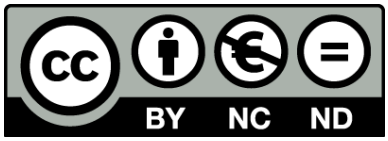
UNIVERSITAT DE BARCELONA

Implementation of a novel analytical framework for large-scale genetic data

Extending the genetic architecture of type 2 diabetes beyond common variants

Sílvia Bonàs Guarch

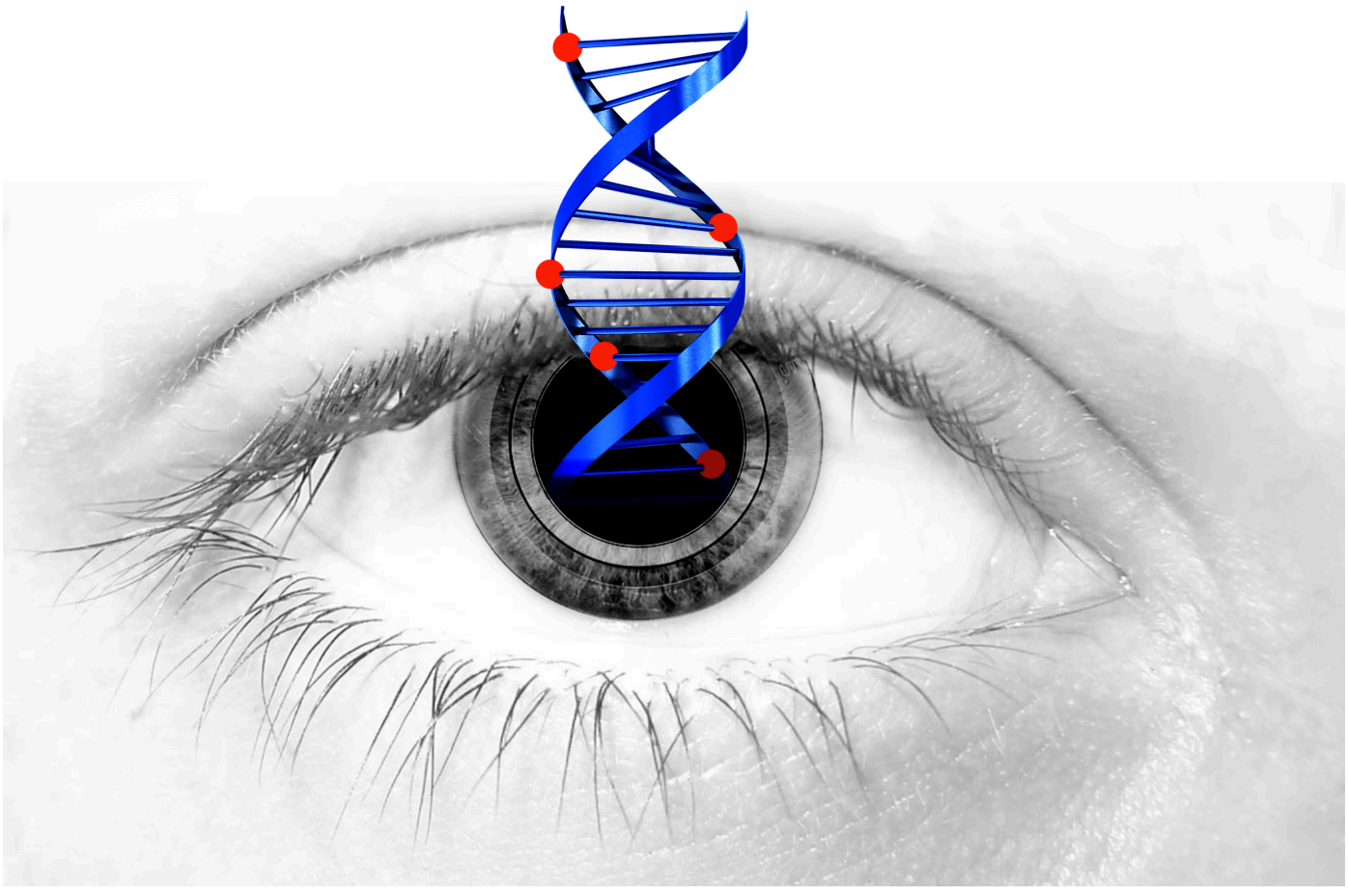


Implementation of a novel analytical framework for large-scale genetic data

Extending the genetic architecture of type 2 diabetes beyond common variants

Silvia Bonàs Guarch

Directed by Dr. Josep M. Mercader and Dr. David Torrents



Facultat de Biologia, Universitat de Barcelona

Implementation of a novel analytical framework for large-scale genetic data

Extending the genetic architecture of type 2 diabetes beyond common variants

Barcelona Supercomputing Center-Centro Nacional de Supercomputación

Programa de Biomedicina EEES H0101

Memòria presentada per Sílvia Bonàs i Guarch per optar al grau de doctora per la Universitat de Barcelona

Doctoranda
Sílvia Bonàs i Guarch

Directors
Josep M. Mercader i David Torrents

Tutor
Modesto Orozco

We're like crystal
We break easy
I'm a poor man
If you leave me
I'm applauded
Then forgotten
It was summer
Now it's autumn

Crystal, New Order

Agraïments

Sóc conscient que el desenvolupament d'aquesta tesi no hauria estat possible sense la interacció amb una heterogeneïtat de cares, a vegades anecdòtica i en altres casos esdevenint maratonianes travessies científiques, que intentaré treure de l'anonimat.

En primer lloc m'agradaria agrair al David Torrents, director del meu grup, haver-me permès realitzar el meu projecte de màster i la tesis al Barcelona Supercomputing Center (BSC). Li vull agrair la seva confiança total i el seu optimisme estoic el qual ha hagut de navegar amb el meu pessimisme a vegades ridícul. M'agradaria també agrair-li múltiples converses i sobretot, la seva empatia, que poc a poc, m'han permès vertebrar quin havia de ser el meu lloc dins el món de la ciència.

En segon lloc, m'agradaria estendre'm amb en Txema, el qual ha sigut el post-doc darrera de cada coma d'aquest treball. Si he arribat mai a tenir alguna visió del que vull aportar dins aquest camp és gràcies a la seva perseverança i confiança en cada pas que he donat. Li vull agrair el seu esforç i el seu inconformisme per anar guanyant posicions en un camp complex i del qual sempre me n'ha fet partícip en cada projecte que ha batallat. Li vull agrair les incomptables oportunitats de formar part en projectes crucials, que m'han permès apuntalar la meva carrera científica. M'agradaria també agrair-li el seu esforç perquè pogués enriquir-me amb l'estada del Broad Institute, en un entorn excel·lent per acabar de madurar com a doctorant. I bàsicament li vull agrair tots els seus consells, que han estat com una petita guia durant tot aquest temps. A nivell personal, crec que cada una de les seves crítiques, e-mails impredictibles, la seva devoció per la feina, cada petit consell, i sí, en algun cas discussions, m'han ajudat a enfrontar-me a tot de pors residuals que sempre m'han inutilitzat. Un dels abismes que sempre he tingut és acabar el doctorat i no estar a l'alçada del que per mi ha estat en Txema com a post-doc.

Cada una de les persones del meu grup també mereix un capítol a part. Però primer m'agradaria dedicar un petit espai a en Brian, la primera persona amb qui vaig establir una conversa al BSC. A en Brian li voldria agrair la seva sinceritat excepcional i estic molt orgullosa del vincle final que hem establert, pel qual ha quedat palès el *Quantenverschränkung*. A continuació, m'agradaria agrair a la Marta Guindo la seva paciència, i molta que n'ha necessitat, per treballar amb mi, la seva motivació infinita i el plaer diàfan que m'ha aportat haver fet un "Team GWAS" amb ella. A la Mercè, tota l'ajuda que sempre m'ha ofert, absolutament immerescuda: des d'una sol·licitud per marxar a Boston a un anàlisi a última hora. Persones com ella són rareses. M'agradaria referir-me a l'Elias Fos, amb el qual va ser un gust treballar en el seu projecte de màster, que ha estat vital per impulsar parts del nostre article principal, i que és un tertulià amb cuir i gas inesgotable. Però per sobre de tot, m'agradaria agrair a tots ells el fet que mai m'ha fet mandra treballar al BSC. És pràcticament impossible si estàs rodejat del seu humor i la seva inacabable energia. M'agradaria també agrair a la Montse tota la seva paciència quan ha de gestionar tots els desastres d'espai que li generem.

