden arrhythmic death. *Circulation.* **138,** 1195–1205 (2018).
283. Antzelevitch, C. *et al.* J-Wave syndromes expert consensus conference report: Emerging concepts and gaps in knowledge. *Journal of Arrhythmia.* **32,** 315–229 (2016).
284. Bezzina, C. R. *et al.* Common variants at *SCN5A-SCN10A* and *HEY2* are associated with Brugada syndrome, a rare disease with high risk of sudden cardiac death. *Nat. Genet.* **45,** 1044–1049 (2013).
285. Probst, V. *et al. SCN5A* mutations and the role of genetic background in the pathophysiology of Brugada syndrome. *Circ. Cardiovasc. Genet.* **2,** 552–557 (2009).

Annex 1

V1: 31/07/12

INFORMED CONSENT FOR THE STUDY OF THE GENETIC BASIS OF HEART DISEASES

(GENCARDIO-RESEARCH)

PRINCIPAL INVESTIGATOR: Ramon Brugada, MD, PhD, FACC, FESC

AIMS

We appreciate your collaboration in the study GENCARDIO-RESEARCH that takes place at the Cardiovascular Genetic Center in Girona. Your participation will help improve our knowledge of heart disease. In recent years significant progress has been made in the investigation of genetic factors in cardiovascular diseases. For such studies it is necessary to collect biological samples (blood, saliva, tissue) and DNA obtained from healthy volunteers and from people with cardiac problems. DNA is the genetic material that is present in all our cells and carries a code in the form of "gene" that determines our personal physical characteristics, like eye color, skin, etc.. These genes can also influence the risk that some people to develop certain diseases. The main objective of this project is to ascertain what features of DNA affect the risk of a person to develop heart disease. We are currently conducting a study that aims to collect and store blood samples to isolate DNA. This DNA is stored in the Cardiovascular Genetics Centre and allows scientists to investigate what genes are associated with heart disease, what diseases are influenced by the environment, and which genes influence the response to certain treatments.

All personal information that is collected or generated by this study will be protected in accordance to the law. To this end we use the measures as detailed below:

PROCES DESCRIPCION

During your participation in the study

- We will inform you about the objectives of the project and answer any questions that you may have.
- To participate in our study you will not receive any financial reward.- The data recorded in your file can be treated statistically for the purpose of research.

V1: 31/07/12

- The data may be provided and processed, anonymously, to a third party who may use it only for research purposes.
- A biological sample of blood, saliva, tissue will be collected in order to extract DNA.
- The Cardiovascular Genetic Center agrees that all information received and all the samples will be codified prior to shipment to outside researchers so the donor cannot be identified.
- Products obtained from the samples will be archived and kept for a minimum period of five years at the Cardiovascular Genetic Center.
- Products obtained from the samples may be used in biomedical research studies conducted by other institutions, national or foreign, provided that: 1) have been considered of scientific interest, 2) that meet the requirements established by the Scientific Committees and External expert Advisors in ethical, economic, environmental, legal and social matters.
- You will be able to know the research studies in which your samples and clinical data have been used, and we will provide you with the results of these studies related to your disease. The Cardiovascular Genetics Center will have copy of the information from each study.
- The Cardiovascular Genetics Center is committed not to sell the samples or data obtained from clinical samples.
- You are entitled to request the destruction of your sample and all the related information which haves been stored at the Cardiovascular Genetic Center.

Contact information:
Dr. Ramon Brugada Terradellas
Cardiovascular Genetic Center
Parc Hospitalari Martí i Julià Carrer Dr. Castany, s/n
Edifici Mancomunitat 2
17190 -Salt- ramon@brugada.org

V1: 31/07/12

DONOR´S STATEMENT

I have been informed by the health professional mentioned below:

- About the advantages and disadvantages of this procedure.

- About the site of collection, storage and processing of the samples and data.

-About the purpose for using my samples and data (genetic studies, public health or statistical that meet all the law requirements, the Advisory Committee of experts in ethical, economic, environmental, legal and Social, and the Scientific Committee).

- That my samples and data will be provided anonymized to the investigators who work with them.

-That, if due to the genetic analysis I have any injury, the Cardiovascular Genetic Center cannot offer me any economic compensation or take care of any medical bills. I will be offered the appropriate treatment as it is provided to the rest of the community.

- That at any time I can revoke my consent and request the destruction of my data and samples stored at the Cardiovascular Genetic Center.

- That at any time I can request information about genetic studies where my samples have been used.

- I understood all the information and I have been able to ask all the questions that I thought appropriate,

I agree to be contacted by the Center´s personnel in order to get additional informationYES___ NO ___

I wish that my DNA sample (or a sample from a deceased family member o minor ofWhom I am responsible) be destroyed once the study has finalized. YES ___ NO ___

V1: 31/07/12

Please, sign parts A, B o C in next page as appropriate.

HEALTH PRFESSIONAL ´S STATEMENT REGARDING PROPERLY INFORMING THE DONOR

Name: Signature:

A. ADULT WITH FULL MENTAL CAPACITY

Patient's name Patient's signature

B. - MINOR OR PATIENT WITH DIMINISHED MENTAL CAPACITY

Minor´s name / patient with diminished mental capacity

Parent´s name/tutor/legal representative

Parent´s signature/tutor/legal representative

C. - DECEASED

Deceased´s name:

Parent´s name /tutor / legal representative

Parent´s signature /tutor / legal representative

Annex 2

Table A-2_1. Forward oligonucleotides used to measure the CTCF binding effects of the 59 CTCF-overlapping variants in luciferase assays.

Chr	Position	Allele*	Sequence (5' - 3')
chr3	37953719	T	TTAAGAGTGGCATGGCTTCCCCCTAGTGGAGGGGAGGA
		A	TTAAGAGTGGCATGGCATCCCCCTAGTGGAGGGGAGGA
chr3	38045036	T	TGTGGGGGCAGTGGCGACCTCTGCTGGCAGCATCCCCA
		A	TGTGGGGGCAGTGGCGACCTCTGCAGGCAGCATCCCCA
chr3	38521504	A	TAGCTGGTTTTTGGTGCCCTCTAATGGCCACATGAAGA
		G	TGGCTGGTTTTTGGTGCCCTCTAATGGCCACATGAAGA
chr3	38780342	G	TTGGCACTTCTCCTGTAGAGGGCGTCCTGGAGCTGAGA
		A	TTGGCACTTCTCCTGTAGAGGGCGTCCTAGAGCTGAGA
chr3	39540276	G	TCTAGAGGTGATGCTGCCCCCTGCCGGACATGGTCGGA
		A	TCTAGAGGTGATGCTGCCCCCTGCCGGACATGGTCAGA
chr3	39854224	A	TTTGGTGTTGTTGCCACATGGTGGCGCCAGGTTGGCTA
		G	TTTGGTGTTGTTGCCGCATGGTGGCGCCAGGTTGGCTA
chr3	39953594	C	TCCCTCTGTGCTCACCTTCCCCCTGCTGGTCTGACTTA
		T	TCCCTTTGTGCTCACCTTCCCCCTGCTGGTCTGACTTA
chr7	79677353	C	TCAAAGACTTATACTGCTATCTAGTGGCTATATTGAGA
		T	TTAAAGACTTATACTGCTATCTAGTGGCTATATTGAGA
chr7	80288990	ATGT	TTTTTCTTCTAGTGACTGCCATCTACTGGTATAATGTA
		A	TTTTTCTTCTAGTGACTGCCATCTACTGGTATAATGTA
chr7	80549633	C	TGTGGGGCACAGCCCAGCAGATGTCACTCTACAGAAGA
		G	TGTGGGGCACAGCGCAGCAGATGTCACTCTACAGAAGA
chr7	80625526	G	TAACCAGGGTTCCCCAGTAGGAGGCAGTCTTGGGGCTA
		A	TAACCAGGGTTCCCCAGTAGAAGGCAGTCTTGGGGCTA
chr7	81076865	TC	TAGAAAGCTGAATGTGCTCCCTAGTGGCCACTGTGTTA
		T	TAGAAAGCTGAATGTGCTCCTAGTGGCCACTGTGTTGA
chr7	81076870	A	TAGAAAGCTGAATGTGCTCCCTAGTGGCCACTGTGTTA
		T	TAGAAAGCTGAATGTGCTCCCTTGTGGCCACTGTGTTA
chr7	81087389	T	TTAGTAAGTTCTATAGCTCCATCTGCTGCTTCATCTTA
		C	TTAGCAAGTTCTATAGCTCCATCTGCTGCTTCATCTTA
chr7	81130725	C	TAAAGTTCCATAGCCACTAAGTGGCAGAATTGGTATTA
		T	TAAAGTTCTATAGCCACTAAGTGGCAGAATTGGTATTA
chr7	82110992	T	TATGGTTCTCATTTCACCCTCTAGTGGTCTTATGACAA
		C	TATGGCTCTCATTTCACCCTCTAGTGGTCTTATGACAA
chr7	82702421	G	TCGGACAGCCTCCCTGCCATCTAGTGATAGCTGGAAAA
		A	TCAGACAGCCTCCCTGCCATCTAGTGATAGCTGGAAAA
chr7	82900741	T	TGAACTGTGTGCAGCACCATCTTGTGACAAATGAAGTA
		G	TGAACTGTGTGCAGCACCATCTTGTGACAAAGGAAGTA
chr10	18883940	C	TCCAGCCCCCAACTACTGCCCTCTGCTGGGCCAGCCTA
		T	TCCAGCCCCCAACTACTGCCCTCTGCTGGGCCAGTCTA
chr10	19457714	C	TCCTCCCTCTCCCTTCATCCCCCTGCTGGTGGCTCAGA
		G	TCCTCCCTCTCCCTTCATCCCCCTGCTGGTGGCTGAGA
chr10	21147418	G	TATTTACAAGGAGCCACTAGATGGTGCTGATTTACTAA
		A	TATTTACAAGGAGCCACTAGATGGTGCTAATTTACTAA
chr11	117122398	A	TAAAGACACTGCCGGCAGGGGGCTCAGAGATGTGAGAA
		T	TAAAGACTCTGCCGGCAGGGGGCTCAGAGATGTGAGAA

Chr	Position	Allele*	Sequence (5' - 3')
chr11	117924278	C	TCCTGCAACCTGAGCGCCCTCTGGTGGCTGAAATGTTA
		T	TTCTGCAACCTGAGCGCCCTCTGGTGGCTGAAATGTTA
chr11	118042498	C	TCGGGCGGAACACGCGTGCCCTCTAGTGGTCACAGAGA
		T	TTGGGCGGAACACGCGTGCCCTCTAGTGGTCACAGAGA
chr11	118549803	C	TACTGCAGGTCCCGACAGGGGGCGGTGGCTCACGCTTA
		T	TACTGCAGGTCCTGACAGGGGGCGGTGGCTCACGCTTA
chr11	118560953	G	TCTCGGCTCACAGCTGCCCCCCGGTGGCTCCGCGGGAA
		GC	TCTCGGCTCACAGCTGCCCCCCGGTGGCTCCGCGGGA
chr11	118886117	C	TAGTGCAGGCTGGCTGCCCTCTTTTGGCCGCGTCTGGA
		T	TAGTGCAGGCTGGCTGCCCTCTTTTGGCCGTGTCTGGA
chr11+	118889370+	C	TGTGCGCAAGTGCAGCAGGTGGCTGCACGGGGGGCGCA
		G	TGTGCGCAAGTGCAGCAGGTGGGTGCACGGGGGGCGCA
chr11	118889378	G	TGTGCGCAAGTGCAGCAGGTGGCTGCACGGGGGGCGCA
		A	TGTGCGCAAGTGCAGCAGGTGGCTGCACGGAGGGCGCA
chr11	119287550	G	TTTCCGAAGTCCGTGGCCCTCTGGTGGCCTCTGTGGCA
		A	TTTCCGAAGTCCGTGGCCCTCTGGTGGCCTCTGTAGCA
chr11	119352239	G	TCGCCGCGGTCGGCCAGGAGGTGGCGCTGTGACTGCAA
		A	TCGCCGCAGTCGGCCAGGAGGTGGCGCTGTGACTGCAA
chr11	119404719	C	TCGCCCGGCTCTCGGGCTCCCCCTGGAGGGCTTTGAAA
		T	TCGCCCGGCTCTTGGGCTCCCCCTGGAGGGCTTTGAAA
chr11	119600183	A	TCAAAGGCAGGCAGTGACCCCTAGCGGCCGGCGGCGCA
		T	TCAATGGCAGGCAGTGACCCCTAGCGGCCGGCGGCGCA
chr11	119612173	C	TGGGACGGGCACGGCGCCACCTAGCGGTCGTGGGCCGA
		T	TGGGACGGGCACGGCGCCATCTAGCGGTCGTGGGCCGA
chr11	120053724	C	TGCCTTTCCCACAGCCCCACCTACTGGGAGAAGCCAGA
		T	TGCCTTTCCCACAGCCTCACCTACTGGGAGAAGCCAGA
chr11	120100704	T	TTGGCCAATGCTGCCACCTGATGTCAGAAAGTGTCCCA
		G	TTGGCCAATGCTGCCACCTGATGGCAGAAAGTGTCCCA
chr11	120173911	C	TTGGAGGGGAAGCTCGCCATCTAGTGGTTGAAATGGGA
		CTGAAG	TTGGAGGGGAAGCTCTGAAGGCCATCTAGTGGTTGAAA
chr11	120173914	C	TCCTGGAGGGGAAGCTCGCCATCTAGTGGTTGAAATGA
		CGGG	TCCTGGAGGGGAAGCTCGCCGGGATCTAGTGGTTGAAA
chr11	120173917	C	TTGGAGGGGAAGCTCGCCATCTAGTGGTTGAAATGGGA
		CT	TTGGAGGGGAAGCTCGCCATCTTAGTGGTTGAAATGGA
chr11	122659691	T	TCTCCTGGGCACATCGCCCTCTGGAGGTGTGCAGTATA
		C	TCTCCCGGGCACATCGCCCTCTGGAGGTGTGCAGTATA
chr11	123036887	G	TACCTCGCAGGCCTGAAGGGGGAGCAGTCGAGCCCAAA
		C	TACCTCGCAGGCCTGAAGCGGGAGCAGTCGAGCCCAAA
chr11	123105986	G	TAAGTGCTCTTGAGTGACTCCTGCTGGTCGTGGATGGA
		T	TAAGTTCTCTTGAGTGACTCCTGCTGGTCGTGGATGGA
chr11	123118102	CT	TGTTGGTTAGATACCGCCCCTTGTGGGCAGCGGGTGA
		C	TGTTGGTTAGATACCGCCCCTGTGGGCAGCGGGTGTA
chr11	123132370	G	TAGCTCAGTGCTAGTGCCACCTAGTGGCTGTCCCGGCA
		C	TAGCTCACTGCTAGTGCCACCTAGTGGCTGTCCCGGCA
chr11	123425811	A	TCTAGTGTCTCTCTAGCGCCCTCTACTGACAAGATATA
		C	TCTAGTGTCTCTCTAGCGCCCTCTACTGACAAGCTATA

Chr	Position	Allele*	Sequence (5' - 3')
chr11	123447866	A	**T**AC**A**TGGGCACTGCTGCCCTCCCGTGGCCAACGGCAG**A**
		G	**T**AC**G**TGGGCACTGCTGCCCTCCCGTGGCCAACGGCAG**A**
chr11	123511861	A	**T**CAGATGGTGACCAGGAGAGGGCGCCAGGGCCCC**A**CC**A**
		AC	**T**CAGATGGTGACCAGGAGAGGGCGCCAGGGCCCC**AC**C**A**
chr11	123511862	C	**T**CAGATGGTGACCAGGAGAGGGCGCCAGGGCCCCA**C**C**A**
		T	**T**CAGATGGTGACCAGGAGAGGGCGCCAGGGCCCCA**T**C**A**
chr11	124539211	G	**T**TTTTCAGATAATCCAGCAG**G**TGGCAGAAGAGGGCAA**A**
		A	**T**TTTTCAGATAATCCAGCAG**A**TGGCAGAAGAGGGCAA**A**
chr11	124648367	T	**T**ACATGATGCTGTTGTCTCCCCC**T**GGAGGAAGAGTTG**A**
		G	**T**ACATGATGCTGTTGTCTCCCCC**G**GGAGGAAGAGTTG**A**
chr11	124984629	G	**T**CTACTACTACTACTGCCGCCTGCTGGAA**G**ATGGGGA**A**
		T	**T**CTACTACTACTACTGCCGCCTGCTGGAA**T**ATGGGGA**A**
chr12	2469918	T	**T**TGTCTCTAC**T**GGCTGTTTGAAGAATTGCAATGCTCG**A**
		A	**T**TGTCTCTAC**A**GGCTGTTTGAAGAATTGCAATGCTCG**A**
chr12	2733860	C	**T**TACAGAGCTCCACAG**C**GCCCCGTGGTGGCCAGTGTC**A**
		T	**T**TACAGAGCTCCACAG**T**GCCCCGTGGTGGCCAGTGTC**A**
chr12	3053956	G	**T**GATGGTTCCATGGC**G**CCACCCACTGTCCAGGAATGC**A**
		A	**T**GATGGTTCCATGGC**A**CCACCCACTGTCCAGGAATGC**A**
chr12	3384802	G	**T**AACCAGCCTGGAGCGCCACCTAGTGGCC**G**TAGGCGG**A**
		A	**T**AACCAGCCTGGAGCGCCACCTAGTGGCC**A**TAGGCGG**A**
chr12	3451418	C	**T**GGCACCAGCTAGACAGCAGGGGG**C**AGGCAAGGCAGA**A**
		T	**T**GGCACCAGCTAGACAGCAGGGGG**T**AGGCAAGGCAGA**A**
chr12	3764916	G	**T**GGGCACCACCCAGCCTGCCTCCTGGAGGCTC**G**GGTA**A**
		C	**T**GGGCACCACCCAGCCTGCCTCCTGGAGGCTC**C**GGTA**A**
chr12	3767720	G	**T**AATGCTGACAAGCCTCCAGAGG**G**CGCTAATTGGAGG**A**
		T	**T**AATGCTGACAAGCCTCCAGAGG**T**CGCTAATTGGAGG**A**
chr12	3913211	G	**T**C**G**AAGCAAGATGCTGCCACCTAGCGTCTGCTGTTCT**A**
		A	**T**C**A**AAGCAAGATGCTGCCACCTAGCGTCTGCTGTTCT**A**

*Top nucleotide corresponds to the reference allele and the bottom nucleotide corresponds to the alternative allele.