Sempre abstrets en els nostres projectes, i ens oblidem que hi ha una persona vetllant perquè tot tiri milles... I no m'oblido de la Paula Cortés, última en arribar però amb la qual hem treballat intensament, a vegades tornant-la boja entre test i test d'associació. Tinc però molt presents cada un dels dinars al Central Park, gràcies i "abajo el fajismo"! Finalment, a tot el departament de Life Science, un plaer.

Durant aquests anys he tingut la oportunitat de col·laborar amb diferents centres de recerca. M'agradaria agrair la confiança que va depositar en el nostre treball el consorci COPSAC i en Klaus Bonnelyke en especial. Aquesta col·laboració ha estat fonamental i ens ha permès optar a molts altres projectes de gran rellevància. Per exemple, ha estat un plaer treballar amb el grup del Jorge Ferrer, una col·laboració que des del primer dia ha estat absolutament enriquidora i esperem que continuï així! D'altra banda, també m'agradaria fer un agraïment al Jose C. Florez, optimista i amb capacitat de lluita excepcionals. Gràcies per ajudar-nos a mirar fites que al principi creia impossibles, i per la oportunitat que em va donar de fer una estada al Broad Institute, com també tot l'entorn i l'aprenentatge que vaig tenir la sort de gaudir.

M'agradaria fer un agraïment per uns amics en especial. A la Casandra, amb la qual hem compartit carreres científiques paral·leles però que ha estat un pulmó per anar trampejant la carrera d'obstacles en què es converteix un doctorat. Un agraïment peculiar és el d'en Miquel. Després de molts anys, encara em té fascinada com degusta com un esmorzar de forquilla la meva mediocritat, que pel que sembla genera expectatives. M'agradaria agrair també a la Marina la seva capacitat per recordar-me que hi ha vida més enllà del doctorat, amb la seva particular però familiar visió de la realitat. D'altra banda, mil gràcies Joel Cockburn per oferir-me el millor nom de guerra que mai tindré, Sílvia Menguele. Ah, i Cristian, gràcies per aguantar-me i no deixar-me respirar, perquè mai perdés les forces i ganes pel camí!

Per acabar, vull agrair als meus pares tot el seu recolzament incondicional, els quals sempre han d'aguantar totes les fílies i fòbies que van aflorant quan les situacions es torcen. Aquesta confiança irreductible la tinc absolutament present. I per acabar, a en Jon, que com en l'esgrima, ha restat pacient amb cada dificultat que hem viscut per donar l'estocada precisa i tranquil·la. El qual ha redescobert, en totes les dimensions inimaginables, qui és la Sílvia.

Mercès a tots.

Abbreviations

1000G,	The 1000 Genomes Project
1000G-Phase1,	The 1000 Genomes Project Release 1
1000G-Phase3,	The 1000 Genomes Project Release 3
2hrGlu,	2 hours after a glucose oral challenge
58C,	1958 British Birth cohort
ABF,	Approximate Baye's Factor
ADA,	American Diabetes Association
Affy,	Affymetrix-array
AMD,	Age-related Macular Degeneration
BMI,	Body Mass Index
CADD,	Combined Annotation Dependent Depletion scoring function
CDCV,	Common Disease Common Variant Model
ChIP-seq,	Chromatin Immunoprecipitation sequencing
CI,	Confidence Interval
CNV,	Copy Number Variant
CNP,	Copy Neutral Polymorphism
COMPSs,	COMP Superscalar
dbGaP,	database of Genotypes and Phentoypes
DEPICT,	Data-driven Expression Prioritized Integration for Complex Traits
d.f,	Degrees of freedom
DIAGRAM,	Diabetes Genetics Replication and Meta-analysis consortium
DM,	Diabetes Mellitus
DNA,	Deoxyribonucleic Acid
DZ,	Dizygotic
EEA,	European Environment Agency
EGA,	European Genome-phenome Archive
ENCODE,	Encyclopedia of DNA Elements Project Consortium
eQTL,	expression Quantitative Trait Locus
ExAC,	Exome Aggregation Consortium
FAO,	Fatty Acid Oxidation

FDR, False Discovery Rate

FFA, Free Fatty Acid

FG, Fasting Glucose

FP, False Positive association

FPG, Fasting Plasma Glucose

GENEVA, Genes and Environment Initiatives in Type 2 Diabetes

GERA, Genetic Epidemiology Research on Adult Health and Aging

GoNL, The Genome of the Netherlands

GT, Genotyped

GTEx, The Genotype-Tissue Expression project

GWAS, Genome-Wide Association Studies

HbA1c, Haemoglobin A1c

HGP, Human Genome Project

HLA, Human Leukocyte Antigen

HMM, Hidden Markov Models

HPC, High Performance Computing

HR, Cox-proportional Hazards Ratio

HRC, Haplotype Reference Consortium

HWE, Hardy-Weinberg Equilibrium

IBD, Identity-By-Descendent

IBS, Identify-By-State

ICD, International Statistical Classification of Diseases and Related Health Problems

IL, Illumina-array

INDEL, Insertion and Deletion

K, Thousand

kb, kilobase pair

LD, Linkage Disequilibrium

LRP, Long-Range Phasing

LSF, Load Sharing Facility

M, Million

Mb, Megabase pair

MAF,	Minor Allele Frequency
MAOS,	Meta-Analysis with Overlapping Subjects
MDS,	MultiDimensional Scaling
MHC,	Major Histocompatibility Complex
MODY,	Maturity-Onset Diabetes of the Young
MPI,	Message Passing Interface
MZ,	Monozygotic
NBS,	UK Blood Service Control Group
NCD,	Non-Communicable Disease
NDDG,	National Diabetes Data Group
NDM,	Neonatal Diabetes Mellitus
NGS,	Next Generation Sequencing
NGT,	Normal Glucose Tolerance
NIH,	National Institutes of Health
NuGENE,	NuGENE Northwestern
OGTT,	Oral Glucose Tolerance Test
OMIM,	The Online Mendelian Inheritance in Man
OpenMP,	Open Multi-Processing
OR,	Odds Ratio
P4,	Predictive, Preventive, Personalized, Participatory
PC,	Principal/New Components
PCA,	Principal Component Analysis
PNDM,	Permanent Neonatal Diabetes Mellitus
QC,	Quality Control
Q-Q,	Quantile-Quantile
Roadmap,	NIH Roadmap Epigenomics Mapping Consortium
RFLP,	Restriction Fragment Length Polymorphism
ROC,	Receiving Operating Characteristic
RPGEH,	Research Program on Genes, Environment, and Health
SE,	Standard Error
SIGMA,	Slim Initiative in Genomic Medicine for the Americas
SLURM,	Simple Linux Utility for Resource Management

SNP,	Single Nucleotide Polymorphism
SNV,	Single Nucleotide Variant
SSMP,	Singapore Sequencing Malay Project
SV,	Structural Variant
T1D,	Type 1 diabetes
T2D,	Type 2 diabetes
T2D Portal,	Type 2 Diabetes Knowledge Portal
TNDM,	Transient Neonatal Diabetes Mellitus
UK10K,	UK10K Consortium
VEP,	Variant Effect Predictor
VMI,	Virtual Machine Image
WES,	Whole-Exome Sequencing
WGS,	Whole-Genome Sequencing
WHO,	World Health Organization
WTCCC,	Wellcome Trust Case Control Consortium

Preface: Genome revolution is shaping biomedical research and clinical medicine

The major landmark in modern genomic and biological research has been the first survey of the entire human genome. On June 2000 the staging of Bill Clinton along with Craig Venter and Francis Collins extolled how genome science would impact our lives by revolutionizing diagnosis, prevention and treatment for a vast number of human diseases (Collins 2010). Since that, we underwent a breathtaking progress in genome science with the unique conjunction of the development of new technologies such as Next Generation Sequencing (NGS) or genotyping arrays (Collins 2010; Hofker et al. 2014) and extensive data sharing initiatives catalysing new discoveries (Kaye et al. 2009; Collins 2010; Hood and Rowen 2013). To underscore the magnitude of this summit, the first sequence from the **Human Genome Project (HGP)** took 13 years and several collaborative efforts from a lace of international public research institutions entailing a 3 billion budget (U.S. Department of Energy & Human Genome Project program). Less than a decade later, NGS technologies have been implemented for clinical diagnosis, we entered in the $1,000 genome era, and the last Illumina sequencer, HiSeq X Ten is capable of producing up to 16 human genomes (1.8 terabases of data) in three days (Hayden 2014).

The success of NGS led to an astonishing rate of growth of sequence data (Koboldt et al. 2013), which is doubling every seven months (Stephens et al. 2015). A downstream consequence has been the rapid accumulation of the number of sequenced genomes of many vertebrates, invertebrates, fungi, plants and microorganisms enabling tackling evolution and genome function through the rationale of comparative genomics (Collins 2010). In addition, the build-up of sequence data of thousands of human subjects contributed to catalogue the genetic differences between individuals, or also called as **genetic variation** (Hofker et al. 2014). There are different types of genetic variation but the most abundant are Single Nucleotide Polymorphisms (SNPs) (Stranger et al. 2011), substitutions of single nucleotides. While the HGP reported around 1.4 M of SNPs (Lander et al. 2001) more than 84 M of SNPs have been described in the new phase 3 release of the 1000 Genomes Project (1000G-Phase3) (Sudmant et al. 2015; The 1000 Genomes Project Consortium et al. 2015). A final example to illustrate the large efforts invested in more accurate descriptions of genetic variation is the last work published from the Exome Aggregation Consortium (ExAC). This study involved the aggregation and analysis of exomic regions through sequencing data of 60,706 individuals (Lek et al. 2016). The disposal of this kind of data showed a widespread mutational recurrence in human genomes, it allowed detecting genes subjected to strong selection depending on the class of mutation and it is expected to facilitate the clinical interpretation of disease-causing variants (Lek et al. 2016). Thus, the accumulation of individual genetic data has empowered researchers to unravel those specific genetic variants associated with disease liability. We also moved from biologically guided candidate single gene-studies involving a few hundreds of individuals towards **hypothesis-free genome-wide analysis**, performing extensive and massive genomic interrogation of thousands of individuals (Relling and Evans 2015; Wang et al. 2015). Piecing these advances all together, we have expanded our understanding of disease pathophysiology. Therefore,

integrating the genetic understanding of the health-status alongside with clinical explorations constitutes the idea beneath personalized medicine. This genomic paradigm shift for clinical medicine provides a **new source of therapeutic breakthroughs and diagnosis** (Hood and Rowen 2013). As an example of this, targeted therapeutics have been resourceful for the treatment of lung cancer: sequence information revealed that tumours carrying specific mutations in the epidermal growth factor receptor (EGFR) were vulnerable to kinase inhibitors, resulting in higher response rates compared to traditional platinum-based chemotherapy (Levy et al. 2012; Swanton and Govindan 2016). Moreover, genetic tests are able to predict which breast cancer patients will benefit from chemotherapy (Innocenti et al. 2011; Gyorffy et al. 2015). Finally, notable successes have been achieved in pharmacogenomics, in which warfarin dose can be adjusted on the basis of genetic polymorphisms placed in *CYP2C8* and *VKORC1C* genes (Collins 2010; Hood and Rowen 2013; Relling and Evans 2015). In line with this, there are large efforts under way to prioritize targeted therapeutics and to optimize drug selection and dosing, such as the Genomics England 100,000 Genomes Project and the US National of Health (NIH) Pharmacogenomics Research Network (Relling and Evans 2015; Wilson and Nicholls 2015).

However, clear successes in clinical decision-making through genomic knowledge are anecdotal due to a **poor understanding of human genetic diseases** (Hofker et al. 2014; Relling and Evans 2015). For instance, **Genome Wide Association Studies (GWAS)** is undoubtedly one of the most important methodological advances emerging from the availability of complete human genome sequences and affordable DNA chips (Visscher et al. 2012; Hofker et al. 2014; Paul et al. 2014). GWAS have been extremely resourceful in identifying genetic variants associated with multiple diseases, but the translation of these results to clinics is sparse (Manolio et al. 2009; Collins 2010; Hofker et al. 2014). Some of the limitations lie on (1) the still small proportion of disease causing genetic factors identified for most complex diseases and (2) a lack of functional characterization and interpretation of disease associated variants, which hampers the identification of the underlying molecular mechanism (Manolio et al. 2009; Hofker et al. 2014).

The genomic revolution has brought new decisive players for the future trend in biomedical research and clinical genetics. The ‘genomical’ challenge is one of the most demanding **Big Data sciences** in all four big computer science domains (data acquisition, storage, distribution and computation). In order to meet this rapid progress of genomic research, the build-up of whole-genome sequences and the emergence of large population biobanks (Stephens et al. 2015) urges a parallel development of computational frameworks. Moreover, a real social concern about data privacy can discourage the participation in genetic studies, which requires a major discussion about the ethical consequences of the return of information to participants seeking for genetic diagnosis (Hood and Rowen 2013; Koboldt et al. 2013). From this brief overview, the agenda of human genomics has clearly many issues to address. In this thesis I translated some of them into the following general goal: setting a

cost-effective genetic research environment through the **implementation of novel analytical and computational methods** in order to **better understand the genetics of Type 2 Diabetes** (T2D). This work is a small glimpse of the frenzied activity in human genomics research and it aims to modestly contribute along with countless research efforts on this broad deployment of P4 medicine (*Predictive, Preventive, Personalized, Participatory)*. In the next sections of this dissertation, I want to spell out this primary focus by providing several concepts that I learned during these years, which prompted this research to successfully achieve the goals of this thesis.

Introduction

1 Disease burden shift: chronic diseases and Type 2 diabetes as the new threats

Life expectancy is continuously increasing as depicted on the latest estimations by the World Health Organization (WHO). Global average life expectancy hit 71.5 years in 2015 according to WHO, and is expected to reach 75 by 2045-2050 (European Environment Agency (EEA) 2015). The economic growth and improvement of social conditions that our society underwent have given access to basic health care and education but they have also favoured unhealthy lifestyles. The final outcome of these developments is an **epidemiological transition** in which non-infectious diseases (non-communicable diseases, NCD) out-weights the disease burden from infectious diseases (European Environment Agency (EEA) 2015). Later reports from WHO attributed 38 M of mortality to NCD (68% of worldwide mortality) in 2012. More than 40% of this mortality in 2012 corresponded to premature deaths under age 70 years, occurring mostly in low/middle income countries, but 28% of them also occur in high-income countries (WHO 2014). Of note, only four main NCDs (cardiovascular diseases, cancers, respiratory diseases and **diabetes mellitus**) are direct responsible for 82% of the whole NCD deaths.

Large inefficient treatment and prevention strategies are predominant for chronic diseases, such as diabetes mellitus (DM). Setting-up a healthcare infrastructure efficient enough to lower this financial burden must be the primary target of our efforts. Thereafter, human genomics has a key role in articulating personalized disease prevention strategies, in the development of new therapeutics and in the improvement of drug efficacy in patients (Collins 2010; Hofker et al. 2014).

2 Historical overview of genetics: where do we come from

Human genetic diseases can be distinguished according to different criteria. In order to explain disease burden, I made a distinction according the mode of transmission as communicable (infectious) or non-communicable. In this example, the criterion chosen is the occurrence of "infection", the action of a pathogenic microorganism for the disease transmission. An extension of the mode of disease transmission was the observation of the **inheritance** of some diseases and traits to the offspring. A key question in biology has been whether phenotypes or physiological traits can be transmitted across generations, and if the underlying causes are biological or environmental (Liu 2007). In this section I briefly traced the history of the genetic field, in which there was a parallel progress of population genetics articulated through several mathematical and statistical works, and molecular biology. Both developments have been critical to empower the study of the inheritance of disease traits.