+Oligonucleotides that could not be cloned.

In the sequence column, the reference allele is highlighted in blue, while the alternative allele is highlighted in orange. The extra nucleotides added to originate a completely different restriction site after plasmid ligation are marked in bold. Chr (chromosome).

Table A-2_2. Reverse oligonucleotides used to measure the CTCF binding effects of the the 59 CTCF-overlapping variants in luciferase assays.

Chr	Position	Allele*	Sequence (5' - 3')
chr3	37953719	T	CTAG**T**CCTCCCCTCCACTAGGGGGAAGCCATGCCACTCTTA**A**GTAC
		A	CTAG**T**CCTCCCCTCCACTAGGGGGATGCCATGCCACTCTTA**A**GTAC
chr3	38045036	T	CTAG**T**GGGGATGCTGCCAGCAGAGGTCGCCACTGCCCCCAC**A**GTAC
		A	CTAG**T**GGGGATGCTGCCTGCAGAGGTCGCCACTGCCCCCAC**A**GTAC
chr3	38521504	A	CTAG**T**CTTCATGTGGCCATTAGAGGGCACCAAAAACCAGCT**A**GTAC
		G	CTAG**T**CTTCATGTGGCCATTAGAGGGCACCAAAAACCAGCC**A**GTAC
chr3	38780342	G	CTAG**T**CTCAGCTCCAGGACGCCCTCTACAGGAGAAGTGCCA**A**GTAC
		A	CTAG**T**CTCAGCTCTAGGACGCCCTCTACAGGAGAAGTGCCA**A**GTAC
chr3	39540276	G	CTAG**T**CCGACCATGTCCGGCAGGGGGCAGCATCACCTCTAG**A**GTAC
		A	CTAG**T**CTGACCATGTCCGGCAGGGGGCAGCATCACCTCTAG**A**GTAC
chr3	39854224	A	CTAG**T**AGCCAACCTGGCGCCACCATGTGGCAACAACACCAA**A**GTAC
		G	CTAG**T**AGCCAACCTGGCGCCACCATGCGGCAACAACACCAA**A**GTAC
chr3	39953594	C	CTAG**T**AAGTCAGACCAGCAGGGGGAAGGTGAGCACAGAGGG**A**GTAC
		T	CTAG**T**AAGTCAGACCAGCAGGGGGAAGGTGAGCACAAAGGG**A**GTAC
chr7	79677353	C	CTAG**T**CTCAATATAGCCACTAGATAGCAGTATAAGTCTTTG**A**GTAC
		T	CTAG**T**CTCAATATAGCCACTAGATAGCAGTATAAGTCTTTA**A**GTAC
chr7	80288990	ATGT	CTAG**T**ACATTATACCAGTAGATGGCAGTCACTAGAAGAAAA**A**GTAC
		A	CTAG**T**ACATTATACCAGTAGATGGCAGTCACTAGAAGAAAA**A**GTAC
chr7	80549633	C	CTAG**T**CTTCTGTAGAGTGACATCTGCTGGGCTGTGCCCCAC**A**GTAC
		G	CTAG**T**CTTCTGTAGAGTGACATCTGCTGCGCTGTGCCCCAC**A**GTAC
chr7	80625526	G	CTAG**T**AGCCCCAAGACTGCCTCCTACTGGGGAACCCTGGTT**A**GTAC
		A	CTAG**T**AGCCCCAAGACTGCCTTCTACTGGGGAACCCTGGTT**A**GTAC

Chr	Position	Alele*	Sequence (5' - 3')
chr7	81076865	TC	CTAG**T**AACACAGTGGCCACTAGGGAGCACATTCAGCTTTCT**A**GTAC
		T	CTAG**T**CAACACAGTGGCCACTAGGAGCACATTCAGCTTTCT**A**GTAC
chr7	81076870	A	CTAG**T**AACACAGTGGCCACTAGGGAGCACATTCAGCTTTCT**A**GTAC
		T	CTAG**T**AACACAGTGGCCACAAGGGAGCACATTCAGCTTTCT**A**GTAC
chr7	81087389	T	CTAG**T**AAGATGAAGCAGCAGATGGAGCTATAGAACTTACTA**A**GTAC
		C	CTAG**T**AAGATGAAGCAGCAGATGGAGCTATAGAACTTGCTA**A**GTAC
chr7	81130725	C	CTAG**T**AATACCAATTCTGCCACTTAGTGGCTATGGAACTTT**A**GTAC
		T	CTAG**T**AATACCAATTCTGCCACTTAGTGGCTATAGAACTTT**A**GTAC
chr7	82110992	T	CTAG**T**TGTCATAAGACCACTAGAGGGTGAAATGAGAACCAT**A**GTAC
		C	CTAG**T**TGTCATAAGACCACTAGAGGGTGAAATGAGAGCCAT**A**GTAC
chr7	82702421	G	CTAG**T**TTTCCAGCTATCACTAGATGGCAGGGAGGCTGTCCG**A**GTAC
		A	CTAG**T**TTTCCAGCTATCACTAGATGGCAGGGAGGCTGTCTG**A**GTAC
chr7	82900741	T	CTAG**T**ACTTCATTTGTCACAAGATGGTGCTGCACACAGTTC**A**GTAC
		G	CTAG**T**ACTTCCTTTGTCACAAGATGGTGCTGCACACAGTTC**A**GTAC
chr10	18883940	C	CTAG**T**AGGCTGGCCCAGCAGAGGGCAGTAGTTGGGGGCTGG**A**GTAC
		T	CTAG**T**AGACTGGCCCAGCAGAGGGCAGTAGTTGGGGGCTGG**A**GTAC
chr10	19457714	C	CTAG**T**CTGAGCCACCAGCAGGGGGATGAAGGGAGAGGGAGG**A**GTAC
		G	CTAG**T**CTCAGCCACCAGCAGGGGGATGAAGGGAGAGGGAGG**A**GTAC
chr10	21147418	G	CTAG**T**TAGTAAATCAGCACCATCTAGTGGCTCCTTGTAAAT**A**GTAC
		A	CTAG**T**TAGTAAATTAGCACCATCTAGTGGCTCCTTGTAAAT**A**GTAC
chr11	117122398	A	CTAG**T**TCTCACATCTCTGAGCCCCCTGCCGGCAGTGTCTTT**A**GTAC
		T	CTAG**T**TCTCACATCTCTGAGCCCCCTGCCGGCAGAGTCTTT**A**GTAC

Chr	Position	Allele*	Sequence (5' - 3')
chr11	117924278	C	CTAG**T**AACATTTCAGCCACCAGAGGGCGCTCAGGTTGCAGG**A**GTAC
		T	CTAG**T**AACATTTCAGCCACCAGAGGGCGCTCAGGTTGCAGA**A**GTAC
chr11	118042498	C	CTAG**T**CTCTGTGACCACTAGAGGGCACGCGTGTTCCGCCCG**A**GTAC
		T	CTAG**T**CTCTGTGACCACTAGAGGGCACGCGTGTTCCGCCCA**A**GTAC
chr11	118549803	C	CTAG**T**AAGCGTGAGCCACCGCCCCCTGTCGGGACCTGCAGT**A**GTAC
		T	CTAG**T**AAGCGTGAGCCACCGCCCCCTGTCAGGACCTGCAGT**A**GTAC
chr11	118560953	G	CTAG**T**TCCCGCGGAGCCACCGGGGGGGCAGCTGTGAGCCGAG**A**GTAC
		GC	CTAG**T**CCCGCGGAGCCACCGGGGGGGCAGCTGTGAGCCGAG**A**GTAC
chr11	118886117	C	CTAG**T**CCAGACGCGGCCAAAAGAGGGCAGCCAGCCTGCACT**A**GTAC
		T	CTAG**T**CCAGACACGGCCAAAAGAGGGCAGCCAGCCTGCACT**A**GTAC
chr11^{+}	118889370^{+}	C	CTAG**T**GCGCCCCCCGTGCAGCCACCTGCTGCACTTGCGCAC**A**GTAC
		G	CTAG**T**GCGCCCCCCGTGCACCCACCTGCTGCACTTGCGCAC**A**GTAC
chr11	118889378	G	CTAG**T**GCGCCCCCCGTGCAGCCACCTGCTGCACTTGCGCAC**A**GTAC
		A	CTAG**T**GCGCCCTCCGTGCAGCCACCTGCTGCACTTGCGCAC**A**GTAC
chr11	119287550	G	CTAG**T**GCCACAGAGGCCACCAGAGGGCCACGGACTTCGGAA**A**GTAC
		A	CTAG**T**GCTACAGAGGCCACCAGAGGGCCACGGACTTCGGAA**A**GTAC
chr11	119352239	G	CTAG**T**TGCAGTCACAGCGCCACCTCCTGGCCGACCGCGGCG**A**GTAC
		A	CTAG**T**TGCAGTCACAGCGCCACCTCCTGGCCGACTGCGGCG**A**GTAC
chr11	119404719	C	CTAG**T**TTCAAAGCCCTCCAGGGGGAGCCCGAGAGCCGGGCG**A**GTAC
		T	CTAG**T**TTCAAAGCCCTCCAGGGGGAGCCCAAGAGCCGGGCG**A**GTAC
chr11	119600183	A	CTAG**T**GCGCCGCCGGCCGCTAGGGGTCACTGCCTGCCTTTG**A**GTAC
		T	CTAG**T**GCGCCGCCGGCCGCTAGGGGTCACTGCCTGCCATTG**A**GTAC

Chr	Position	Allele*	Sequence (5' - 3')
chr11	119612173	C	CTAG**T**CGGCCCACGACCGCTAGGTGGCGCCGTGCCCGTCCC**A**GTAC
		T	CTAG**T**CGGCCCACGACCGCTAGATGGCGCCGTGCCCGTCCC**A**GTAC
chr11	120053724	C	CTAG**T**CTGGCTTCTCCCAGTAGGTGGGGCTGTGGGAAAGGC**A**GTAC
		T	CTAG**T**CTGGCTTCTCCCAGTAGGTGAGGCTGTGGGAAAGGC**A**GTAC
chr11	120100704	T	CTAG**T**GGGACACTTTCTGACATCAGGTGGCAGCATTGGCCA**A**GTAC
		G	CTAG**T**GGGACACTTTCTGCCATCAGGTGGCAGCATTGGCCA**A**GTAC
chr11	120173911	C	CTAG**T**CCCATTTCAACCACTAGATGGCGAGCTTCCCCTCCA**A**GTAC
		CTGAAG	CTAG**T**TTCAACCACTAGATGGCCTTCAGAGCTTCCCCTCCA**A**GTAC
chr11	120173914	C	CTAG**T**CATTTCAACCACTAGATGGCGAGCTTCCCCTCCAGG**A**GTAC
		CGGG	CTAG**T**TTCAACCACTAGATCCCGGCGAGCTTCCCCTCCAGG**A**GTAC
chr11	120173917	C	CTAG**T**CCCATTTCAACCACTAGATGGCGAGCTTCCCCTCCA**A**GTAC
		CT	CTAG**T**CCATTTCAACCACTAAGATGGCGAGCTTCCCCTCCA**A**GTAC
chr11	122659691	T	CTAG**T**ATACTGCACACCTCCAGAGGGCGATGTGCCCAGGAG**A**GTAC
		C	CTAG**T**ATACTGCACACCTCCAGAGGGCGATGTGCCCGGGAG**A**GTAC
chr11	123036887	G	CTAG**T**TTGGGCTCGACTGCTCCCCCTTCAGGCCTGCGAGGT**A**GTAC
		C	CTAG**T**TTGGGCTCGACTGCTCCCGCTTCAGGCCTGCGAGGT**A**GTAC
chr11	123105986	G	CTAG**T**CCATCCACGACCAGCAGGAGTCACTCAAGAGCACTT**A**GTAC
		T	CTAG**T**CCATCCACGACCAGCAGGAGTCACTCAAGAGAACTT**A**GTAC
chr11	123118102	CT	CTAG**T**CACCCGCTGCCCACAAGGGGGCGGTATCTAACCAAC**A**GTAC
		C	CTAG**T**ACACCCGCTGCCCACAGGGGGCGGTATCTAACCAAC**A**GTAC
chr11	123132370	G	CTAG**T**GCCGGGACAGCCACTAGGTGGCACTAGCACTGAGCT**A**GTAC
		C	CTAG**T**GCCGGGACAGCCACTAGGTGGCACTAGCAGTGAGCT**A**GTAC

Chr	Position	Allele*	Sequence (5' - 3')
chr11	123425811	A	CTAG**T**ATATCTTGTCAGTAGAGGGCGCTAGAGAGACACTAG**A**GTAC
		C	CTAG**T**ATAGCTTGTCAGTAGAGGGCGCTAGAGAGACACTAG**A**GTAC
chr11	123447866	A	CTAG**T**CTGCCGTTGGCCACGGGAGGGCAGCAGTGCCCATGT**A**GTAC
		G	CTAG**T**CTGCCGTTGGCCACGGGAGGGCAGCAGTGCCCACGT**A**GTAC
chr11	123511861	A	CTAG**T**GGTGGGGCCCTGGCGCCCTCTCCTGGTCACCATCTG**A**GTAC
		AC	CTAG**T**GGTGGGGCCCTGGCGCCCTCTCCTGGTCACCATCTG**A**GTAC
chr11	123511862	C	CTAG**T**GGTGGGGCCCTGGCGCCCTCTCCTGGTCACCATCTG**A**GTAC
		T	CTAG**T**GATGGGGCCCTGGCGCCCTCTCCTGGTCACCATCTG**A**GTAC
chr11	124539211	G	CTAG**T**TTGCCCTCTTCTGCCACCTGCTGGATTATCTGAAAA**A**GTAC
		A	CTAG**T**TTGCCCTCTTCTGCCATCTGCTGGATTATCTGAAAA**A**GTAC
chr11	124648367	T	CTAG**T**CAACTCTTCCTCCAGGGGGAGACAACAGCATCATGT**A**GTAC
		G	CTAG**T**CAACTCTTCCTCCCGGGGGAGACAACAGCATCATGT**A**GTAC
chr11	124984629	G	CTAG**T**TCCCCATCTTCCAGCAGGCGGCAGTAGTAGTAGTAG**A**GTAC
		T	CTAG**T**TCCCCATATTCCAGCAGGCGGCAGTAGTAGTAGTAG**A**GTAC
chr12	2469918	T	CTAG**T**CGAGCATTGCAATTCTTCAAACAGCCAGTAGAGACA**A**GTAC
		A	CTAG**T**CGAGCATTGCAATTCTTCAAACAGCCTGTAGAGACA**A**GTAC
chr12	2733860	C	CTAG**T**GACACTGGCCACCACGGGGCGCTGTGGAGCTCTGTA**A**GTAC
		T	CTAG**T**GACACTGGCCACCACGGGGCACTGTGGAGCTCTGTA**A**GTAC
chr12	3053956	G	CTAG**T**GCATTCCTGGACAGTGGGTGGCGCCATGGAACCATC**A**GTAC
		A	CTAG**T**GCATTCCTGGACAGTGGGTGGTGCCATGGAACCATC**A**GTAC
chr12	3384802	G	CTAG**T**CCGCCTACGGCCACTAGGTGGCGCTCCAGGCTGGTT**A**GTAC
		A	CTAG**T**CCGCCTATGGCCACTAGGTGGCGCTCCAGGCTGGTT**A**GTAC