2.1 From Hippocrates until the foundations of population genetics

Genetics is the science focused on the study of genes, genetic variations between individuals and inheritance (National Institutes of Health (US) 2007), concept that draws its ideas from the Ancient

Greece (Cobb 2006; Liu 2007). However, the history of classical genetics begins with **Gregor Johann Mendel** (1822-1884), who statistically studied inheritance for the first time. In 1866, this Austrian Augustinian monk published his study on pea plants in which he detailed how certain phenotypes or traits are transmitted to the offspring following certain rules (Cobb 2006), the "Mendelian laws of inheritance". Mendel outlined a mathematical framework explaining how a trait is passed from parents to their progeny through a **genotype** (the offspring receives a genetic unit from each parent), and how this genetic material is able to create new variations (Liu 2007). However, Mendel's work was brushed aside until the 20th century. Contemporarily, **Charles Darwin** (1809-1882) was unable to adequately incorporate inheritance in his theory of **evolution by natural selection** (Darwin 1859; Darwin 1868; Charlesworth and Charlesworth 2009). In order to justify heritable variability, which is indispensable for selection, he articulated his own theory of "pangenesis" (also proposed by Hippocrates): all kinds of variation occurring during lifetime are transmitted by means of *gemmules*. He suggested that all parts of the body throw off *gemmules* at different developmental stages and if any part underwent any kind of modification, it would be transmitted to the offspring (Darwin 1868). This hypothesis provides an explanation for the inheritance of acquired characters (Cobb 2006; Liu 2007; Charlesworth and Charlesworth 2009; Liu and Li 2012).

The birth of genetics is tied to the publication of the independent works on plant hybridization from Hugo de Vries (1848-1935), Carl Correns (1864-1933) and Erik Tschermak (1871-1962), that corroborated and rediscovered Mendel's work (Haynes 1998). De Vries asserted as Mendel that inheritance is driven by discrete particles and he also suggested that exact hereditary units named "pangenes" (or genes) were behind equivalent characteristics from similar species (De Vries 1889; Lenay 2000). De Vries also introduced the term "**mutation"** when suggesting how new species are the result of preexisting ones and the sudden appearance of inheritable variations, or mutations (De Vries 1901-1903; Lenay 2000). It was not until 1905 when Bateson (1861-1926), chief popularizer of Mendel's ideas, coined the word "genetics" in order to describe the study of heredity and the new phenomena of genetic variation (Haynes 1998). However, Mendel's theory was not easily embraced by the scientific community. The dominant view of inheritance was **"biometry"**, originated with Karl Pearson (1857-1936), based on the statistic analysis of continuously varying traits and gradual evolution from Darwinism (Rice 2014). On the other side, Mendelians such as Bateson argued that single strong mutations were beneath major adaptive changes, and thus, they were primarily interested in the inheritance of discrete traits and the identification of driver strong allelic effects (Stranger et al. 2011; Rice 2014). Thereafter, there was a large controversy between Mendelians and Biometricians that confronted Mendelian particulate inheritance in contrast to quantitative genetics used on continuously varying traits.

Population genetics was conceived as a need to reconcile Mendel with Darwin. This emerging field has been crucial to understand genetic variation within-species and gene mapping for human

diseases (Wakeley 2005; Crow 2010). Ronald Fisher (1890-1962) mathematically showed that continuous variation might arise from the combined action of multiple independent genes with small contributions, resulting in an approximate normal distribution for this given continuous trait (Fisher 1918; Fisher 1930). J.B.S. Haldane (1892-1964) articulated a mathematical framework for the origin and fall of genetic variation driving evolution (Haldane 1932). Finally, Sewall Wright (1889-1988) emphasized the role of genetic drift (i.e. random fluctuation of genetic diversity) in evolution as an evolutionary force altering the composition of genetic characters in a population through random sampling (Crow 2010). Fisher, Haldane and Wright provided the mathematical groundwork of the **"modern synthesis"** of Darwin's natural selection for evolution and Mendel's law of inheritance (Bowler 2003; Crow 2010; Charlesworth and Charlesworth 2016). This is the foundational moment of population genetics (Stranger et al. 2011), which is defined as the discipline focused on describing how evolutionary forces modify the genetic composition in a population (Charlesworth and Charlesworth 2016).

This theoretical core was extended with previous and further discoveries. For instance, in the second half of the 20th century, Motoo Kimura (1924-1994) along with James Crow (1916-2012) brought back the debate between Wright and Fisher about the role of genetic drift. The authors stated that genetic variability is mainly driven by neutral mutations and genetic drift (Kimura and Crow 1964). These successive works reduced the evolutionary process to manageable parameters such as mutation, drift, selection and recombination, which can be empirically estimated.

Another remarkable principle of population genetics was the **Hardy-Weinberg Equilibrium (HWE)**, independently conceived by G.H. Hardy (1877-1947) and W. Weinberg (1862-1937) in 1908 (Hardy 1908; Weinberg 1908). This principle answered one of the most challenging opponents of the evolution by natural selection proposed by Darwin, the "blending inheritance". According to this hypothesis, random mating would cancel out genetic variation, homogenizing trait variation, overriding natural selection as an evolutionary driving force. The Hardy-Weinberg principle shows how genetic variation is not lost in a population under Mendelian inheritance. The first take-home message was that frequencies of genetic variants are stable over time in the absence of evolutionary forces. Second, for each genetic unit, the distribution of genotypes in the next generation for diploid organisms can be predicted by a simple equation based on the frequencies of possible gametes in the population (Hardy 1908; Weinberg 1908; Wigginton et al. 2005).

Another constitutional principle that determined our ability to track the underlying causes of inherited diseases was **genetic linkage**. Genetic linkage is the **physical association of inherited genetic units** (Stranger et al. 2011), which contradicted Mendel's law of independent segregation of different trait characteristics (Lobo and Shaw 2008). By studying inheritance of two traits (colour and shape) in sweet peas plants, Bateson and Reginald Punnett (1875-1967) realized that the ratios from the phenotypic combinations of the crossings deviated from Mendel's law (increased occurrence of

purple-long and red-round peas). They deduced that a physical coupling mechanism connected the genetic characters from these two traits (Bateson et al. 1905). Later, using fruit fly as an organism model, T. H. Morgan (1866-1945) discovered that a white-eyed mutant phenotype was tied to males (Morgan 1910). He reasoned that this alteration should be placed in the sex chromosome and he argued that if a trait is physically coupled to a specific chromosome, this should be true for others. In 1911, Morgan suggested for **traits segregating together** that their respective genetic characters should reside **close in the same pair of homologous chromosomes** (Morgan 1911). Morgan also suggested that during meiosis, homologous chromosomes exchanges some parts, what we know as **genetic recombination** (Griffiths et al. 2000a; Lobo and Shaw 2008). This knowledge was crucial for presenting genes as physical entities that underwent genetic recombination and can be specifically placed on chromosomes. This work and other studies of linkage opened the way to gene mapping, which allowed unravelling the basis of inherited diseases (Lobo and Shaw 2008).

2.2 The DNA era in molecular biology

To summarize, there was a vivid progress on conceptualizing inheritance and evolution but it still remained obscure how these genetic characters were molecularly transmitted. Morphological structures or "chromosomes" were identified by observing cell division (Flemming 1965), which served Morgan to show how specific genes are physically attached. However, how this information was organized and of what actually consists, was a mystery. Thus, the parallel deployment of molecular biology to this theoretical progress was indispensable. Actually, the second half of the 20th century is known as the DNA era. Friedrich Miescher (1844-1895) identified the "nuclein" substance from white-cell nucleus in 1869, now known as DNA (DeoxyriboNucleic Acid) (Dahm 2010). Shortly afterwards, DNA material was identified as the molecule behind the inheritance (Hershey and Chase 1952; Griffiths et al. 2000b). With the progress of X-ray crystallography, a great focus of study was placed on unveiling the tridimensional structure of complex biological molecules. Maurice Wilkins (1916-2004) and Rosalind Franklin (1920-1958) contributed with X-ray studies to the research of DNA molecules, and the latter one produced the first picture of DNA fibres (Griffiths et al. 2000c). In 1950, Erwin Chargaff (1905-2002) reported equal base ratios in any DNA sample which suggested to Watson that bases on each DNA strand were paired (Chargaff et al. 1950; Griffiths et al. 2000c). In conjunction with the unauthorized glimpses of Franklin's images showing the double-stranded helical structure, James Watson and Francis Crick were on the verge of reporting the model of DNA: the double helix. The double helix model was published in 1953, and it has been underscored as one of the most significant discoveries of the 20th century (Watson and Crick 1953; Griffiths et al. 2000c). Soon afterwards, Francis Crick conceived the central dogma of molecular biology, the principle to understand the relationship between DNA and proteins (Crick 1970).

Once the structure and function of DNA were discovered, research was redirected to decipher the DNA sequence. Fred Sanger (1918-2013), who had technically enabled reading the sequence of

protein chains in 1950, succeeded in discovering the DNA sequence of a bacteriophage ϕX174 in 1977 (Sanger et al. 1977; Hutchison 2007). However, scaling DNA sequencing from the 5,386 bases of the ϕX174 phage to the ~3 billion base pairs (bp) of the human genome was a huge technical challenge, which required assembling a large-scale international project, the HGP. The culmination of this project is a milestone in molecular biology and medicine: the HGP accelerated the development of high-throughput sequencing technologies, building-up an astonishing number of sequenced genomes, and catalysing the study of genetic variation and the inheritance of diseases and traits.

3 Genetic variants and their contribution to human genetic diseases

In this section I want to describe the different types of genetic variation and the structural properties, such as linkage disequilibrium, that have been crucial to track inheritance of traits and diseases.

3.1 Genetic variation: remnants of our history

Genetic information is stored in each of our cells as molecules of DNA, a 3 billion-long sequence of nucleotides (A, T, C and G, that stands for Adenine, Thymine, Cytosine and Guanine). As diploid organisms, we have two copies of this molecule that we inherited from our respective parents, which differ between them. Each specific physical position in the genome is called a *locus* (pl. *loci*), which can encompass a large region such as a gene or be narrowed to a particular base pair position. Alternate forms of each *locu*s are referred as an *allele*. However, the term is loosely used to name the alternate forms of a specific base pair position. For instance, for a *locus*$_J$, the “G” is the most fixed DNA base in the population, but some individuals have the alternative “T” DNA base. Each “G” and “T” forms are *alleles*. For a given position, the two alleles inherited from each parent are called a *genotype*. Genotypes can be homozygous when an individual inherits identical alleles from each parent or heterozygous, when each parent transmitted different *alleles* for a given *locus.*

Genetic variation corresponds to the naturally occurring **differences among individuals** in a population, which are gathered in our genotypes. Of note, the latest estimations from the 1000G-Phase3 release highlighted that a typical genome differs from the human reference sequence in 4.1 M to 5 M sites (The 1000 Genomes Project Consortium et al. 2015).

By estimating the proportion of variant sites at a population level we can track remnants of human evolution. The average proportion of variant sites is not homogenous across populations: populations with African ancestry retained the highest number of variant sites compared to other populations, which is in concordance with the out-of-Africa human origin model (Stranger et al. 2011; The 1000 Genomes Project Consortium et al. 2015). This hypothesis suggests that modern humans that were originated in Africa replaced non-African populations and it has been widely accepted since it was proposed in the late 1980s (Cann et al. 1987; Stringer and Andrews 1988; Wilson and Cann 1992). Additionally, the lower proportion of genetic variation in humans compared to other apes has been

shaped by the out-of-Africa migration of our ancestors. During their expansion over the globe, our ancestors encountered new environments leading to novel adaptions or founder effects that were followed by population bottlenecks (Jorde and Wooding 2004; Lachance and Tishkoff 2013). In these last 10,000 years, our environments and quality of life underwent drastic changes that have been translated into an asymmetry between our genomes and our current environments. An example of that is the "thrifty genotype" hypothesis, which was proposed to explain, for example, the high prevalence of T2D: alleles associated with fat deposits and the increase of the risk for T2D, were advantageous for early hunter-gatherers. However, in our social framework these genetic variants favoured what has become a threat to the subsistence of our health infrastructures (Lachance and Tishkoff 2013; Segurel et al. 2013). This hypothesis is not free of controversy, and last studies pointed to opposite scenarios. However, our incomplete understanding of the whole set of genetic factors modifying T2D susceptibility, as it occurs for the majority of common diseases, makes denser studies of the evolution of T2D susceptibility still a challenge (Segurel et al. 2013).

3.2 Types of genetic variation

Genetic variation takes many forms ranging from the narrowest to the largest scale in: (a) Single Nucleotide Variants (SNVs), (b) Insertions and deletions (INDELs), (d) Tandem Repeats, (e) variable number of copies of a segment of DNA sequence (Copy Number Variants, CNVs), (f) inversions and translocations (Copy Neutral Variants) of these segments and other large structural events that can even lead to chromosomal aneuploidies (Ku et al. 2010; Baker 2012; Zhao et al. 2013) (Figure 1).

However, there is still a lack of consensus in the classification of genetic variations and the criteria available are neutral, without referring to the association with a phenotype or a disease. One useful criterion is the **size of DNA sequence** that these variants encompass. We coin the term "structural variation" (SV) for genetic variants involving segments covering more than 100 bp (the number is arbitrary, earlier definitions used a 1 kilobase pair (kb) cut-off because of the ability to detect of smaller variants) (Baker 2012). Within the "structural variant" category, we have alterations that are quantitative such as copy number variants while copy neutral variants are positional (translocations) and orientational (inversions). CNVs have been limited to segments of DNA ranging smaller than 5 Mb (megabase pair) whereas large structural variants responds to alterations involving more that 5 Mb of DNA sequence (Ku et al. 2010; Zhao et al. 2013) as is represented in Figure 1.

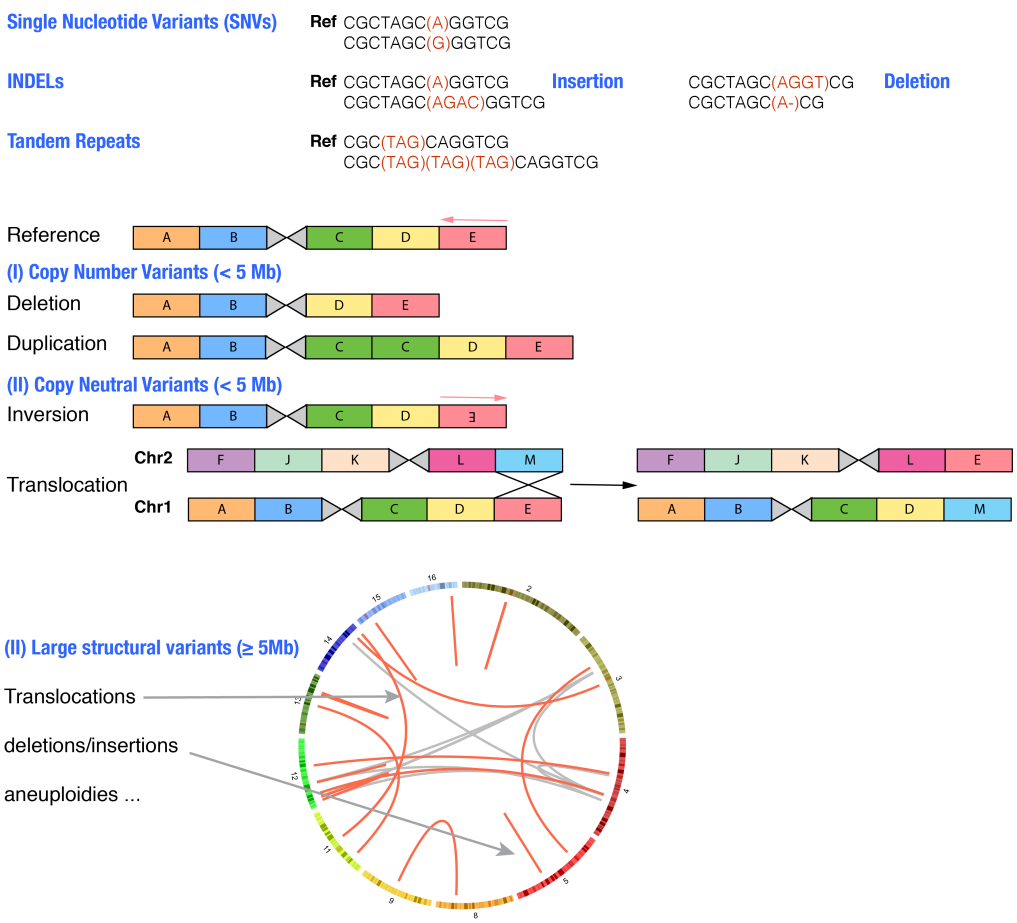


Figure 1. Diversity of genetic variation. Based on the size of DNA sequence of the variant, we can distinguish between Single Nucleotide Variants, short insertions and deletions (INDELs) and tandem repeats, which entail segments shorter than 1kb. On the other hand, we have structural variants, categorized as Copy Number/Neutral variants sizing less than 5Mb and large chromosomal rearrangements displayed on a "circos" representation of structural variants (where the circle corresponds to chromosomes and the inner lines, to structural variants).