Chr	Position	Alele*	Sequence (5' - 3')
chr12	3451418	C	CTAG**T**TCTGCCTTGCCTGCCCCCTGCTGTCTAGCTGGTGCC**A**GTAC
		T	CTAG**T**TCTGCCTTGCCTACCCCCTGCTGTCTAGCTGGTGCC**A**GTAC
chr12	3764916	G	CTAG**T**TACCCGAGCCTCCAGGAGGCAGGCTGGGTGGTGCCC**A**GTAC
		C	CTAG**T**TACCGGAGCCTCCAGGAGGCAGGCTGGGTGGTGCCC**A**GTAC
chr12	3767720	G	CTAG**T**CCTCCAATTAGCGCCCTCTGGAGGCTTGTCAGCATT**A**GTAC
		T	CTAG**T**CCTCCAATTAGCGACCTCTGGAGGCTTGTCAGCATT**A**GTAC
chr12	3913211	G	CTAG**T**AGAACAGCAGACGCTAGGTGGCAGCATCTTGCTTCG**A**GTAC
		A	CTAG**T**AGAACAGCAGACGCTAGGTGGCAGCATCTTGCTTTG**A**GTAC

*Top nucleotide corresponds to the reference allele and the bottom nucleotide corresponds to the alternative allele. [+]Oligonucleotides that could not be cloned.

The extra nucleotides added to originate a completely different restriction site after plasmid ligation are marked in bold. The 5' and 3' overhangs required for the plasmid ligation are highlighted in red.

Annex 3

Table A-3. Total list of 1,293 Regulome-seq regions for all 6 locus considered. The genomic position, the locus and the length of each region is indicated. The regulatory feature used to select each region is also shown.

Chr	Start	End	Locus	Length (bp)	Regulatory feature
chr3	37427943	37428712	SCN5A	769	DHS
chr3	37443823	37444337	SCN5A	514	DHS
chr3	37493013	37494810	SCN5A	1,797	DHS-CTCF-H3K4me3
chr3	37495418	37495994	SCN5A	576	DHS-CTCF
chr3	37497806	37498434	SCN5A	628	DHS
chr3	37499027	37499809	SCN5A	782	DHS
chr3	37534480	37535106	SCN5A	626	DHS
chr3	37567801	37568692	SCN5A	891	DHS
chr3	37579498	37580304	SCN5A	806	DHS
chr3	37588310	37588974	SCN5A	664	DHS-CTCF
chr3	37595900	37596050	SCN5A	150	CTCF
chr3	37600975	37601488	SCN5A	513	DHS
chr3	37612668	37613330	SCN5A	662	DHS-CTCF
chr3	37727880	37728030	SCN5A	150	CTCF
chr3	37742281	37742684	SCN5A	403	DHS-CTCF
chr3	37754497	37755389	SCN5A	892	DHS
chr3	37783168	37783523	SCN5A	355	DHS-CTCF
chr3	37804163	37804795	SCN5A	632	DHS
chr3	37806927	37807930	SCN5A	1,003	DHS
chr3	37809862	37810363	SCN5A	501	DHS
chr3	37812317	37813658	SCN5A	1,341	DHS
chr3	37820775	37821624	SCN5A	849	DHS-CTCF
chr3	37823887	37824513	SCN5A	626	DHS
chr3	37828452	37829619	SCN5A	1,167	DHS
chr3	37845566	37845921	SCN5A	355	DHS-CTCF
chr3	37854032	37854380	SCN5A	348	DHS
chr3	37856038	37856836	SCN5A	798	DHS-CTCF
chr3	37864678	37865068	SCN5A	390	DHS
chr3	37875380	37875530	SCN5A	150	CTCF
chr3	37896810	37897550	SCN5A	740	DHS-CTCF
chr3	37901580	37905343	SCN5A	3,763	DHS-H3K4me3
chr3	37913832	37914333	SCN5A	501	DHS
chr3	37934366	37934997	SCN5A	631	DHS
chr3	37945952	37946722	SCN5A	770	DHS
chr3	37947723	37948349	SCN5A	626	DHS
chr3	37950802	37951429	SCN5A	627	DHS-CTCF
chr3	37953569	37954298	SCN5A	729	DHS-CTCF
chr3	37964068	37965204	SCN5A	1,136	DHS

Chr	Start	End	Locus	Length (bp)	Regulatory feature
chr3	37968074	37969521	SCN5A	1,447	DHS
chr3	37976267	37978681	SCN5A	2,414	DHS
chr3	37986916	37988386	SCN5A	1,470	DHS
chr3	37991288	37992059	SCN5A	771	DHS
chr3	38008693	38009295	SCN5A	602	DHS-CTCF
chr3	38015926	38016698	SCN5A	772	DHS-CTCF
chr3	38027500	38027650	SCN5A	150	DHS
chr3	38034975	38036782	SCN5A	1,807	DHS-H3K4me3
chr3	38038109	38039010	SCN5A	901	DHS-CTCF
chr3	38039797	38041069	SCN5A	1,272	DHS-H3K4me3
chr3	38043080	38043230	SCN5A	150	CTCF
chr3	38044599	38047290	SCN5A	2,691	DHS-CTCF
chr3	38064922	38067603	SCN5A	2,681	DHS-CTCF-H3K4me3
chr3	38069976	38072101	SCN5A	2,125	DHS-H3K4me3
chr3	38080370	38082019	SCN5A	1,649	DHS-CTCF-H3K4me3
chr3	38125888	38126533	SCN5A	645	DHS
chr3	38159291	38160024	SCN5A	733	DHS-CTCF
chr3	38172937	38173651	SCN5A	714	DHS-CTCF
chr3	38177251	38181540	SCN5A	4,289	DHS-H3K4me3
chr3	38192780	38192930	SCN5A	150	CTCF
chr3	38195452	38196078	SCN5A	626	DHS
chr3	38205557	38208432	SCN5A	2,875	DHS-H3K4me3
chr3	38216071	38216572	SCN5A	501	DHS
chr3	38269304	38270217	SCN5A	913	DHS
chr3	38270604	38270880	SCN5A	276	DHS
chr3	38288700	38289430	SCN5A	730	DHS-CTCF
chr3	38323428	38324059	SCN5A	631	CTCF
chr3	38325244	38325745	SCN5A	501	DHS
chr3	38339535	38340231	SCN5A	696	CTCF
chr3	38358740	38358980	SCN5A	240	DHS
chr3	38369022	38369395	SCN5A	373	DHS
chr3	38376088	38376715	SCN5A	627	DHS
chr3	38385360	38385510	SCN5A	150	CTCF
chr3	38387300	38387450	SCN5A	150	CTCF
chr3	38387750	38389551	SCN5A	1,801	DHS-CTCF-H3K4me3
chr3	38390475	38390734	SCN5A	259	DHS
chr3	38394691	38395025	SCN5A	334	DHS
chr3	38396271	38397048	SCN5A	777	DHS
chr3	38438835	38439305	SCN5A	470	DHS
chr3	38445553	38446361	SCN5A	808	DHS
chr3	38480008	38480327	SCN5A	319	DHS

Chr	Start	End	Locus	Length (bp)	Regulatory feature
chr3	38494654	38497994	SCN5A	3,340	DHS-H3K4me3
chr3	38500549	38501114	SCN5A	565	DHS
chr3	38521195	38521986	SCN5A	791	DHS-CTCF
chr3	38537325	38539127	SCN5A	1,802	DHS-CTCF-H3K4me3
chr3	38562740	38563523	SCN5A	783	DHS-CTCF
chr3	38569980	38570130	SCN5A	150	CTCF
chr3	38583313	38583757	SCN5A	444	DHS
chr3	38601692	38602557	SCN5A	865	CTCF
chr3	38633773	38634073	SCN5A	300	CTCF
chr3	38636832	38637554	SCN5A	722	DHS
chr3	38665260	38665410	SCN5A	150	CTCF
chr3	38669317	38669969	SCN5A	652	CTCF
chr3	38687900	38688740	SCN5A	840	DHS-CTCF
chr3	38690364	38691966	SCN5A	1,602	DHS-H3K4me3
chr3	38692380	38693256	SCN5A	876	DHS-CTCF
chr3	38696990	38697746	SCN5A	756	DHS
chr3	38704069	38704575	SCN5A	506	DHS
chr3	38749129	38749729	SCN5A	600	CTCF
chr3	38767078	38767946	SCN5A	868	DHS
chr3	38771841	38772213	SCN5A	372	DHS
chr3	38772991	38773619	SCN5A	628	DHS
chr3	38777404	38777704	SCN5A	300	CTCF
chr3	38780029	38780719	SCN5A	690	DHS-CTCF
chr3	38823640	38823790	SCN5A	150	CTCF
chr3	38823800	38823950	SCN5A	150	CTCF
chr3	38847191	38847692	SCN5A	501	DHS
chr3	38870510	38871519	SCN5A	1,009	DHS
chr3	38875419	38876312	SCN5A	893	DHS
chr3	38877180	38877330	SCN5A	150	CTCF
chr3	38882503	38882883	SCN5A	380	DHS-CTCF
chr3	38885329	38885901	SCN5A	572	DHS-CTCF
chr3	38894451	38894922	SCN5A	471	DHS
chr3	38931387	38931873	SCN5A	486	DHS
chr3	39002281	39003014	SCN5A	733	DHS-CTCF
chr3	39025760	39025910	SCN5A	150	CTCF
chr3	39033855	39034346	SCN5A	491	DHS
chr3	39041276	39041993	SCN5A	717	CTCF
chr3	39054840	39055209	SCN5A	369	DHS
chr3	39072177	39072779	SCN5A	602	CTCF
chr3	39079492	39080374	SCN5A	882	DHS
chr3	39092560	39094805	SCN5A	2,245	DHS-H3K4me3

Chr	Start	End	Locus	Length (bp)	Regulatory feature
chr3	39106677	39107239	SCN5A	562	DHS-CTCF
chr3	39121100	39121250	SCN5A	150	CTCF
chr3	39130446	39130795	SCN5A	349	DHS
chr3	39132503	39132845	SCN5A	342	DHS
chr3	39138339	39138576	SCN5A	237	DHS
chr3	39142340	39143124	SCN5A	784	DHS-CTCF
chr3	39144740	39144890	SCN5A	150	CTCF
chr3	39147495	39150076	SCN5A	2,581	DHS-H3K4me3
chr3	39166658	39167256	SCN5A	598	DHS-CTCF
chr3	39167399	39167680	SCN5A	281	DHS
chr3	39167907	39168600	SCN5A	693	DHS-CTCF
chr3	39180602	39180949	SCN5A	347	DHS
chr3	39187680	39187830	SCN5A	150	CTCF
chr3	39188230	39190128	SCN5A	1,898	DHS
chr3	39191913	39195869	SCN5A	3,956	DHS-H3K4me3
chr3	39207351	39209356	SCN5A	2,005	DHS-CTCF
chr3	39218720	39220476	SCN5A	1,756	DHS-H3K4me3
chr3	39221913	39223077	SCN5A	1,164	DHS-CTCF-H3K4me3
chr3	39231140	39231886	SCN5A	746	DHS-CTCF
chr3	39243465	39244107	SCN5A	642	DHS-CTCF
chr3	39249822	39250515	SCN5A	693	DHS-CTCF
chr3	39302356	39303055	SCN5A	699	DHS-CTCF
chr3	39314600	39315070	SCN5A	470	CTCF
chr3	39329705	39330449	SCN5A	744	DHS-CTCF
chr3	39333209	39334020	SCN5A	811	DHS-CTCF
chr3	39403515	39404306	SCN5A	791	DHS
chr3	39419517	39420129	SCN5A	612	DHS-CTCF
chr3	39423995	39426653	SCN5A	2,658	DHS-CTCF-H3K4me3
chr3	39446978	39450150	SCN5A	3,172	DHS-CTCF-H3K4me3
chr3	39452001	39452401	SCN5A	400	DHS
chr3	39455079	39455409	SCN5A	330	DHS
chr3	39458660	39459442	SCN5A	782	DHS-CTCF
chr3	39470105	39470734	SCN5A	629	DHS
chr3	39475096	39475597	SCN5A	501	DHS
chr3	39485120	39485310	SCN5A	190	CTCF
chr3	39489191	39489926	SCN5A	735	DHS-CTCF
chr3	39539981	39540644	SCN5A	663	DHS-CTCF
chr3	39543517	39544641	SCN5A	1,124	DHS-H3K4me3
chr3	39547744	39548430	SCN5A	686	DHS
chr3	39572981	39573900	SCN5A	919	DHS
chr3	39583208	39583870	SCN5A	662	DHS-CTCF

Chr	Start	End	Locus	Length (bp)	Regulatory feature
chr3	39589740	39590306	SCN5A	566	DHS
chr3	39679270	39679949	SCN5A	679	DHS-CTCF
chr3	39680787	39681538	SCN5A	751	DHS
chr3	39685488	39686248	SCN5A	760	DHS
chr3	39709817	39710566	SCN5A	749	DHS-CTCF
chr3	39712655	39713466	SCN5A	811	DHS
chr3	39746916	39747544	SCN5A	628	DHS
chr3	39748619	39749404	SCN5A	785	DHS-CTCF
chr3	39765696	39766246	SCN5A	550	DHS-CTCF
chr3	39769000	39769150	SCN5A	150	DHS
chr3	39846485	39846907	SCN5A	422	DHS
chr3	39850090	39852135	SCN5A	2,045	DHS-H3K4me3
chr3	39853923	39854640	SCN5A	717	CTCF
chr3	39855519	39855876	SCN5A	357	DHS
chr3	39949282	39950033	SCN5A	751	DHS
chr3	39952364	39953808	SCN5A	1,444	DHS
chr3	40002056	40003233	SCN5A	1,177	DHS
chr3	40003998	40004533	SCN5A	535	DHS
chr3	40022876	40023648	SCN5A	772	DHS
chr3	40032028	40032451	SCN5A	423	DHS-CTCF
chr3	40084729	40085254	SCN5A	525	DHS
chr3	40085890	40086518	SCN5A	628	DHS
chr3	40088988	40089271	SCN5A	283	DHS
chr7	79049576	79050350	CACNA2D1	774	DHS
chr7	79080560	79085112	CACNA2D1	4,552	DHS-H3K4me3
chr7	79090941	79091442	CACNA2D1	501	DHS
chr7	79124803	79125819	CACNA2D1	1,016	DHS
chr7	79131420	79131870	CACNA2D1	450	CTCF
chr7	79134200	79134350	CACNA2D1	150	CTCF
chr7	79135685	79136311	CACNA2D1	626	DHS
chr7	79183228	79183729	CACNA2D1	501	DHS
chr7	79239580	79240767	CACNA2D1	1,187	DHS
chr7	79281809	79282310	CACNA2D1	501	DHS
chr7	79321142	79321910	CACNA2D1	768	DHS
chr7	79371041	79371396	CACNA2D1	355	CTCF
chr7	79399290	79399916	CACNA2D1	626	CTCF
chr7	79407996	79409149	CACNA2D1	1,153	CTCF
chr7	79410829	79411862	CACNA2D1	1,033	CTCF
chr7	79416847	79417475	CACNA2D1	628	CTCF
chr7	79418817	79419444	CACNA2D1	627	CTCF
chr7	79435656	79436011	CACNA2D1	355	DHS-CTCF