The frequency of these genetic variants within the population is an alternative way to categorize genetic variation. "Polymorphism" is an umbrella term for all kind of genetic variation accounting for population frequencies above 1%, which now also includes Copy Number Polymorphisms (CNP) a part from Single Nucleotide Polymorphisms or SNPs.

For each polymorphism, we can distinguish between the "major" (highest frequency in the studied population) and the "minor" allele (variant form in the population) on the basis of the frequency in general populations. Therefore, polymorphisms are traditionally classified as common when the frequency of the **minor** or rare allele (Minor Allele Frequency, MAF) remains above 5% (Ku et al. 2010). Huge efforts for cataloguing genetic variation through sequencing studies such as the 1000G Project (The 1000 Genomes Project Consortium et al. 2010; The 1000 Genomes Project Consortium et al. 2012; The 1000 Genomes Project Consortium et al. 2015) or the UK10K Consortium (UK10K Consortium et al. 2015) have enlarged the landscape of genetic polymorphisms. There is plenty of attention whether low-frequency (1% ≤ MAF < 5%) and rare variants (0.1% ≤ MAF < 1%) contribute

to disease susceptibility and phenotype variation (Huang et al. 2015; The 1000 Genomes Project Consortium et al. 2015). Nevertheless, these allele frequency boundaries are standards to facilitate the interpretation and the study of genetic variation and disease and trait outcomes. The accumulation of genetic individual data like in the last call set from the whole exome-sequence data of 60,706 individuals of the ExAC consortium (Lek et al. 2016) has enlarged the landscape of genetic variation towards more rare genetic variants. The availability of personal genomes would ultimately convey to the characterization of more private and individual genetic variants (Lupski et al. 2011). Therefore, with the advent of personal genomes these artificial boundaries can fall, which will favour a continuum conceptualization of genetic variation in terms of allele frequency.

In summary, we can conceive genetic variation according to different criteria but in any case, these boundaries are agnostic of the molecular mechanisms that the genetic variants might mediate. These classifications are only based on patterns of DNA sequence changes (Ku et al. 2010).

3.3 Linkage disequilibrium: breaking down the correlation patterns of human genetic variation

As mentioned above, the most abundant type of the 0.1% of sites that are variable in a typical human genome is by far composed by SNPs, which have become the suitable markers to explore the relationship between our genotype and inherited diseases. However, this large collection of millions of SNPs is not providing unique and independent information. In fact, population genetic forces have brought structure to our genomes, which is reflected in the occurrence of **linkage disequilibrium (LD)** a non-random association of alleles from different *loci* (Slatkin 2008). This **correlation structure** varies across the genome and populations (Frazer et al. 2009), and also depends on the physical exchange of DNA during meiosis, also called as **recombination.** Recombination events in the genome are confined in hotspots, which determine the boundaries between blocks of linked alleles from different *loci* (Daly et al. 2001; Wall and Pritchard 2003). Closer markers are less likely to suffer from a recombination event, thus alleles at different *loci* but spatially close will be transmitted together from parents to offspring (Crawford and Nickerson 2005; Frazer et al. 2009; Ku et al. 2010).

To exemplify this correlation, two SNPs are in LD if by observing a specific allele A for the first SNP, there is more chances to observe a specific allele B for the second SNP. Thus, these two alleles are entangled by the LD correlation. This correlation can be mathematically estimated or quantified by *D*, the coefficient of linkage disequilibrium, described by the frequency of gametes carrying simultaneously the pair of alleles A and B at two *loci* (p_{AB}) and the frequencies of these alleles (p_A and p_B) (Slatkin 2008).

$$D_{AB} = p_{AB} - p_A * p_B$$

However, this descriptive statistic was inconvenient to compare LD across different pairs of alleles because the possible values of *D* strictly depend on the allele frequencies. The normalisation of *D*

was *D',* a ratio based on the maximum possible absolute *D* value (D_{max}) according to the observed allele frequencies (Lewontin 1964).

$$D' = \frac{D}{D_{max}}$$

Another useful measure of LD is the r^2, which is similar to *D',* a correlation coefficient ranging from 0 to 1 expressed as (Slatkin 2008):

$$r^2 = \frac{D^2}{p_A * (1 - p_A) * p_B * (1 - p_B)}$$

The study of LD extended the definition of "haplotype" to refer to the combination of correlated alleles from different markers at the same chromosome, which are inherited together. Therefore, this knowledge allowed determining regions with almost no evidence of recombination, accounting for a set of markers in high LD, that were called "haplotype blocks". In addition, it was noted that these blocks were separated by hotspots of recombination (Crawford and Nickerson 2005; Hofker et al. 2014). The discovery of haplotype blocks posed the following hypothesis: in order to assess genome-wide which genetic variants are associated with a certain disease, testing a single variant per block was informative enough (Slatkin 2008) as illustrated in Figure 2.

The **International HapMap Project (International HapMap 2003),** following the HGP, was a pioneer huge collective effort fuelled by the opportunity to describe the human genome in terms of haplotype blocks, focused on describing common genetic variation and informing about which SNPs remain linked during chromosomal recombination and inherited together across all the genome. The Phase I an II of the project catalogued ~3 M markers in 269 individuals from four populations (Yoruba, Japanese, Han Chinese and Utah residents with European ancestry). In the Phase III, 1.5 M genetic variants were genotyped in a larger set of samples including seven additional populations (Slatkin 2008). The HapMap project provided an extremely useful report that guided the genetic studies of inherited diseases: the majority of variants within the HapMap project (MAF ≥ 5%) were adequately captured by half a million of SNPs (Frazer et al. 2009; Hofker et al. 2014). Therefore, studying the genetics of T2D susceptibility was not tied to genotyping millions of variants. In fact, researchers were able to capture those signals correlated with a disease phenotype just handling half a million of proxy SNPs or tagSNPs, which economically enabled the GWAS approach.

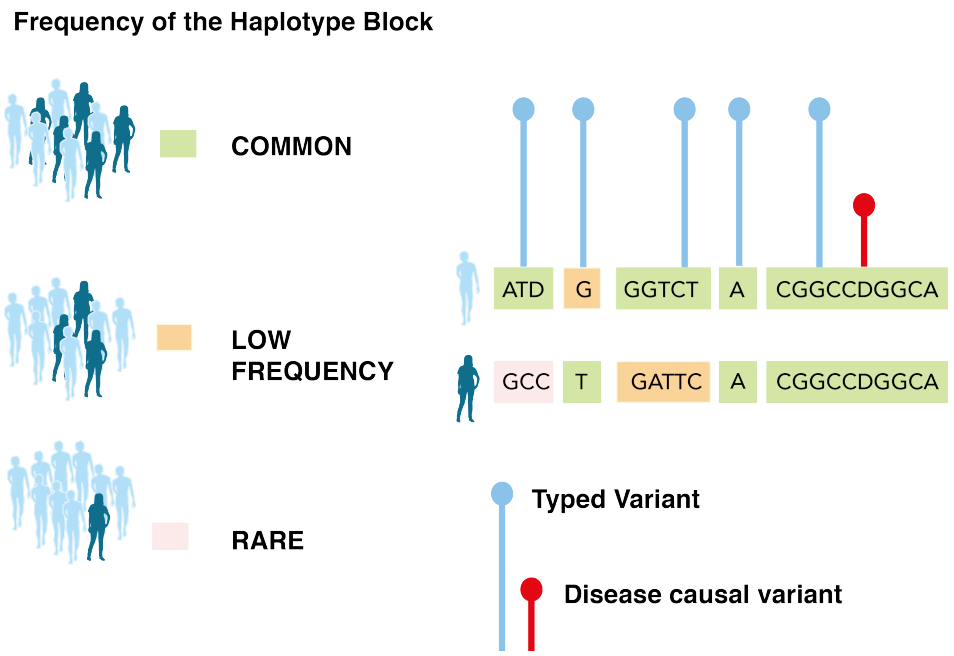


Figure 2. The structure of the genome in haplotype blocks. At the right, a single genomic region for two individuals has been illustrated as haplotype blocks (boxes), which are determined by measures of LD. Each letter corresponds to an allele from an SNP (A, G, T or C) or from an insertion/deletion (I/D). Blue lines highlighted those tagSNPs required to be genotyped and the red line, underscores a disease causal variant within a populated haplotype block. The left part of the figure shows the frequency of the haplotypes depending on the colour.

4 Characterization of human inherited diseases: heritability and genetic architecture

Human genetic diseases are highly heterogeneous but some historical classifiers were established to provide a theoretical basis to study them. This next section was written with the aim to comprehensibly describe how we have characterized the genetic basis of human inherited diseases, which has determined our methodologies to understand them

4.1 Heritability: quantifying the genetic contribution to trait transmission

A crucial observation that synthesizes the concept of inheritance was that for most human traits, relatives tend to be more alike compared to random individuals from the population. This resemblance among relatives fostered the study of inheritance of traits and diseases, which allowed us to assemble a new concept named as “heritability” (Visscher et al. 2008). Resemblance can arise from common environmental and inherited factors, and heritability addresses the partitioning of this resemblance into nature and nurture. **Heritability** allows us to compare the importance of genetics against environment in explaining the trait variation enclosed in a population. Heritability measures how much variability of a specific trait is controlled by genetic differences (Visscher et al. 2008; Tenesa and Haley 2013). Technically, heritability is formulated as a ratio of variances: the proportion

of the total variance of a particular measurement (a phenotype) in a population, which is attributable to genetic variation (Visscher et al. 2008; Wray and Visscher 2008).

Traditionally, heritability was estimated by means of a **(i) within-family design**, looking at the correlation of full or half siblings, by a **(ii) pedigree design** such as the regression of offspring on parental phenotypes or the observed differences in the correlation of monozygotic (MZ) and dizygotic (DZ) twin pairs, and by a **(iii) population design** based on the genetic similarity between distant relatives (Vinkhuyzen et al. 2013). Moreover, heritability may also be estimated by means of adoption studies, including MZ twins reared apart and non-biological relatives reared together (Sham and Cherny 2011; Tenesa and Haley 2013). Of note, the emergence of genome-wide SNP data has been crucial to overcome the confounding of genes and environment that led to biased estimations of heritability. Using a population of unrelated people, for which only the proportion of genetic variance explained by SNPs is captured, may avoid inflated estimations of heritability due to environmental factors shared between related individuals, among other factors (Vinkhuyzen et al. 2013; Zaitlen et al. 2014).

Heritability has been extremely crucial to provide meaningful and appropriate comparisons of traits and it is an informative indicator of the efficiency of gene-mapping or the prediction of genetic risk in human disease studies (Visscher et al. 2008). The highest the heritability, the easier it should be to identify genetic risk factors for the disease.

4.2 Genetic architecture of human diseases: disentangling genotype-phenotype relationships

The study of human genetics has unlocked the underlying genetic basis of a vast number of Human Genetic diseases and its activity has been extremely intensified during this last decade. All the research efforts devoted to link genetic variants and genes with specific disease phenotypes benefited from the creation of public databases such as The Online Mendelian Inheritance in Man (OMIM) (Amberger et al. 2009; Amberger et al. 2015). OMIM accounts for 23,603 entries (accessed July 20th, 2016) and the NHGRI-EBI Catalog of published genome-wide association studies reported 23,058 unique SNP-trait associations across 2,502 studies (accessed July 20th, 2016) (Welter et al. 2014).

The relationship between genotype and phenotype has always been hard to decipher, but some basic models that helped to conceive the genetic structure beneath human genetic diseases are represented in Figure 3. These simple schemes summarize part of what we designate as **"genetic architecture**", that depends on the number, frequencies and effect sizes of disease causal variants (Flannick et al. 2016). Shifts in the parameters beneath the genetic architecture led to the traditional classification of human genetic diseases: monogenic and complex (polygenic) diseases.

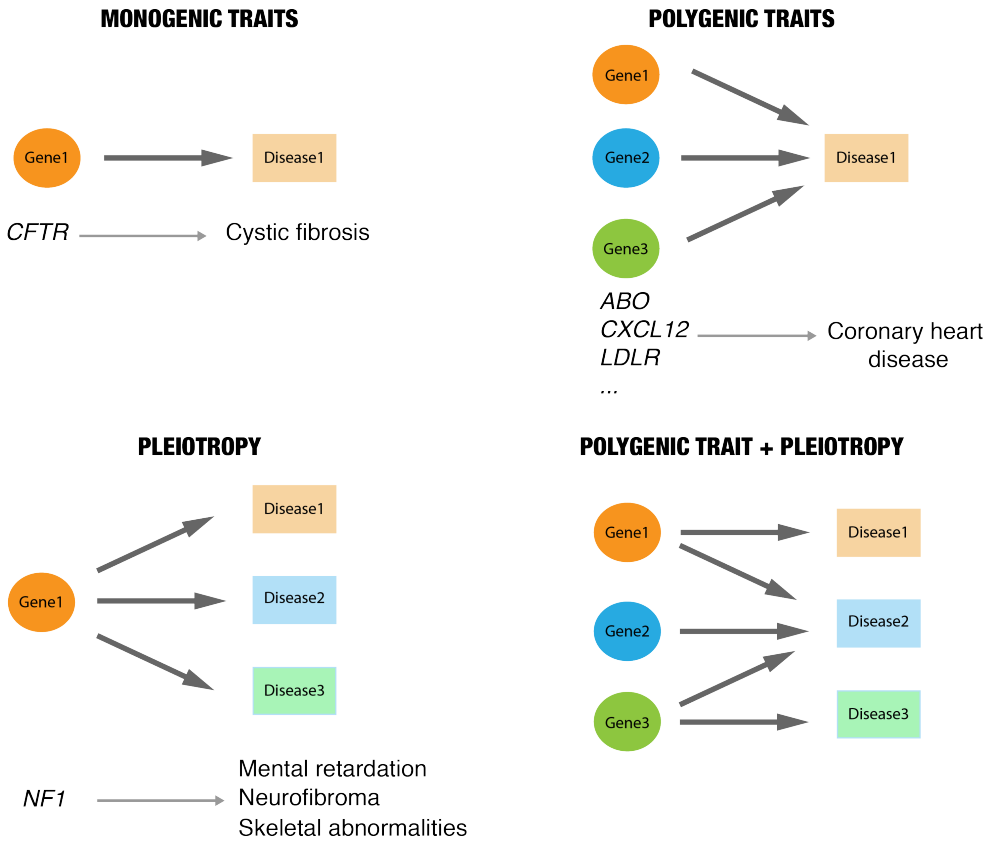


Figure 3. Genetic models behind human genetic diseases. From top to bottom, left to right, traits and diseases can be shaped as (a) Monogenic traits caused by single-gene defects, (b) Polygenic traits requiring the involvement of multiple gene defects, (c) Pleiotropy in which a single gene or genetic variant yield to different phenotypic manifestations and (c) a model based on polygenic and pleiotropic effects.

4.2.1 Mendelian o monogenic diseases

Most of our knowledge about human genetic diseases answered the genetics behind monogenic diseases such as Huntington disease (Gusella et al. 1983; MacDonald et al. 1993) or Cystic Fibrosis (Riordan et al. 1989a; Rommens et al. 1989; Janssens and van Duijn 2008). These disorders are characterized by rare mutations interfering with the specific **function of a single gene** (Janssens and van Duijn 2008). They are also called “Mendelian” diseases because they segregate according to Mendelian laws following several models of inheritance: autosomal dominant, recessive, co-dominant, sex-linked (Health 2010). Thus, one deleterious variant or defect is sufficient to cause pathogenic phenotypes, observed by a remarkably higher risk of disease in carriers of these mutations compared to non-carriers (Janssens and van Duijn 2008). The genetic architecture of monogenic diseases has been conceived to involve genetic variants with huge effect sizes and a negligible intervention of the environment, although variable penetrance can occur. Moreover, highly

penetrant[1] variants tend to lower allele frequencies, and thus, monogenic disorders are attributed to rare individual mutations or for instance, mutations enclosed in a familiar lineage (see Figure 4) (Cooper et al. 2013; Flannick et al. 2016).