Chr	Start	End	Locus	Length (bp)	Regulatory feature
chr7	79449536	79450313	CACNA2D1	777	DHS-H3K4me3
chr7	79452439	79452940	CACNA2D1	501	DHS
chr7	79483697	79484485	CACNA2D1	788	DHS
chr7	79505917	79506835	CACNA2D1	918	DHS
chr7	79562224	79563160	CACNA2D1	936	DHS
chr7	79565546	79566304	CACNA2D1	758	DHS
chr7	79643721	79644222	CACNA2D1	501	DHS
chr7	79648333	79648967	CACNA2D1	634	DHS-CTCF
chr7	79677174	79677529	CACNA2D1	355	CTCF
chr7	79689533	79689888	CACNA2D1	355	DHS-CTCF
chr7	79692618	79693246	CACNA2D1	628	DHS-H3K4me3
chr7	79699600	79700101	CACNA2D1	501	DHS
chr7	79707294	79708311	CACNA2D1	1,017	DHS-H3K4me3
chr7	79716334	79717557	CACNA2D1	1,223	DHS
chr7	79724873	79725228	CACNA2D1	355	DHS-CTCF
chr7	79728000	79728429	CACNA2D1	429	CTCF
chr7	79762760	79762910	CACNA2D1	150	CTCF
chr7	79763312	79766650	CACNA2D1	3,338	DHS-CTCF-H3K4me3
chr7	79774885	79775859	CACNA2D1	974	DHS
chr7	79776654	79777670	CACNA2D1	1,016	DHS
chr7	79779842	79781241	CACNA2D1	1,399	DHS
chr7	79783846	79784636	CACNA2D1	790	DHS
chr7	79836200	79836350	CACNA2D1	150	CTCF
chr7	79851159	79851786	CACNA2D1	627	DHS
chr7	79873678	79874179	CACNA2D1	501	DHS
chr7	79892481	79892982	CACNA2D1	501	DHS
chr7	79902556	79903190	CACNA2D1	634	DHS
chr7	79906794	79907422	CACNA2D1	628	DHS
chr7	79914104	79914605	CACNA2D1	501	DHS
chr7	79918081	79918450	CACNA2D1	369	DHS-CTCF
chr7	79936689	79937190	CACNA2D1	501	DHS
chr7	79941398	79942025	CACNA2D1	627	DHS
chr7	79964220	79964991	CACNA2D1	771	DHS
chr7	79975937	79976713	CACNA2D1	776	DHS-CTCF
chr7	79986765	79987391	CACNA2D1	626	DHS
chr7	80008284	80008785	CACNA2D1	501	DHS
chr7	80021581	80021936	CACNA2D1	355	DHS
chr7	80057460	80057610	CACNA2D1	150	DHS-CTCF
chr7	80057738	80058239	CACNA2D1	501	DHS-CTCF
chr7	80069412	80070553	CACNA2D1	1,141	DHS
chr7	80083065	80083420	CACNA2D1	355	DHS-CTCF

Chr	Start	End	Locus	Length (bp)	Regulatory feature
chr7	80096153	80096906	CACNA2D1	753	DHS
chr7	80127670	80129231	CACNA2D1	1,561	DHS
chr7	80167532	80168159	CACNA2D1	627	DHS
chr7	80171756	80172385	CACNA2D1	629	DHS
chr7	80214035	80214661	CACNA2D1	626	DHS
chr7	80240916	80241682	CACNA2D1	766	DHS
chr7	80244100	80244794	CACNA2D1	694	DHS
chr7	80251777	80252405	CACNA2D1	628	DHS
chr7	80263538	80264352	CACNA2D1	814	DHS
chr7	80288805	80289210	CACNA2D1	405	DHS-CTCF
chr7	80291075	80291430	CACNA2D1	355	CTCF
chr7	80310589	80310944	CACNA2D1	355	DHS-CTCF
chr7	80328227	80328995	CACNA2D1	768	DHS-CTCF
chr7	80350952	80352898	CACNA2D1	1,946	DHS
chr7	80361973	80362328	CACNA2D1	355	CTCF
chr7	80410361	80411516	CACNA2D1	1,155	DHS
chr7	80456155	80456922	CACNA2D1	767	DHS
chr7	80512523	80513540	CACNA2D1	1,017	DHS
chr7	80529869	80530807	CACNA2D1	938	DHS
chr7	80545860	80546010	CACNA2D1	150	DHS-CTCF
chr7	80546103	80546458	CACNA2D1	355	DHS-CTCF
chr7	80546882	80549309	CACNA2D1	2,427	DHS-H3K4me3
chr7	80549447	80549802	CACNA2D1	355	DHS-CTCF
chr7	80570346	80572407	CACNA2D1	2,061	DHS
chr7	80625320	80625694	CACNA2D1	374	CTCF
chr7	80642282	80643056	CACNA2D1	774	DHS
chr7	80740453	80741373	CACNA2D1	920	DHS
chr7	80790669	80791295	CACNA2D1	626	DHS
chr7	80802904	80803829	CACNA2D1	925	DHS-H3K4me3
chr7	80839730	80841509	CACNA2D1	1,779	DHS
chr7	80873600	80873750	CACNA2D1	150	DHS-CTCF
chr7	80886350	80886705	CACNA2D1	355	DHS-CTCF
chr7	80914217	80914999	CACNA2D1	782	DHS
chr7	80928205	80929217	CACNA2D1	1,012	DHS
chr7	80996317	80997580	CACNA2D1	1,263	DHS
chr7	81059580	81059730	CACNA2D1	150	DHS
chr7	81059885	81060793	CACNA2D1	908	DHS
chr7	81076640	81077141	CACNA2D1	501	DHS-CTCF
chr7	81087234	81087589	CACNA2D1	355	DHS-CTCF
chr7	81099221	81100010	CACNA2D1	789	DHS-CTCF
chr7	81111907	81112262	CACNA2D1	355	DHS-CTCF

Chr	Start	End	Locus	Length (bp)	Regulatory feature
chr7	81113510	81114136	CACNA2D1	626	DHS
chr7	81128664	81129562	CACNA2D1	898	DHS-CTCF
chr7	81130337	81131185	CACNA2D1	848	DHS-CTCF
chr7	81203702	81205151	CACNA2D1	1,449	DHS
chr7	81213171	81214674	CACNA2D1	1,503	DHS
chr7	81224035	81224662	CACNA2D1	627	DHS
chr7	81229679	81230835	CACNA2D1	1,156	DHS
chr7	81233022	81234449	CACNA2D1	1,427	DHS
chr7	81238286	81238787	CACNA2D1	501	DHS
chr7	81246034	81248278	CACNA2D1	2,244	DHS
chr7	81255234	81255861	CACNA2D1	627	DHS
chr7	81256691	81257456	CACNA2D1	765	DHS
chr7	81267137	81269435	CACNA2D1	2,298	DHS
chr7	81314244	81315123	CACNA2D1	879	DHS
chr7	81319680	81320234	CACNA2D1	554	DHS
chr7	81345354	81345709	CACNA2D1	355	DHS-CTCF
chr7	81354115	81355728	CACNA2D1	1,613	DHS
chr7	81391100	81391250	CACNA2D1	150	CTCF
chr7	81391340	81391490	CACNA2D1	150	CTCF
chr7	81392604	81400096	CACNA2D1	7,492	DHS-CTCF-H3K4me3
chr7	81462702	81463203	CACNA2D1	501	DHS
chr7	81474941	81476342	CACNA2D1	1,401	DHS-H3K4me3
chr7	81485020	81485170	CACNA2D1	150	CTCF
chr7	81502608	81503427	CACNA2D1	819	DHS
chr7	81511005	81512317	CACNA2D1	1,312	DHS
chr7	81563719	81564632	CACNA2D1	913	DHS
chr7	81585627	81586398	CACNA2D1	771	DHS-CTCF
chr7	81589293	81589794	CACNA2D1	501	DHS
chr7	81641549	81642324	CACNA2D1	775	DHS
chr7	81646580	81647081	CACNA2D1	501	DHS-CTCF
chr7	81665703	81666330	CACNA2D1	627	DHS
chr7	81668860	81669630	CACNA2D1	770	DHS
chr7	81673810	81674165	CACNA2D1	355	DHS-CTCF
chr7	81678558	81680667	CACNA2D1	2,109	DHS
chr7	81685538	81686188	CACNA2D1	650	DHS
chr7	81686709	81687335	CACNA2D1	626	DHS
chr7	81738469	81739229	CACNA2D1	760	DHS
chr7	81770605	81771231	CACNA2D1	626	DHS
chr7	81794020	81794725	CACNA2D1	705	DHS-CTCF
chr7	81801331	81801686	CACNA2D1	355	DHS-CTCF
chr7	81812922	81813558	CACNA2D1	636	DHS

Chr	Start	End	Locus	Length (bp)	Regulatory feature
chr7	81823320	81823946	CACNA2D1	626	DHS
chr7	81833957	81834584	CACNA2D1	627	DHS
chr7	81863773	81864128	CACNA2D1	355	DHS-CTCF
chr7	81877179	81877534	CACNA2D1	355	DHS-CTCF
chr7	81915440	81915590	CACNA2D1	150	DHS
chr7	81949026	81949664	CACNA2D1	638	DHS
chr7	81965280	81965450	CACNA2D1	170	DHS-CTCF
chr7	81977520	81977670	CACNA2D1	150	DHS
chr7	81991379	81992007	CACNA2D1	628	DHS
chr7	82057464	82058220	CACNA2D1	756	DHS
chr7	82070327	82074411	CACNA2D1	4,084	DHS-CTCF-H3K4me3
chr7	82110825	82111180	CACNA2D1	355	DHS-CTCF
chr7	82135543	82136555	CACNA2D1	1,012	DHS-CTCF
chr7	82151168	82151669	CACNA2D1	501	DHS
chr7	82156824	82158794	CACNA2D1	1,970	DHS
chr7	82159560	82160320	CACNA2D1	760	DHS
chr7	82171813	82172702	CACNA2D1	889	DHS
chr7	82192014	82192901	CACNA2D1	887	DHS
chr7	82201515	82202141	CACNA2D1	626	DHS-CTCF
chr7	82224120	82224270	CACNA2D1	150	DHS
chr7	82228085	82228845	CACNA2D1	760	DHS
chr7	82234577	82234932	CACNA2D1	355	DHS-CTCF
chr7	82253442	82254198	CACNA2D1	756	DHS
chr7	82356131	82356757	CACNA2D1	626	DHS-CTCF
chr7	82358764	82359119	CACNA2D1	355	DHS-CTCF
chr7	82369582	82369937	CACNA2D1	355	DHS-CTCF
chr7	82419482	82420109	CACNA2D1	627	DHS
chr7	82434621	82435530	CACNA2D1	909	DHS-H3K4me3
chr7	82439859	82440487	CACNA2D1	628	DHS-CTCF
chr7	82445680	82445830	CACNA2D1	150	CTCF
chr7	82634900	82635050	CACNA2D1	150	DHS-CTCF
chr7	82702280	82702550	CACNA2D1	270	DHS-CTCF
chr7	82791371	82792749	CACNA2D1	1,378	DHS-H3K4me3
chr7	82900565	82900920	CACNA2D1	355	DHS-H3K4me3
chr7	82977297	82977925	CACNA2D1	628	DHS
chr7	82995080	82995490	CACNA2D1	410	CTCF
chr7	83040327	83040682	CACNA2D1	355	DHS-CTCF
chr7	83042887	83043388	CACNA2D1	501	DHS
chr7	83097412	83098180	CACNA2D1	768	DHS
chr7	83131781	83132282	CACNA2D1	501	DHS
chr7	83143229	83143584	CACNA2D1	355	DHS-CTCF

Chr	Start	End	Locus	Length (bp)	Regulatory feature
chr7	83277726	83278732	CACNA2D1	1,006	DHS-H3K4me3
chr7	83285297	83285652	CACNA2D1	355	DHS-CTCF
chr7	83306402	83307028	CACNA2D1	626	DHS
chr7	83563482	83563930	CACNA2D1	448	CTCF
chr7	83565249	83565878	CACNA2D1	629	DHS-CTCF
chr7	83576566	83576921	CACNA2D1	355	CTCF
chr7	83586593	83587094	CACNA2D1	501	DHS
chr7	83605577	83605932	CACNA2D1	355	DHS-CTCF
chr7	83650624	83651254	CACNA2D1	630	DHS
chr7	83654405	83655336	CACNA2D1	931	DHS
chr7	83671371	83671999	CACNA2D1	628	DHS
chr7	83672905	83673532	CACNA2D1	627	DHS
chr7	83702863	83703489	CACNA2D1	626	DHS
chr7	83721493	83722121	CACNA2D1	628	DHS-H3K4me3
chr7	83730164	83730530	CACNA2D1	366	DHS-CTCF
chr7	83731415	83732041	CACNA2D1	626	DHS
chr7	83818137	83818492	CACNA2D1	355	DHS-CTCF
chr7	83822240	83822390	CACNA2D1	150	CTCF
chr7	83822520	83822670	CACNA2D1	150	CTCF
chr7	83822780	83825981	CACNA2D1	3,201	DHS-H3K4me3
chr7	83847955	83848456	CACNA2D1	501	DHS
chr7	83862160	83862310	CACNA2D1	150	DHS
chr7	83866621	83867397	CACNA2D1	776	DHS
chr7	83877040	83877190	CACNA2D1	150	DHS
chr7	83880659	83881287	CACNA2D1	628	DHS
chr7	83887880	83888381	CACNA2D1	501	DHS
chr7	83900764	83901265	CACNA2D1	501	DHS
chr7	83965479	83967445	CACNA2D1	1,966	DHS
chr7	83968580	83969081	CACNA2D1	501	DHS
chr7	84001240	84002553	CACNA2D1	1,313	DHS
chr7	84041515	84042678	CACNA2D1	1,163	DHS
chr7	84083960	84084110	CACNA2D1	150	CTCF
chr7	84154290	84154916	CACNA2D1	626	DHS
chr7	84158663	84159289	CACNA2D1	626	DHS
chr7	84244139	84244494	CACNA2D1	355	DHS-CTCF
chr7	84267220	84267846	CACNA2D1	626	DHS
chr7	84277091	84277718	CACNA2D1	627	DHS
chr7	84326778	84327279	CACNA2D1	501	DHS
chr7	84330500	84331717	CACNA2D1	1,217	DHS
chr7	84358569	84359070	CACNA2D1	501	DHS
chr7	84420149	84420650	CACNA2D1	501	DHS

Chr	Start	End	Locus	Length (bp)	Regulatory feature
chr7	84569001	84569655	CACNA2D1	654	DHS
chr7	84614880	84615030	CACNA2D1	150	DHS-CTCF
chr7	84628040	84628190	CACNA2D1	150	DHS-CTCF
chr7	84644520	84644670	CACNA2D1	150	CTCF
chr7	84650439	84650794	CACNA2D1	355	DHS-CTCF
chr7	84652408	84652770	CACNA2D1	362	DHS-CTCF
chr7	84660089	84660745	CACNA2D1	656	DHS-CTCF
chr7	84670907	84671408	CACNA2D1	501	DHS-CTCF
chr10	16851336	16851691	CACNB2	355	DHS-CTCF
chr10	16857877	16860639	CACNB2	2,762	DHS-H3K4me3
chr10	16869860	16870010	CACNB2	150	DHS-CTCF
chr10	16870025	16871423	CACNB2	1,398	DHS
chr10	16874644	16875399	CACNB2	755	DHS
chr10	16885102	16885865	CACNB2	763	DHS-CTCF
chr10	16895175	16895803	CACNB2	628	DHS-CTCF
chr10	16929333	16930108	CACNB2	775	DHS
chr10	16933288	16934462	CACNB2	1,174	DHS-CTCF
chr10	16946494	16947391	CACNB2	897	DHS
chr10	16957440	16957590	CACNB2	150	CTCF
chr10	16958320	16958470	CACNB2	150	CTCF
chr10	16961520	16961730	CACNB2	210	DHS-CTCF
chr10	16975980	16976130	CACNB2	150	DHS-CTCF
chr10	16985800	16985950	CACNB2	150	DHS-CTCF
chr10	16986345	16986846	CACNB2	501	DHS
chr10	16992814	16993710	CACNB2	896	DHS
chr10	16994666	16995030	CACNB2	364	DHS-CTCF
chr10	17007930	17011098	CACNB2	3,168	DHS
chr10	17028171	17030622	CACNB2	2,451	DHS
chr10	17034496	17035350	CACNB2	854	DHS-H3K4me3
chr10	17037812	17039383	CACNB2	1,571	DHS
chr10	17041870	17042646	CACNB2	776	DHS
chr10	17043554	17045836	CACNB2	2,282	DHS
chr10	17048672	17050462	CACNB2	1,790	DHS
chr10	17051059	17052071	CACNB2	1,012	DHS
chr10	17064589	17066509	CACNB2	1,920	DHS
chr10	17067315	17068385	CACNB2	1,070	DHS
chr10	17071216	17072423	CACNB2	1,207	DHS
chr10	17075798	17076554	CACNB2	756	DHS
chr10	17080587	17081215	CACNB2	628	DHS
chr10	17098539	17099377	CACNB2	838	DHS
chr10	17101002	17101914	CACNB2	912	DHS

Chr	Start	End	Locus	Length (bp)	Regulatory feature
chr10	17103896	17105089	CACNB2	1,193	DHS
chr10	17117281	17118261	CACNB2	980	DHS
chr10	17155719	17156506	CACNB2	787	DHS
chr10	17157583	17157938	CACNB2	355	DHS-CTCF
chr10	17189975	17190602	CACNB2	627	DHS
chr10	17213067	17213710	CACNB2	643	DHS
chr10	17225109	17225464	CACNB2	355	CTCF
chr10	17241127	17244216	CACNB2	3,089	DHS-H3K4me3
chr10	17248242	17248743	CACNB2	501	DHS
chr10	17256241	17260384	CACNB2	4,143	DHS
chr10	17269091	17277740	CACNB2	8,649	DHS-CTCF-H3K4me3
chr10	17278291	17278792	CACNB2	501	DHS
chr10	17279793	17281585	CACNB2	1,792	DHS
chr10	17291783	17292138	CACNB2	355	DHS
chr10	17299447	17300236	CACNB2	789	DHS
chr10	17332320	17332470	CACNB2	150	CTCF
chr10	17390375	17391472	CACNB2	1,097	DHS-H3K4me3
chr10	17417940	17418344	CACNB2	404	CTCF
chr10	17423432	17424187	CACNB2	755	DHS
chr10	17425829	17426733	CACNB2	904	DHS
chr10	17451169	17451524	CACNB2	355	DHS-CTCF
chr10	17451920	17452070	CACNB2	150	CTCF
chr10	17472673	17473174	CACNB2	501	DHS-CTCF
chr10	17495168	17497095	CACNB2	1,927	DHS-H3K4me3
chr10	17516627	17516982	CACNB2	355	DHS-CTCF
chr10	17546809	17547438	CACNB2	629	DHS
chr10	17554115	17555327	CACNB2	1,212	DHS
chr10	17569316	17570228	CACNB2	912	DHS
chr10	17588952	17589453	CACNB2	501	DHS
chr10	17590703	17591204	CACNB2	501	DHS
chr10	17604396	17605477	CACNB2	1,081	DHS
chr10	17641112	17641467	CACNB2	355	DHS-CTCF
chr10	17649506	17649861	CACNB2	355	DHS-CTCF
chr10	17658350	17660302	CACNB2	1,952	DHS-H3K4me3
chr10	17684834	17687861	CACNB2	3,027	DHS-H3K4me3
chr10	17707559	17708315	CACNB2	756	DHS
chr10	17715406	17716358	CACNB2	952	DHS
chr10	17779375	17779730	CACNB2	355	DHS
chr10	17784866	17785221	CACNB2	355	DHS
chr10	17878784	17879542	CACNB2	758	DHS
chr10	17889696	17890051	CACNB2	355	DHS

Chr	Start	End	Locus	Length (bp)	Regulatory feature
chr10	17914721	17915076	CACNB2	355	DHS
chr10	17952208	17952563	CACNB2	355	DHS
chr10	18026357	18026712	CACNB2	355	DHS
chr10	18031821	18032176	CACNB2	355	DHS
chr10	18076541	18077042	CACNB2	501	DHS
chr10	18125872	18126498	CACNB2	626	DHS
chr10	18136600	18136989	CACNB2	389	CTCF
chr10	18161636	18161991	CACNB2	355	DHS
chr10	18199093	18199448	CACNB2	355	DHS
chr10	18270492	18271143	CACNB2	651	DHS
chr10	18294369	18295157	CACNB2	788	DHS
chr10	18348778	18349826	CACNB2	1,048	DHS
chr10	18393196	18393551	CACNB2	355	DHS-CTCF
chr10	18428994	18430832	CACNB2	1,838	DHS-H3K4me3
chr10	18434337	18435128	CACNB2	791	DHS
chr10	18447669	18448170	CACNB2	501	DHS
chr10	18453292	18454476	CACNB2	1,184	DHS
chr10	18468721	18469347	CACNB2	626	DHS
chr10	18503832	18504333	CACNB2	501	DHS-CTCF
chr10	18505453	18505808	CACNB2	355	DHS-CTCF
chr10	18522740	18522890	CACNB2	150	CTCF
chr10	18527800	18528199	CACNB2	399	DHS-CTCF
chr10	18601674	18602029	CACNB2	355	DHS-CTCF
chr10	18629465	18630356	CACNB2	891	DHS-H3K4me3
chr10	18769560	18769710	CACNB2	150	CTCF
chr10	18786728	18787488	CACNB2	760	DHS
chr10	18811771	18812581	CACNB2	810	DHS
chr10	18872501	18872856	CACNB2	355	DHS-CTCF
chr10	18883724	18884079	CACNB2	355	CTCF
chr10	18890420	18890570	CACNB2	150	CTCF
chr10	18896381	18897008	CACNB2	627	DHS
chr10	18899276	18900044	CACNB2	768	DHS
chr10	18905260	18905624	CACNB2	364	DHS-CTCF
chr10	18912505	18913132	CACNB2	627	DHS
chr10	18916215	18917240	CACNB2	1,025	DHS-H3K4me3
chr10	18938400	18941308	CACNB2	2,908	DHS-H3K4me3
chr10	18946753	18950341	CACNB2	3,588	DHS-H3K4me3
chr10	18955751	18956377	CACNB2	626	DHS
chr10	18970091	18971527	CACNB2	1,436	DHS
chr10	18973412	18974041	CACNB2	629	DHS
chr10	18978428	18978783	CACNB2	355	DHS-CTCF

Chr	Start	End	Locus	Length (bp)	Regulatory feature
chr10	18988600	18988750	CACNB2	150	CTCF
chr10	19001843	19002615	CACNB2	772	DHS
chr10	19082980	19083130	CACNB2	150	CTCF
chr10	19457660	19457810	CACNB2	150	CTCF
chr10	19614394	19615021	CACNB2	627	DHS
chr10	19636528	19636883	CACNB2	355	DHS-CTCF
chr10	19699980	19700130	CACNB2	150	CTCF
chr10	19701472	19701827	CACNB2	355	DHS-CTCF
chr10	19777575	19778624	CACNB2	1,049	DHS-CTCF-H3K4me3
chr10	19787293	19787794	CACNB2	501	DHS
chr10	19909740	19910138	CACNB2	398	DHS-CTCF
chr10	19918140	19919231	CACNB2	1,091	DHS-CTCF
chr10	19927088	19928358	CACNB2	1,270	DHS-H3K4me3
chr10	19940441	19940942	CACNB2	501	DHS
chr10	19943475	19943976	CACNB2	501	DHS
chr10	19952440	19952590	CACNB2	150	CTCF
chr10	19972965	19973410	CACNB2	445	DHS-CTCF
chr10	20002706	20003061	CACNB2	355	DHS-CTCF
chr10	20008014	20008942	CACNB2	928	DHS
chr10	20019010	20019637	CACNB2	627	DHS
chr10	20037868	20038914	CACNB2	1,046	DHS
chr10	20074217	20074718	CACNB2	501	DHS
chr10	20077030	20077385	CACNB2	355	DHS-CTCF
chr10	20100410	20100911	CACNB2	501	DHS-CTCF
chr10	20103930	20108755	CACNB2	4,825	DHS-CTCF-H3K4me3
chr10	20115564	20116190	CACNB2	626	DHS
chr10	20117187	20118116	CACNB2	929	DHS
chr10	20132720	20132870	CACNB2	150	DHS
chr10	20136196	20136823	CACNB2	627	DHS
chr10	20207452	20208210	CACNB2	758	DHS
chr10	20210180	20212068	CACNB2	1,888	DHS
chr10	20218202	20219085	CACNB2	883	DHS
chr10	20220234	20221213	CACNB2	979	DHS
chr10	20259337	20260899	CACNB2	1,562	DHS
chr10	20282735	20283616	CACNB2	881	DHS
chr10	20292673	20293676	CACNB2	1,003	DHS
chr10	20328915	20329416	CACNB2	501	DHS
chr10	20389040	20389666	CACNB2	626	DHS
chr10	20394691	20395474	CACNB2	783	DHS
chr10	20399520	20399670	CACNB2	150	CTCF
chr10	20836820	20836970	CACNB2	150	CTCF

Chr	Start	End	Locus	Length (bp)	Regulatory feature
chr10	20887483	20887984	CACNB2	501	DHS
chr10	21057796	21058151	CACNB2	355	DHS-CTCF
chr10	21144366	21144721	CACNB2	355	DHS-CTCF
chr10	21147217	21147572	CACNB2	355	DHS-CTCF
chr11	116642324	116644855	SCN2B	2,531	DHS-H3K4me3
chr11	116653414	116653915	SCN2B	501	DHS
chr11	116657281	116659700	SCN2B	2,419	DHS-H3K4me3
chr11	116661640	116661790	SCN2B	150	DHS-CTCF
chr11	116661892	116662519	SCN2B	627	DHS
chr11	116679404	116680690	SCN2B	1,286	DHS
chr11	116699700	116699850	SCN2B	150	CTCF
chr11	116700130	116700756	SCN2B	626	DHS-CTCF
chr11	116706341	116707371	SCN2B	1,030	DHS-CTCF
chr11	116711019	116711374	SCN2B	355	DHS
chr11	116722955	116723582	SCN2B	627	DHS
chr11	116724135	116725028	SCN2B	893	DHS
chr11	116732081	116732857	SCN2B	776	DHS
chr11	116741765	116742540	SCN2B	775	DHS
chr11	116744781	116745546	SCN2B	765	DHS-CTCF
chr11	116750496	116751122	SCN2B	626	DHS
chr11	116758599	116759226	SCN2B	627	DHS-CTCF
chr11	116765805	116766160	SCN2B	355	DHS
chr11	116781030	116781531	SCN2B	501	DHS
chr11	116783667	116784293	SCN2B	626	CTCF
chr11	116797898	116798253	SCN2B	355	DHS-H3K4me3
chr11	116800304	116801542	SCN2B	1,238	DHS
chr11	116854942	116855992	SCN2B	1,050	DHS
chr11	116857097	116857901	SCN2B	804	DHS
chr11	116860576	116861202	SCN2B	626	DHS
chr11	116881368	116882683	SCN2B	1,315	DHS
chr11	116897419	116898184	SCN2B	765	DHS
chr11	116916440	116917225	SCN2B	785	DHS
chr11	116929412	116930038	SCN2B	626	DHS
chr11	116940526	116941305	SCN2B	779	DHS
chr11	116942039	116943988	SCN2B	1,949	DHS
chr11	116952700	116952850	SCN2B	150	CTCF
chr11	116952920	116953070	SCN2B	150	CTCF
chr11	116953935	116954692	SCN2B	757	DHS
chr11	116966180	116970173	SCN2B	3,993	DHS-CTCF-H3K4me3
chr11	116976631	116977132	SCN2B	501	DHS-CTCF
chr11	117005841	117006342	SCN2B	501	DHS-CTCF

Chr	Start	End	Locus	Length (bp)	Regulatory feature
chr11	117009081	117010042	SCN2B	961	DHS-CTCF
chr11	117014051	117017764	SCN2B	3,713	DHS-CTCF-H3K4me3
chr11	117031428	117032055	SCN2B	627	DHS
chr11	117042188	117044436	SCN2B	2,248	DHS
chr11	117048653	117051455	SCN2B	2,802	DHS-H3K4me3
chr11	117051760	117052488	SCN2B	728	DHS-CTCF
chr11	117068100	117074486	SCN2B	6,386	DHS-CTCF-H3K4me3
chr11	117076260	117076657	SCN2B	397	DHS-CTCF
chr11	117077459	117077814	SCN2B	355	DHS-CTCF
chr11	117079458	117081050	SCN2B	1,592	DHS-CTCF
chr11	117087401	117089865	SCN2B	2,464	DHS
chr11	117090937	117093501	SCN2B	2,564	DHS
chr11	117097088	117097720	SCN2B	632	DHS-CTCF
chr11	117101124	117104615	SCN2B	3,491	DHS-CTCF-H3K4me3
chr11	117109995	117110622	SCN2B	627	DHS-CTCF
chr11	117122280	117122430	SCN2B	150	CTCF
chr11	117130103	117130735	SCN2B	632	DHS
chr11	117144365	117145117	SCN2B	752	DHS
chr11	117150319	117150953	SCN2B	634	DHS
chr11	117151699	117152591	SCN2B	892	DHS
chr11	117170956	117172087	SCN2B	1,131	DHS
chr11	117184782	117187690	SCN2B	2,908	DHS-CTCF-H3K4me3
chr11	117190068	117190694	SCN2B	626	DHS
chr11	117198053	117200235	SCN2B	2,182	DHS-H3K4me3
chr11	117260071	117260881	SCN2B	810	DHS
chr11	117296899	117297254	SCN2B	355	DHS-CTCF
chr11	117313680	117313830	SCN2B	150	CTCF
chr11	117314243	117314598	SCN2B	355	DHS-CTCF
chr11	117350160	117350310	SCN2B	150	CTCF
chr11	117489820	117489970	SCN2B	150	CTCF
chr11	117491960	117492315	SCN2B	355	CTCF
chr11	117502140	117502290	SCN2B	150	CTCF
chr11	117601100	117601250	SCN2B	150	CTCF
chr11	117678187	117678542	SCN2B	355	DHS-CTCF
chr11	117678720	117678870	SCN2B	150	DHS-CTCF
chr11	117680000	117680378	SCN2B	378	DHS-CTCF
chr11	117680460	117680610	SCN2B	150	DHS-CTCF
chr11	117688193	117690352	SCN2B	2,159	DHS-CTCF
chr11	117692736	117693091	SCN2B	355	CTCF
chr11	117693900	117694130	SCN2B	230	CTCF
chr11	117694366	117696090	SCN2B	1,724	DHS

Chr	Start	End	Locus	Length (bp)	Regulatory feature
chr11	117699215	117699843	SCN2B	628	DHS
chr11	117714590	117715217	SCN2B	627	DHS-CTCF
chr11	117746689	117748025	SCN2B	1,336	DHS-H3K4me3
chr11	117756330	117757099	SCN2B	769	DHS-H3K4me3
chr11	117775291	117775646	SCN2B	355	DHS-CTCF
chr11	117777885	117778511	SCN2B	626	DHS
chr11	117816180	117816330	SCN2B	150	CTCF
chr11	117817326	117818129	SCN2B	803	DHS-CTCF
chr11	117821080	117821350	SCN2B	270	CTCF
chr11	117856772	117857524	SCN2B	752	DHS-H3K4me3
chr11	117873287	117873913	SCN2B	626	DHS
chr11	117877160	117877935	SCN2B	775	DHS
chr11	117923907	117924541	SCN2B	634	DHS-CTCF
chr11	117939160	117939310	SCN2B	150	CTCF
chr11	118015929	118017064	SCN2B	1,135	DHS-CTCF-H3K4me3
chr11	118023083	118024441	SCN2B	1,358	DHS-H3K4me3
chr11	118027304	118027659	SCN2B	355	CTCF
chr11	118042349	118042704	SCN2B	355	DHS-CTCF
chr11	118046920	118047070	SCN2B	150	DHS
chr11	118065800	118066301	SCN2B	501	DHS
chr11	118068600	118068750	SCN2B	150	DHS-CTCF
chr11	118068988	118069343	SCN2B	355	DHS-CTCF
chr11	118081671	118082297	SCN2B	626	DHS
chr11	118122501	118123515	SCN2B	1,014	DHS-H3K4me3
chr11	118129160	118129538	SCN2B	378	CTCF
chr11	118134605	118135501	SCN2B	896	DHS-H3K4me3
chr11	118145860	118146505	SCN2B	645	DHS
chr11	118161580	118162027	SCN2B	447	DHS-CTCF
chr11	118166560	118166915	SCN2B	355	DHS-CTCF
chr11	118186260	118186410	SCN2B	150	CTCF
chr11	118187059	118187875	SCN2B	816	DHS
chr11	118223700	118224175	SCN2B	475	DHS-CTCF
chr11	118229720	118231473	SCN2B	1,753	DHS-H3K4me3
chr11	118269474	118270808	SCN2B	1,334	DHS
chr11	118271094	118274014	SCN2B	2,920	DHS-H3K4me3
chr11	118286159	118286952	SCN2B	793	DHS
chr11	118304846	118310106	SCN2B	5,260	DHS-H3K4me3
chr11	118356040	118356210	SCN2B	170	CTCF
chr11	118356220	118356370	SCN2B	150	CTCF
chr11	118359101	118359730	SCN2B	629	DHS-CTCF
chr11	118393116	118394021	SCN2B	905	DHS-CTCF

Chr	Start	End	Locus	Length (bp)	Regulatory feature
chr11	118395000	118395150	SCN2B	150	CTCF
chr11	118400932	118402591	SCN2B	1,659	DHS-H3K4me3
chr11	118435844	118437081	SCN2B	1,237	DHS-CTCF-H3K4me3
chr11	118442316	118446044	SCN2B	3,728	DHS-H3K4me3
chr11	118465707	118466334	SCN2B	627	DHS-CTCF
chr11	118472663	118473710	SCN2B	1,047	DHS
chr11	118475665	118477120	SCN2B	1,455	DHS
chr11	118477864	118482446	SCN2B	4,582	DHS-CTCF-H3K4me3
chr11	118483015	118483811	SCN2B	796	DHS
chr11	118488551	118490880	SCN2B	2,329	DHS-H3K4me3
chr11	118491516	118494259	SCN2B	2,743	DHS-H3K4me3
chr11	118505132	118506063	SCN2B	931	DHS
chr11	118511480	118511630	SCN2B	150	CTCF
chr11	118529500	118529650	SCN2B	150	CTCF
chr11	118529911	118530809	SCN2B	898	DHS-CTCF
chr11	118543576	118543931	SCN2B	355	DHS-CTCF
chr11	118549633	118549988	SCN2B	355	DHS-CTCF
chr11	118559940	118561477	SCN2B	1,537	DHS-CTCF-H3K4me3
chr11	118567740	118567890	SCN2B	150	DHS
chr11	118585965	118586594	SCN2B	629	DHS
chr11	118590877	118591503	SCN2B	626	DHS
chr11	118641057	118641685	SCN2B	628	DHS
chr11	118659275	118664156	SCN2B	4,881	DHS-CTCF-H3K4me3
chr11	118740463	118741789	SCN2B	1,326	DHS-CTCF
chr11	118758182	118759190	SCN2B	1,008	DHS
chr11	118760445	118760946	SCN2B	501	DHS
chr11	118770200	118770350	SCN2B	150	CTCF
chr11	118772198	118772553	SCN2B	355	CTCF
chr11	118777607	118778363	SCN2B	756	DHS-CTCF
chr11	118779480	118783592	SCN2B	4,112	DHS-H3K4me3
chr11	118787164	118797267	SCN2B	10,103	DHS-CTCF-H3K4me3
chr11	118798535	118801374	SCN2B	2,839	DHS-CTCF-H3K4me3
chr11	118827016	118827817	SCN2B	801	DHS-CTCF
chr11	118850859	118851749	SCN2B	890	DHS-H3K4me3
chr11	118867523	118870065	SCN2B	2,542	DHS-H3K4me3
chr11	118886060	118886210	SCN2B	150	CTCF
chr11	118887372	118891264	SCN2B	3,892	DHS-H3K4me3
chr11	118900553	118902178	SCN2B	1,625	DHS-CTCF-H3K4me3
chr11	118915540	118915730	SCN2B	190	CTCF
chr11	118926398	118929250	SCN2B	2,852	DHS-CTCF-H3K4me3
chr11	118933467	118933968	SCN2B	501	DHS

Chr	Start	End	Locus	Length (bp)	Regulatory feature
chr11	118936133	118936488	SCN2B	355	DHS-CTCF
chr11	118937320	118939773	SCN2B	2,453	DHS-H3K4me3
chr11	118955051	118956962	SCN2B	1,911	DHS-H3K4me3
chr11	118963598	118966806	SCN2B	3,208	DHS-CTCF-H3K4me3
chr11	118970965	118973928	SCN2B	2,963	DHS-CTCF-H3K4me3
chr11	118977212	118979915	SCN2B	2,703	DHS-H3K4me3
chr11	118991138	118993558	SCN2B	2,420	DHS-H3K4me3
chr11	119001338	119001964	SCN2B	626	DHS
chr11	119005109	119005464	SCN2B	355	CTCF
chr11	119015520	119016149	SCN2B	629	DHS-CTCF
chr11	119016180	119016330	SCN2B	150	DHS-CTCF
chr11	119019308	119021074	SCN2B	1,766	H3K4me3
chr11	119024340	119024490	SCN2B	150	DHS-CTCF
chr11	119026843	119027198	SCN2B	355	DHS-CTCF
chr11	119038928	119040635	SCN2B	1,707	DHS-CTCF-H3K4me3
chr11	119066356	119068122	SCN2B	1,766	DHS-CTCF
chr11	119076031	119078602	SCN2B	2,571	DHS-CTCF-H3K4me3
chr11	119088992	119089766	SCN2B	774	DHS
chr11	119133344	119133699	SCN2B	355	DHS-CTCF
chr11	119153035	119153790	SCN2B	755	DHS
chr11	119155278	119155633	SCN2B	355	DHS-CTCF
chr11	119165940	119166090	SCN2B	150	DHS-CTCF
chr11	119177072	119177573	SCN2B	501	DHS-CTCF
chr11	119178140	119178290	SCN2B	150	CTCF
chr11	119185798	119188534	SCN2B	2,736	DHS-H3K4me3
chr11	119188985	119194864	SCN2B	5,879	DHS-CTCF-H3K4me3
chr11	119204704	119206626	SCN2B	1,922	DHS-H3K4me3
chr11	119208249	119212117	SCN2B	3,868	DHS-CTCF-H3K4me3
chr11	119223552	119224053	SCN2B	501	DHS
chr11	119227016	119227930	SCN2B	914	DHS-CTCF-H3K4me3
chr11	119231686	119232503	SCN2B	817	DHS
chr11	119234626	119235256	SCN2B	630	DHS-H3K4me3
chr11	119242600	119242750	SCN2B	150	CTCF
chr11	119244775	119245276	SCN2B	501	DHS-CTCF
chr11	119246126	119246899	SCN2B	773	DHS
chr11	119251691	119253197	SCN2B	1,506	DHS-CTCF-H3K4me3
chr11	119260586	119260941	SCN2B	355	CTCF
chr11	119287480	119287650	SCN2B	170	CTCF
chr11	119287860	119288945	SCN2B	1,085	DHS
chr11	119292038	119294524	SCN2B	2,486	DHS-H3K4me3
chr11	119297953	119298582	SCN2B	629	DHS

Chr	Start	End	Locus	Length (bp)	Regulatory feature
chr11	119303056	119304370	SCN2B	1,314	DHS-H3K4me3
chr11	119311729	119312484	SCN2B	755	DHS
chr11	119331234	119331993	SCN2B	759	DHS
chr11	119340978	119341737	SCN2B	759	DHS
chr11	119345140	119345290	SCN2B	150	CTCF
chr11	119345328	119345963	SCN2B	635	DHS-CTCF
chr11	119346080	119346230	SCN2B	150	DHS-CTCF
chr11	119350380	119350751	SCN2B	371	DHS-CTCF
chr11	119351761	119352539	SCN2B	778	DHS-CTCF
chr11	119356160	119356310	SCN2B	150	DHS-CTCF
chr11	119362466	119362821	SCN2B	355	DHS-CTCF
chr11	119380495	119381450	SCN2B	955	DHS
chr11	119404453	119405216	SCN2B	763	DHS-CTCF
chr11	119438578	119439615	SCN2B	1,037	DHS
chr11	119443880	119444030	SCN2B	150	CTCF
chr11	119454276	119456392	SCN2B	2,116	DHS-H3K4me3
chr11	119465358	119465989	SCN2B	631	DHS
chr11	119469288	119469915	SCN2B	627	DHS
chr11	119473656	119474011	SCN2B	355	CTCF
chr11	119486090	119486850	SCN2B	760	DHS
chr11	119494460	119494610	SCN2B	150	CTCF
chr11	119494740	119495105	SCN2B	365	CTCF
chr11	119512700	119513098	SCN2B	398	CTCF
chr11	119536991	119537619	SCN2B	628	DHS-CTCF
chr11	119541536	119542941	SCN2B	1,405	DHS
chr11	119552243	119552598	SCN2B	355	DHS
chr11	119555016	119555517	SCN2B	501	DHS
chr11	119562053	119562410	SCN2B	357	CTCF
chr11	119577646	119578001	SCN2B	355	CTCF
chr11	119580172	119580976	SCN2B	804	DHS
chr11	119591796	119592151	SCN2B	355	DHS-CTCF
chr11	119597808	119601352	SCN2B	3,544	DHS-CTCF-H3K4me3
chr11	119611980	119612357	SCN2B	377	CTCF
chr11	119612705	119613801	SCN2B	1,096	DHS-H3K4me3
chr11	119645100	119645250	SCN2B	150	DHS
chr11	119662355	119663117	SCN2B	762	DHS
chr11	119665882	119666533	SCN2B	651	DHS
chr11	119853577	119854205	SCN2B	628	DHS
chr11	119869968	119870323	SCN2B	355	CTCF
chr11	119883840	119884310	SCN2B	470	CTCF
chr11	119938624	119939395	SCN2B	771	DHS

Chr	Start	End	Locus	Length (bp)	Regulatory feature
chr11	119950501	119950856	SCN2B	355	CTCF
chr11	119978941	119979296	SCN2B	355	DHS-CTCF
chr11	119983280	119983430	SCN2B	150	CTCF
chr11	119986300	119986672	SCN2B	372	CTCF
chr11	120022096	120022860	SCN2B	764	DHS
chr11	120039317	120040602	SCN2B	1,285	DHS-CTCF-H3K4me3
chr11	120042767	120043426	SCN2B	659	DHS-CTCF
chr11	120051392	120051893	SCN2B	501	DHS-CTCF
chr11	120053540	120053901	SCN2B	361	DHS-CTCF
chr11	120055582	120056466	SCN2B	884	DHS
chr11	120076220	120076721	SCN2B	501	DHS
chr11	120079412	120080038	SCN2B	626	DHS
chr11	120080725	120084177	SCN2B	3,452	DHS-H3K4me3
chr11	120086067	120086883	SCN2B	816	DHS
chr11	120099680	120099950	SCN2B	270	DHS-CTCF
chr11	120100538	120100893	SCN2B	355	DHS-CTCF
chr11	120105733	120106088	SCN2B	355	DHS-CTCF
chr11	120110900	120111361	SCN2B	461	CTCF
chr11	120129500	120129650	SCN2B	150	CTCF
chr11	120142927	120143553	SCN2B	626	DHS
chr11	120170580	120171081	SCN2B	501	DHS
chr11	120173668	120174169	SCN2B	501	DHS
chr11	122471958	122472755	SCN3B	797	DHS
chr11	122476079	122476705	SCN3B	626	DHS
chr11	122499371	122500183	SCN3B	812	DHS-CTCF
chr11	122503905	122504554	SCN3B	649	DHS-CTCF
chr11	122512950	122513853	SCN3B	903	DHS
chr11	122517625	122518253	SCN3B	628	DHS
chr11	122519800	122520301	SCN3B	501	DHS
chr11	122524466	122525232	SCN3B	766	DHS-CTCF
chr11	122526031	122528044	SCN3B	2,013	DHS-H3K4me3
chr11	122561313	122562454	SCN3B	1,141	DHS-H3K4me3
chr11	122568191	122568962	SCN3B	771	DHS
chr11	122579826	122580625	SCN3B	799	DHS
chr11	122582483	122583189	SCN3B	706	DHS-CTCF
chr11	122592238	122593157	SCN3B	919	DHS
chr11	122602126	122602879	SCN3B	753	DHS
chr11	122603849	122604730	SCN3B	881	DHS
chr11	122612120	122613069	SCN3B	949	DHS
chr11	122619437	122620366	SCN3B	929	DHS
chr11	122625284	122625639	SCN3B	355	CTCF

Chr	Start	End	Locus	Length (bp)	Regulatory feature
chr11	122625860	122626010	SCN3B	150	CTCF
chr11	122626043	122626679	SCN3B	636	DHS
chr11	122627260	122627410	SCN3B	150	DHS
chr11	122643574	122644202	SCN3B	628	DHS
chr11	122649818	122650319	SCN3B	501	DHS
chr11	122652030	122652658	SCN3B	628	DHS-CTCF
chr11	122655007	122655830	SCN3B	823	DHS
chr11	122655980	122656130	SCN3B	150	CTCF
chr11	122659511	122659866	SCN3B	355	DHS-CTCF
chr11	122666760	122667203	SCN3B	443	DHS
chr11	122678906	122679712	SCN3B	806	CTCF
chr11	122685112	122685467	SCN3B	355	DHS
chr11	122687585	122688339	SCN3B	754	DHS
chr11	122699868	122700369	SCN3B	501	DHS
chr11	122720700	122721428	SCN3B	728	DHS-CTCF
chr11	122724439	122725203	SCN3B	764	DHS-H3K4me3
chr11	122727088	122728019	SCN3B	931	DHS-H3K4me3
chr11	122734448	122734850	SCN3B	402	DHS-CTCF
chr11	122753221	122754743	SCN3B	1,522	DHS-H3K4me3
chr11	122768468	122769094	SCN3B	626	DHS
chr11	122774940	122775090	SCN3B	150	DHS
chr11	122781514	122781869	SCN3B	355	DHS-CTCF
chr11	122852345	122852972	SCN3B	627	DHS
chr11	122854449	122855907	SCN3B	1,458	H3K4me3
chr11	122891709	122892064	SCN3B	355	DHS-CTCF
chr11	122893841	122894196	SCN3B	355	DHS-CTCF
chr11	122894460	122894610	SCN3B	150	CTCF
chr11	122904528	122904883	SCN3B	355	DHs-CTCF
chr11	122918175	122918934	SCN3B	759	DHS
chr11	122924820	122925596	SCN3B	776	DHS-CTCF
chr11	122928500	122928650	SCN3B	150	DHS
chr11	122929045	122935064	SCN3B	6,019	DHS-H3K4me3
chr11	122936115	122936939	SCN3B	824	DHS
chr11	122953925	122954857	SCN3B	932	DHS
chr11	122971034	122971922	SCN3B	888	DHS
chr11	122973894	122974520	SCN3B	626	DHS-CTCF
chr11	122981940	122983437	SCN3B	1,497	DHS-CTCF
chr11	122987036	122987677	SCN3B	641	DHS
chr11	122989043	122989827	SCN3B	784	DHS
chr11	122992735	122993974	SCN3B	1,239	DHS
chr11	123000506	123001007	SCN3B	501	DHS

Chr	Start	End	Locus	Length (bp)	Regulatory feature
chr11	123006576	123007203	SCN3B	627	DHS
chr11	123007814	123009030	SCN3B	1,216	DHS
chr11	123014150	123014651	SCN3B	501	DHS
chr11	123015686	123016613	SCN3B	927	DHS
chr11	123034460	123034610	SCN3B	150	CTCF
chr11	123036220	123036370	SCN3B	150	CTCF
chr11	123036676	123037732	SCN3B	1,056	DHS-CTCF
chr11	123038759	123039528	SCN3B	769	DHS
chr11	123043804	123046099	SCN3B	2,295	DHS
chr11	123050638	123051527	SCN3B	889	DHS
chr11	123058347	123060165	SCN3B	1,818	DHS-H3K4me3
chr11	123062827	123067075	SCN3B	4,248	DHS-H3K4me3
chr11	123068272	123069153	SCN3B	881	DHS
chr11	123072595	123073801	SCN3B	1,206	DHS
chr11	123077318	123077947	SCN3B	629	DHS
chr11	123091595	123092636	SCN3B	1,041	DHS
chr11	123101930	123102558	SCN3B	628	DHS
chr11	123105824	123107304	SCN3B	1,480	DHS-CTCF
chr11	123108827	123109717	SCN3B	890	DHS
chr11	123117669	123118433	SCN3B	764	DHS-CTCF
chr11	123122079	123122984	SCN3B	905	DHS
chr11	123131927	123132718	SCN3B	791	DHS-CTCF
chr11	123171805	123173403	SCN3B	1,598	DHS-CTCF-H3K4me3
chr11	123173520	123173750	SCN3B	230	DHS
chr11	123175740	123175890	SCN3B	150	CTCF
chr11	123178624	123179427	SCN3B	803	DHS
chr11	123186492	123187253	SCN3B	761	DHS
chr11	123228540	123229671	SCN3B	1,131	DHS-H3K4me3
chr11	123277954	123278580	SCN3B	626	DHS-CTCF
chr11	123298723	123299477	SCN3B	754	DHS
chr11	123300446	123302976	SCN3B	2,530	DHS-H3K4me3
chr11	123305014	123305369	SCN3B	355	DHS-CTCF
chr11	123321127	123322015	SCN3B	888	DHS
chr11	123324625	123325448	SCN3B	823	DHS
chr11	123327724	123328514	SCN3B	790	DHS
chr11	123339506	123340550	SCN3B	1,044	DHS
chr11	123344596	123346854	SCN3B	2,258	DHS
chr11	123349002	123349628	SCN3B	626	DHS-CTCF
chr11	123354142	123354643	SCN3B	501	DHS
chr11	123361140	123361290	SCN3B	150	CTCF
chr11	123380280	123380906	SCN3B	626	DHS

Chr	Start	End	Locus	Length (bp)	Regulatory feature
chr11	123381825	123382455	SCN3B	630	DHS-CTCF
chr11	123416945	123417904	SCN3B	959	DHS
chr11	123425623	123425978	SCN3B	355	DHS-CTCF
chr11	123429340	123429490	SCN3B	150	DHS-CTCF
chr11	123430124	123432454	SCN3B	2,330	DHS-CTCF-H3K4me3
chr11	123447712	123448067	SCN3B	355	DHS-CTCF
chr11	123451279	123451634	SCN3B	355	DHS-CTCF
chr11	123451840	123451990	SCN3B	150	CTCF
chr11	123460860	123461662	SCN3B	802	DHS
chr11	123468620	123468770	SCN3B	150	CTCF
chr11	123471604	123472514	SCN3B	910	DHS
chr11	123486992	123487762	SCN3B	770	DHS
chr11	123495672	123496312	SCN3B	640	DHS
chr11	123499759	123500521	SCN3B	762	DHS
chr11	123511670	123512025	SCN3B	355	DHS-CTCF
chr11	123524415	123526046	SCN3B	1,631	DHS-H3K4me3
chr11	123567000	123567150	SCN3B	150	CTCF
chr11	123578211	123578712	SCN3B	501	DHS-CTCF
chr11	123581940	123582441	SCN3B	501	DHS-CTCF
chr11	123610840	123613439	SCN3B	2,599	DHS-H3K4me3
chr11	123844340	123844998	SCN3B	658	DHS
chr11	123940041	123940981	SCN3B	940	DHS-CTCF-H3K4me3
chr11	123941904	123942540	SCN3B	636	DHS-CTCF
chr11	123942680	123942830	SCN3B	150	CTCF
chr11	123946597	123947494	SCN3B	897	H3K4me3
chr11	123978857	123979212	SCN3B	355	DHS-CTCF
chr11	123985624	123987404	SCN3B	1,780	DHS-H3K4me3
chr11	123994300	123994450	SCN3B	150	DHS
chr11	124044644	124045145	SCN3B	501	DHS-CTCF
chr11	124065538	124066301	SCN3B	763	DHS
chr11	124146001	124146627	SCN3B	626	DHS
chr11	124155611	124156362	SCN3B	751	DHS-CTCF
chr11	124271983	124272836	SCN3B	853	DHS-CTCF
chr11	124285237	124285592	SCN3B	355	DHS-CTCF
chr11	124312323	124312678	SCN3B	355	DHS
chr11	124337219	124338132	SCN3B	913	DHS
chr11	124340212	124341243	SCN3B	1,031	DHS-H3K4me3
chr11	124351129	124352442	SCN3B	1,313	DHS-CTCF
chr11	124405613	124407033	SCN3B	1,420	DHS
chr11	124407608	124408656	SCN3B	1,048	DHS
chr11	124448426	124449224	SCN3B	798	DHS-CTCF

Chr	Start	End	Locus	Length (bp)	Regulatory feature
chr11	124456094	124456595	SCN3B	501	CTCF
chr11	124491964	124494318	SCN3B	2,354	DHS-H3K4me3
chr11	124503286	124503641	SCN3B	355	DHS-CTCF
chr11	124513233	124514998	SCN3B	1,765	DHS-CTCF
chr11	124539015	124539370	SCN3B	355	DHS-CTCF
chr11	124542452	124544685	SCN3B	2,233	DHS-H3K4me3
chr11	124546087	124546879	SCN3B	792	DHS
chr11	124587410	124588730	SCN3B	1,320	DHS-H3K4me3
chr11	124595413	124596319	SCN3B	906	DHS
chr11	124609227	124610866	SCN3B	1,639	DHS-CTCF-H3K4me3
chr11	124611060	124611210	SCN3B	150	CTCF
chr11	124615243	124617445	SCN3B	2,202	DHS-CTCF-H3K4me3
chr11	124621644	124622426	SCN3B	782	H3K4me3
chr11	124628139	124630128	SCN3B	1,989	DHS-CTCF-H3K4me3
chr11	124631404	124633420	SCN3B	2,016	DHS-CTCF-H3K4me3
chr11	124639757	124640384	SCN3B	627	DHS
chr11	124648209	124648564	SCN3B	355	CTCF
chr11	124663597	124663952	SCN3B	355	DHS-CTCF
chr11	124668552	124671606	SCN3B	3,054	DHS-H3K4me3
chr11	124706189	124707959	SCN3B	1,770	DHS-CTCF
chr11	124708991	124710110	SCN3B	1,119	H3K4me3
chr11	124712976	124713477	SCN3B	501	H3K4me3
chr11	124733081	124733711	SCN3B	630	DHS
chr11	124735770	124736551	SCN3B	781	CTCF
chr11	124736960	124737250	SCN3B	290	CTCF
chr11	124737360	124737510	SCN3B	150	CTCF
chr11	124737629	124737984	SCN3B	355	CTCF
chr11	124738220	124738370	SCN3B	150	CTCF
chr11	124738880	124739030	SCN3B	150	CTCF
chr11	124745527	124747981	SCN3B	2,454	DHS-CTCF-H3K4me3
chr11	124767939	124769643	SCN3B	1,704	DHS
chr11	124770210	124771007	SCN3B	797	DHS
chr11	124786769	124787543	SCN3B	774	DHS
chr11	124790619	124791758	SCN3B	1,139	DHS-H3K4me3
chr11	124804588	124805475	SCN3B	887	DHS
chr11	124813280	124813430	SCN3B	150	DHS-H3K4me3
chr11	124823307	124824840	SCN3B	1,533	DHS-CTCF
chr11	124831858	124832484	SCN3B	626	DHS
chr11	124840940	124841090	SCN3B	150	DHS
chr11	124906697	124907198	SCN3B	501	DHS-H3K4me3
chr11	124932391	124934033	SCN3B	1,642	DHS

Chr	Start	End	Locus	Length (bp)	Regulatory feature
chr11	124940390	124941147	SCN3B	757	DHS
chr11	124952851	124953477	SCN3B	626	DHS-CTCF
chr11	124961057	124962071	SCN3B	1,014	DHS-H3K4me3
chr11	124980724	124982406	SCN3B	1,682	CTCF
chr11	124984456	124984811	SCN3B	355	DHS-CTCF
chr11	124993991	124994510	SCN3B	519	DHS
chr11	124999660	124999810	SCN3B	150	DHS-CTCF
chr11	125011300	125011450	SCN3B	150	DHS-CTCF
chr11	125011620	125012002	SCN3B	382	DHS-CTCF
chr11	125033818	125041313	SCN3B	7,495	DHS-CTCF-H3K4me3
chr11	125062022	125062649	SCN3B	627	DHS
chr11	125081437	125081792	SCN3B	355	CTCF
chr11	125084546	125085348	SCN3B	802	DHS
chr11	125094422	125095049	SCN3B	627	DHS-CTCF
chr11	125110832	125111583	SCN3B	751	DHS
chr11	125115026	125115783	SCN3B	757	DHS
chr11	125126880	125127030	SCN3B	150	DHS-CTCF
chr11	125127119	125127474	SCN3B	355	CTCF
chr11	125132588	125134069	SCN3B	1,481	DHS
chr11	125145675	125146314	SCN3B	639	DHS
chr11	125157255	125157890	SCN3B	635	DHS
chr11	125170282	125171342	SCN3B	1,060	DHS
chr11	125177640	125177790	SCN3B	150	CTCF
chr11	125184996	125185799	SCN3B	803	DHS
chr11	125208580	125209782	SCN3B	1,202	DHS
chr11	125213461	125213962	SCN3B	501	DHS
chr11	125215555	125216623	SCN3B	1,068	DHS-CTCF
chr11	125218260	125218410	SCN3B	150	DHS-CTCF
chr11	125218660	125219015	SCN3B	355	DHS-CTCF
chr11	125225133	125225488	SCN3B	355	DHS
chr11	125234460	125234730	SCN3B	270	DHS
chr11	125243428	125244197	SCN3B	769	DHS
chr11	125260558	125261494	SCN3B	936	DHS-CTCF
chr11	125274493	125275276	SCN3B	783	CTCF
chr11	125293147	125293502	SCN3B	355	DHS-CTCF
chr11	125298600	125298750	SCN3B	150	DHS-CTCF
chr11	125299040	125299395	SCN3B	355	DHS-CTCF
chr11	125303352	125303707	SCN3B	355	DHS-CTCF
chr11	125322443	125323069	SCN3B	626	DHS-CTCF
chr11	125327722	125328477	SCN3B	755	DHS
chr11	125364138	125366835	SCN3B	2,697	DHS-H3K4me3

Chr	Start	End	Locus	Length (bp)	Regulatory feature
chr11	125379071	125379924	SCN3B	853	DHS
chr11	125384109	125385160	SCN3B	1,051	DHS
chr11	125386943	125387796	SCN3B	853	DHS-H3K4me3
chr11	125438835	125440695	SCN3B	1,860	DHS-CTCF-H3K4me3
chr11	125461074	125464031	SCN3B	2,957	DHS-H3K4me3
chr11	125473700	125473850	SCN3B	150	CTCF
chr11	125494859	125497791	SCN3B	2,932	DHS-CTCF-H3K4me3
chr11	125549349	125549704	SCN3B	355	CTCF
chr11	125594026	125594381	SCN3B	355	DHS-CTCF
chr11	125620560	125620710	SCN3B	150	CTCF
chr11	125642212	125643922	SCN3B	1,710	DHS
chr11	125739460	125739690	SCN3B	230	DHS
chr11	125741331	125741686	SCN3B	355	DHS-CTCF
chr11	125743594	125744911	SCN3B	1,317	DHS-CTCF-H3K4me3
chr11	125756865	125758861	SCN3B	1,996	DHS-H3K4me3
chr11	125772659	125773627	SCN3B	968	DHS-H3K4me3
chr11	125773860	125774150	SCN3B	290	DHS-H3K4me3
chr11	125774260	125774410	SCN3B	150	DHS-H3K4me3
chr12	1098728	1101858	CACNA1C	3,130	DHS-H3K4me3
chr12	1123019	1123662	CACNA1C	643	DHS
chr12	1175442	1175797	CACNA1C	355	DHS-CTCF
chr12	1192898	1193399	CACNA1C	501	DHS
chr12	1202226	1202854	CACNA1C	628	DHS-CTCF
chr12	1227112	1227613	CACNA1C	501	DHS-CTCF
chr12	1313806	1314560	CACNA1C	754	DHS-H3K4me3
chr12	1390876	1391654	CACNA1C	778	DHS
chr12	1415171	1415797	CACNA1C	626	DHS
chr12	1427672	1428667	CACNA1C	995	DHS
chr12	1538793	1539735	CACNA1C	942	DHS
chr12	1553033	1553811	CACNA1C	778	DHS-H3K4me3
chr12	1572354	1573666	CACNA1C	1,312	DHS
chr12	1579093	1579720	CACNA1C	627	DHS
chr12	1583020	1584606	CACNA1C	1,586	DHS
chr12	1598025	1598656	CACNA1C	631	DHS
chr12	1626282	1626637	CACNA1C	355	DHS-CTCF
chr12	1628140	1628910	CACNA1C	770	DHS
chr12	1638890	1640965	CACNA1C	2,075	H3K4me3
chr12	1642300	1642450	CACNA1C	150	DHS
chr12	1654150	1654816	CACNA1C	666	DHS-CTCF
chr12	1684856	1686649	CACNA1C	1,793	DHS
chr12	1686920	1687070	CACNA1C	150	CTCF

Chr	Start	End	Locus	Length (bp)	Regulatory feature
chr12	1687120	1687270	CACNA1C	150	CTCF
chr12	1692539	1693040	CACNA1C	501	DHS-CTCF
chr12	1699099	1705044	CACNA1C	5,945	DHS-H3K4me3
chr12	1714132	1715789	CACNA1C	1,657	DHS-CTCF-H3K4me3
chr12	1717340	1718247	CACNA1C	907	DHS
chr12	1720772	1721655	CACNA1C	883	DHS
chr12	1731895	1732521	CACNA1C	626	DHS
chr12	1734328	1735103	CACNA1C	775	DHS
chr12	1738605	1740156	CACNA1C	1,551	DHS-CTCF-H3K4me3
chr12	1740440	1740590	CACNA1C	150	DHS
chr12	1743699	1744338	CACNA1C	639	DHS-CTCF
chr12	1753560	1754317	CACNA1C	757	DHS
chr12	1759148	1759930	CACNA1C	782	DHS
chr12	1761932	1762559	CACNA1C	627	DHS-CTCF
chr12	1769341	1769842	CACNA1C	501	DHS
chr12	1770435	1772579	CACNA1C	2,144	DHS-H3K4me3
chr12	1772780	1772930	CACNA1C	150	DHS-CTCF
chr12	1775540	1775690	CACNA1C	150	CTCF
chr12	1777923	1778695	CACNA1C	772	DHS
chr12	1799177	1801651	CACNA1C	2,474	DHS-H3K4me3
chr12	1805204	1805705	CACNA1C	501	DHS
chr12	1820587	1821603	CACNA1C	1,016	DHS
chr12	1828243	1828869	CACNA1C	626	DHS
chr12	1879009	1879510	CACNA1C	501	DHS
chr12	1885563	1886330	CACNA1C	767	DHS
chr12	1886934	1887435	CACNA1C	501	DHS
chr12	1888782	1889539	CACNA1C	757	DHS
chr12	1900220	1900721	CACNA1C	501	DHS-CTCF
chr12	1905039	1907016	CACNA1C	1,977	DHS-CTCF-H3K4me3
chr12	1909791	1910146	CACNA1C	355	DHS-CTCF
chr12	1913452	1914241	CACNA1C	789	DHS
chr12	1915397	1916023	CACNA1C	626	DHS
chr12	1918990	1919888	CACNA1C	898	DHS
chr12	1920775	1922595	CACNA1C	1,820	DHS
chr12	1940141	1940776	CACNA1C	635	DHS
chr12	1960140	1960960	CACNA1C	820	DHS
chr12	1973949	1974603	CACNA1C	654	H3K4me3
chr12	2036819	2037740	CACNA1C	921	DHS
chr12	2049071	2049572	CACNA1C	501	DHS-CTCF
chr12	2056188	2056816	CACNA1C	628	DHS
chr12	2102080	2102230	CACNA1C	150	DHS

Chr	Start	End	Locus	Length (bp)	Regulatory feature
chr12	2102563	2103804	CACNA1C	1,241	DHS
chr12	2112285	2114334	CACNA1C	2,049	DHS-H3K4me3
chr12	2118820	2118970	CACNA1C	150	DHS-CTCF
chr12	2123140	2123290	CACNA1C	150	DHS
chr12	2143916	2144762	CACNA1C	846	DHS-CTCF-H3K4me3
chr12	2145460	2145710	CACNA1C	250	CTCF
chr12	2158910	2159688	CACNA1C	778	DHS
chr12	2160945	2164778	CACNA1C	3,833	DHS-CTCF-H3K4me3
chr12	2166373	2166728	CACNA1C	355	DHS-CTCF
chr12	2172994	2173785	CACNA1C	791	DHS-H3K4me3
chr12	2185078	2186001	CACNA1C	923	DHS-H3K4me3
chr12	2193540	2194295	CACNA1C	755	DHS
chr12	2262906	2263707	CACNA1C	801	DHS
chr12	2270472	2271270	CACNA1C	798	DHS
chr12	2271988	2273561	CACNA1C	1,573	DHS
chr12	2279274	2280156	CACNA1C	882	DHS
chr12	2289150	2290301	CACNA1C	1,151	DHS
chr12	2292945	2293973	CACNA1C	1,028	DHS-CTCF
chr12	2307268	2307895	CACNA1C	627	DHS
chr12	2325084	2325897	CACNA1C	813	DHS
chr12	2339187	2339814	CACNA1C	627	DHS-CTCF-H3K4me3
chr12	2340257	2341015	CACNA1C	758	DHS-CTCF-H3K4me3
chr12	2353031	2353793	CACNA1C	762	DHS
chr12	2365000	2365230	CACNA1C	230	DHS-CTCF
chr12	2365513	2366572	CACNA1C	1,059	DHS
chr12	2367123	2367880	CACNA1C	757	DHS
chr12	2368526	2369152	CACNA1C	626	DHS
chr12	2373425	2375287	CACNA1C	1,862	DHS
chr12	2376612	2377379	CACNA1C	767	DHS
chr12	2378048	2378674	CACNA1C	626	DHS
chr12	2382518	2382873	CACNA1C	355	DHS-CTCF
chr12	2391520	2391670	CACNA1C	150	DHS-CTCF
chr12	2392672	2394627	CACNA1C	1,955	DHS-CTCF
chr12	2397129	2397630	CACNA1C	501	DHS
chr12	2400369	2400724	CACNA1C	355	DHS-CTCF
chr12	2403505	2405525	CACNA1C	2,020	DHS
chr12	2441358	2441859	CACNA1C	501	DHS-CTCF
chr12	2445300	2446246	CACNA1C	946	DHS
chr12	2448170	2448671	CACNA1C	501	DHS
chr12	2450830	2451592	CACNA1C	762	DHS
chr12	2467116	2468095	CACNA1C	979	DHS

Chr	Start	End	Locus	Length (bp)	Regulatory feature
chr12	2468260	2468410	CACNA1C	150	CTCF
chr12	2469540	2470096	CACNA1C	556	DHS-CTCF
chr12	2491610	2491965	CACNA1C	355	CTCF
chr12	2493720	2494631	CACNA1C	911	DHS
chr12	2504542	2504897	CACNA1C	355	DHS
chr12	2505275	2506321	CACNA1C	1,046	CTCF
chr12	2515817	2516318	CACNA1C	501	CTCF
chr12	2603402	2604169	CACNA1C	767	DHS-CTCF
chr12	2635540	2635690	CACNA1C	150	DHS
chr12	2657420	2657570	CACNA1C	150	CTCF
chr12	2692160	2692574	CACNA1C	414	DHS-CTCF
chr12	2697560	2697710	CACNA1C	150	CTCF
chr12	2722557	2723444	CACNA1C	887	DHS
chr12	2724620	2725429	CACNA1C	809	DHS
chr12	2733668	2734023	CACNA1C	355	CTCF
chr12	2749320	2749470	CACNA1C	150	CTCF
chr12	2749580	2750709	CACNA1C	1,129	DHS-CTCF
chr12	2783540	2783957	CACNA1C	417	CTCF
chr12	2791760	2792436	CACNA1C	676	DHS-CTCF
chr12	2800023	2801497	CACNA1C	1,474	DHS-CTCF-H3K4me3
chr12	2802560	2802710	CACNA1C	150	CTCF
chr12	2842800	2842950	CACNA1C	150	CTCF
chr12	2848833	2849476	CACNA1C	643	DHS-CTCF
chr12	2861696	2862335	CACNA1C	639	DHS
chr12	2892912	2893691	CACNA1C	779	DHS-CTCF
chr12	2903149	2906035	CACNA1C	2,886	DHS-H3K4me3
chr12	2906833	2907593	CACNA1C	760	DHS-CTCF
chr12	2907700	2907850	CACNA1C	150	CTCF
chr12	2913454	2914879	CACNA1C	1,425	DHS
chr12	2921184	2923272	CACNA1C	2,088	DHS-CTCF-H3K4me3
chr12	2943792	2944682	CACNA1C	890	DHS-CTCF-H3K4me3
chr12	2954262	2955761	CACNA1C	1,499	DHS-CTCF
chr12	2962695	2963323	CACNA1C	628	DHS-CTCF
chr12	2985398	2988088	CACNA1C	2,690	DHS-H3K4me3
chr12	2994408	2994763	CACNA1C	355	DHS-CTCF
chr12	2999423	3001369	CACNA1C	1,946	DHS-H3K4me3
chr12	3002099	3002600	CACNA1C	501	DHS
chr12	3011365	3011992	CACNA1C	627	DHS
chr12	3053508	3054119	CACNA1C	611	DHS-CTCF
chr12	3067132	3070730	CACNA1C	3,598	DHS-H3K4me3
chr12	3073151	3073788	CACNA1C	637	DHS-H3K4me3

Chr	Start	End	Locus	Length (bp)	Regulatory feature
chr12	3107853	3108619	CACNA1C	766	DHS
chr12	3109729	3110084	CACNA1C	355	CTCF
chr12	3113856	3114211	CACNA1C	355	CTCF
chr12	3143071	3143827	CACNA1C	756	DHS-CTCF
chr12	3145456	3145811	CACNA1C	355	DHS-CTCF
chr12	3150646	3151273	CACNA1C	627	DHS-CTCF
chr12	3155586	3156087	CACNA1C	501	DHS
chr12	3182571	3183644	CACNA1C	1,073	DHS-CTCF
chr12	3185634	3187880	CACNA1C	2,246	DHS-H3K4me3
chr12	3206135	3207381	CACNA1C	1,246	DHS-H3K4me3
chr12	3225887	3227382	CACNA1C	1,495	DHS
chr12	3228685	3229575	CACNA1C	890	DHS
chr12	3232016	3233054	CACNA1C	1,038	DHS-CTCF
chr12	3236940	3237569	CACNA1C	629	DHS
chr12	3238481	3239107	CACNA1C	626	DHS
chr12	3242105	3242734	CACNA1C	629	DHS
chr12	3244573	3245213	CACNA1C	640	DHS-CTCF
chr12	3246332	3247085	CACNA1C	753	DHS
chr12	3260023	3261584	CACNA1C	1,561	DHS
chr12	3277645	3278413	CACNA1C	768	DHS
chr12	3279812	3280564	CACNA1C	752	DHS
chr12	3294020	3294170	CACNA1C	150	DHS
chr12	3308399	3310497	CACNA1C	2,098	DHS-H3K4me3
chr12	3311260	3311410	CACNA1C	150	CTCF
chr12	3311520	3311710	CACNA1C	190	CTCF
chr12	3312029	3312530	CACNA1C	501	DHS-CTCF
chr12	3312780	3312930	CACNA1C	150	CTCF
chr12	3317220	3317370	CACNA1C	150	CTCF
chr12	3320718	3321219	CACNA1C	501	CTCF
chr12	3326540	3328408	CACNA1C	1,868	DHS
chr12	3345444	3346515	CACNA1C	1,071	DHS
chr12	3361408	3362080	CACNA1C	672	DHS
chr12	3371180	3371330	CACNA1C	150	DHS-CTCF
chr12	3384495	3385131	CACNA1C	636	DHS-CTCF
chr12	3402868	3403494	CACNA1C	626	DHS
chr12	3404813	3405442	CACNA1C	629	DHS
chr12	3408520	3408670	CACNA1C	150	CTCF
chr12	3408984	3410071	CACNA1C	1,087	DHS-CTCF
chr12	3413060	3413210	CACNA1C	150	DHS
chr12	3413417	3414186	CACNA1C	769	DHS
chr12	3422770	3423125	CACNA1C	355	DHS-CTCF

Chr	Start	End	Locus	Length (bp)	Regulatory feature
chr12	3423380	3423530	CACNA1C	150	CTCF
chr12	3424020	3424170	CACNA1C	150	CTCF
chr12	3424233	3424985	CACNA1C	752	DHS
chr12	3451200	3451555	CACNA1C	355	DHS-CTCF
chr12	3474915	3475671	CACNA1C	756	H3K4me3
chr12	3632340	3632690	CACNA1C	350	CTCF
chr12	3636292	3636793	CACNA1C	501	DHS
chr12	3663580	3663730	CACNA1C	150	CTCF
chr12	3750477	3751266	CACNA1C	789	DHS
chr12	3764689	3765044	CACNA1C	355	DHS-CTCF
chr12	3767553	3767908	CACNA1C	355	DHS-CTCF
chr12	3768140	3768290	CACNA1C	150	CTCF
chr12	3776780	3776930	CACNA1C	150	CTCF
chr12	3790440	3790827	CACNA1C	387	DHS-CTCF
chr12	3792829	3794454	CACNA1C	1,625	DHS
chr12	3795820	3795970	CACNA1C	150	CTCF
chr12	3807236	3808250	CACNA1C	1,014	DHS
chr12	3813640	3814681	CACNA1C	1,041	DHS
chr12	3837229	3837982	CACNA1C	753	DHS
chr12	3852761	3853571	CACNA1C	810	DHS
chr12	3853820	3853970	CACNA1C	150	CTCF
chr12	3861130	3863223	CACNA1C	2,093	DHS-CTCF-H3K4me3
chr12	3886395	3886750	CACNA1C	355	DHS-CTCF
chr12	3906853	3907208	CACNA1C	355	CTCF
chr12	3912983	3913484	CACNA1C	501	DHS-CTCF

Chr (chromosome)

Annex 4

Table A-4. List of 59 CTCF-overlapping variants. The genomic position, the reference and alternative alleles and the number of individuals sharing the variant and the genomic coordinates of the CTCF binding site containing each variant is also shown. The position of the variant relative to CTCF motif is also shown.

Chr	Pos	Ref	Alt	Individuals	CTCF binding site			Position relative to motif
chr3	37953719	T	A	4	chr3	37953703	37953739	Non-Core
chr3	38045036	T	A	2	chr3	38045012	38045048	Core
chr3	38521504	A	G	1	chr3	38521503	38521539	Far
chr3	38780342	G	A	1	chr3	38780314	38780350	Closer
chr3	39540276	G	A	1	chr3	39540241	39540277	Far
chr3	39854224	A	G	2	chr3	39854209	39854245	Non-Core
chr3	39953594	C	T	69	chr3	39953589	39953625	Far
chr7	79677353	C	T	3	chr7	79677352	79677388	Far
chr7	80288990	ATGT	A	53	chr7	80288957	80288993	Closer
chr7	80549633	C	G	1	chr7	80549620	80549656	Core
chr7	80625526	G	A	54	chr7	80625506	80625542	Non-Core
chr7	81076865	TC	T	1	chr7	81076848	81076884	Non-Core
chr7	81076870	A	T	1	chr7	81076848	81076884	Non-Core
chr7	81087389	T	C	46	chr7	81087385	81087421	Far
chr7	81130725	C	T	5	chr7	81130717	81130753	Closer
chr7	82110992	T	C	66	chr7	82110987	82111023	Far
chr7	82702421	G	A	1	chr7	82702419	82702455	Far
chr7	82900741	T	G	5	chr7	82900710	82900746	Closer
chr10	18883940	C	T	61	chr10	18883906	18883942	Closer
chr10	19457714	C	G	1	chr10	19457680	19457716	Closer
chr10	21147418	G	A	30	chr10	21147390	21147426	Non-Core
chr11	117122398	A	T	1	chr11	117122391	117122427	Closer
chr11	117924278	C	T	75	chr11	117924277	117924313	Far
chr11	118042498	C	T	2	chr11	118042497	118042533	Far

Chr	Pos	Ref	Alt	Individuals	CTCF binding site			Position relative to motif
chr11	118549803	C	T	42	chr11	118549791	118549827	Core
chr11	118560953	G	GC	37	chr11	118560938	118560974	Core
chr11	118886117	C	T	43	chr11	118886087	118886123	Closer
chr11[+]	118889370[+]	C	G	1	chr11	118889348	118889384	Non-Core
chr11	118889378	G	A	49	chr11	118889348	118889384	Closer
chr11	119287550	G	A	2	chr11	119287516	119287552	Far
chr11	119352239	G	A	4	chr11	119352232	119352268	Closer
chr11	119404719	C	T	2	chr11	119404707	119404743	Non-Core
chr11	119600183	A	T	2	chr11	119600179	119600215	Far
chr11	119612173	C	T	3	chr11	119612154	119612190	Non-Core
chr11	120053724	C	T	1	chr11	120053708	120053744	Core
chr11	120100704	T	G	1	chr11	120100681	120100717	Core
chr11	120173911	C	CTGAAG	9	chr11	120173897	120173933	Non-Core
chr11	120173914	C	CGGG	9	chr11	120173895	120173931	Non-Core
chr11	120173917	C	CT	9	chr11	120173897	120173933	Core
chr11	122659691	T	C	64	chr11	122659686	122659722	Far
chr11	123036887	G	C	1	chr11	123036869	123036905	Non-Core
chr11	123105986	G	T	2	chr11	123105981	123106017	Far
chr11	123118102	CT	C	1	chr11	123118082	123118118	Core
chr11	123132370	G	C	1	chr11	123132363	123132399	Closer
chr11	123425811	A	C	1	chr11	123425778	123425814	Closer
chr11	123447866	A	G	1	chr11	123447863	123447899	Far
chr11	123511861	A	AC	2	chr11	123511827	123511863	Far
chr11	123511862	C	T	1	chr11	123511827	123511863	Far
chr11	124539211	G	A	1	chr11	124539191	124539227	Non-Core
chr11	124648367	T	G	1	chr11	124648344	124648380	Core

Chr	Pos	Ref	Alt	Individuals	CTCF binding site			Position relative to motif
chr11	124984629	G	T	7	chr11	124984600	124984636	Non-Core
chr12	2469918	T	A	1	chr12	2469908	2469944	Core
chr12	2733860	C	T	1	chr12	2733844	2733880	Non-Core
chr12	3053956	G	A	1	chr12	3053941	3053977	Non-Core
chr12	3384802	G	A	1	chr12	3384773	3384809	Non-Core
chr12	3451418	C	T	1	chr12	3451394	3451430	Non-Core
chr12	3764916	G	C	21	chr12	3764884	3764920	Closer
chr12	3767720	G	T	20	chr12	3767697	3767733	Core
chr12	3913211	G	A	2	chr12	3913209	3913245	Far

[+]Variant that could not be tested in luciferase reporter assays (see methods section 2.11).

Chr (chromosome), Pos (position), Ref (reference allele), Alt (alternative allel). For more details regarding the position relative to motif see Results **Figure 86**.