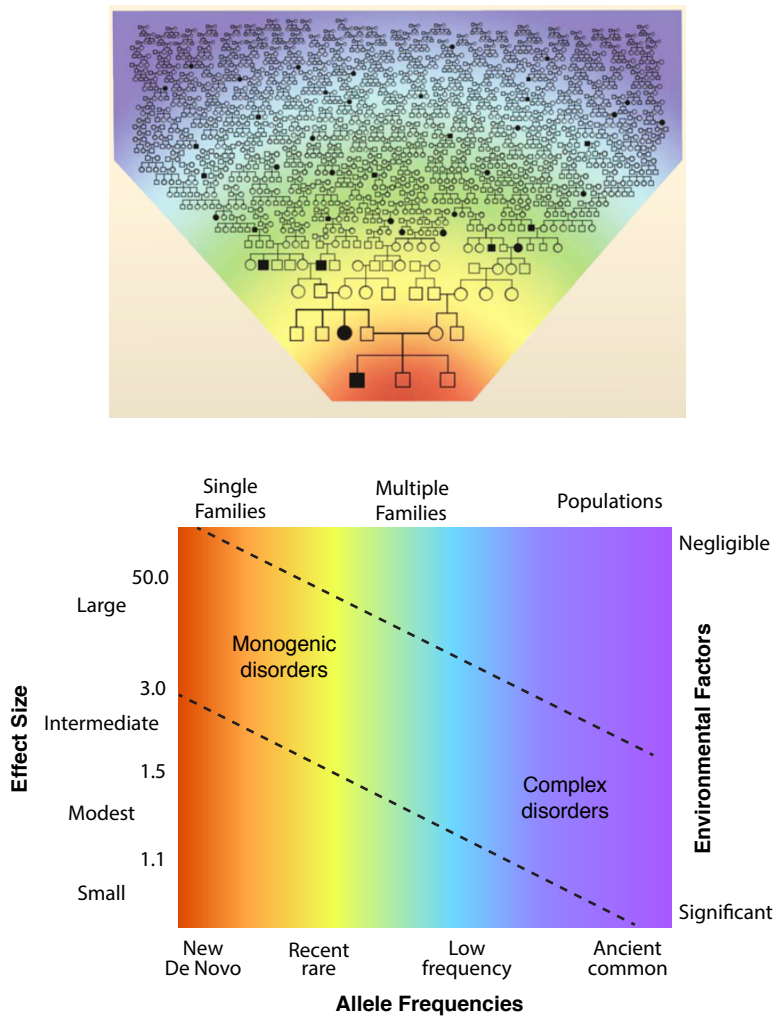


Figure 4. Genetic architecture of human genetic diseases. The number, the allele frequency and the effect size of the disease-causal variants determine the genetic architecture. At the top, a coloured gradient representing a large population ancestry until a single familiar structure is shown (obtained from Lupski, J.R. et al., 2011). At the bottom, using four axes (effect size, allele frequency, involvement of environmental factors and occurrence in single families or populations) complex and monogenic diseases have been placed in a continuum phenotype in a landscape delimited by historical criteria for classifying human genetic diseases.

The study of single-gene disorders had an unquestioned value in the enrichment of our understanding of gene function, regulation mechanisms, human phenotypes and body physiology and vice versa. These advances allowed the discovery of novel therapeutic and diagnostic strategies and improving the patient care for rare and complex diseases commonly occurring in the population (common form of a disease) (Chong et al. 2015; Flannick et al. 2016). For instance, mutations in

[1] Penetrance corresponds to the percentage of individuals carrying a particular mutation or genotype that also develops a certain disorder or exhibits a certain phenotype.

genes altering renal salt reabsorption underlay monogenic disorders such as Bartter and Gitelman's syndromes (increased blood pressure) or the Liddle syndrome (decreased blood pressure). The understanding of Bartter's syndrome, a life threatening disorder characterized by hypotension, for which mutations in the *KCNJ1* gene have been attributed, allowed proposing ROMK, the product of *KCNJ1,* as a novel target for hypertension and heart failure (Garcia and Kaczorowski 2014).

Traditionally, linkage mapping combined with Sanger sequencing has been applied to identify the genetic defects beneath a large fraction of monogenic disorders (discussed afterwards) (Chong et al. 2015; Flannick et al. 2016). NGS techniques allowed turning a blind eye to prior biology and accelerating the identification of genes underlying monogenic phenotypes. Thus, the pace of disease gene discovery has increased from ~166 to 236 between the time periods of 2005-2009 and 2010-2014 (Chong et al. 2015) that resulted in 2,937 genes reported for 4,163 monogenic phenotypes. However, for ~50% of all known monogenic disorders, the underlying causal genes are still unknown. In connection with disease burden, aggregating clinically identified monogenic phenotypes and all congenital anomalies sums up to 8 M births worldwide that present a "serious" genetic condition, which represents $5 M healthcare expenditure per child during their lifetime in United States (Chong et al. 2015).

4.2.2 Complex diseases

Complex polygenic diseases result from an intricate **interplay** between **genetic, environmental and lifestyle factors** and represent a vast public health impact (Buchanan et al. 2006). In contrast with monogenic diseases, there is no action of a distinctive Mendelian inheritance pattern. However, complex diseases have a tendency to cluster in families, which is in agreement with their significant multifactorial genetic component that has been arduous to identify until the last decade (Manolio et al. 2009; Flannick et al. 2016). As shown in Figure 3, we need to move away from single gene-disorder associations, a kind of 'genetic deterministic' rationale and we need to rely on a new concept, 'genetic predisposition': a single variant showing a modest or weak effect size is not able by itself to directly cause a complex disease phenotype (Buchanan et al. 2006).

The genetic architecture of complex diseases is shaped by the action of numerous low penetrant variants from multiple *loci*, which synergistically modulate disease susceptibility, in conjunction with an environmental component (Manolio et al. 2009). During this last decade, the most accepted model for the genetics of complex diseases has been steeped in the common disease-common variant **(CDCV) hypothesis** built on the basis of the infinitesimal rationale proposed by Fisher: multiple genetic variants commonly occurring in the population (MAF ≥ 1-5%) have individual modest effects on disease susceptibility. However, in a cooperative manner, they grant substantial risk of manifesting the complex disease phenotype (Reich and Lander 2001; Manolio et al. 2009; Lowe and Reddy 2015) (see Figure 4).

Nevertheless, the small fraction of estimated heritability, ranging from 5 to 20%, explained for the majority of common diseases (Agarwala et al. 2013) brought researchers to reconsider this model. Alternative hypotheses, such as the infinitesimal model, stressed the involvement of a larger number of common variations contributing with **small increments to the risk of disease** (Manolio et al. 2009; Gibson 2011). Some authors also supported that **rare variants** showing large effect-sizes (Pritchard 2001) are a key source to increase the fraction of explained heritability, specially considering the underestimated abundance of rare and private variants arising from the analysis of whole-genome sequences (Lupski et al. 2011). In line with that, novel trends have conceived complex diseases as a collection of multiple, even hundreds of rare monogenic sub-phenotypes driven by rare variants. In this context, GWAS results would only reflect **synthetic associations** in which rare variants happen to segregate with common GWAS signals (Dickson et al. 2010). Moreover, the term **"clan genomics"** has been used to argue that common diseases can be driven by a unique combination of rare alleles of recent origin clustered in a family lineage (Lupski et al. 2011). Furthermore, **epistatic** gene-gene and **gene-environment** interactions were also suggested to explain a fraction of the aetiology of complex diseases (Schork 1997; Manolio et al. 2009; Gibson 2011).

Additionally, in order to get deeper insights into disease aetiology, reducing phenotypic heterogeneity is fundamental such as in approaches based on using **intermediate phenotypes**, quantitative measures of a disease characteristic (Buchanan et al. 2006; Wang et al. 2012a). Additionally, for apparent monogenic diseases such as sickle cell anaemia, the heterogeneity of mutations and the action of genetic variation in unlinked modifier genes, are able to modify disease penetrance, which explains the clinical heterogeneity of some monogenic disorders (Cooper et al. 2013). These examples are pushing us to redefine the boundaries between monogenic and complex diseases. Applying techniques for the study of single-gene disorders may facilitate the study of genes behind intermediate phenotypes related with complex traits. But also, the methodology behind the study of the genetic architecture of complex diseases can help us to comprehend the multigenic nature from monogenic disorders (Cooper et al. 2013; Tallapragada et al. 2015; Flannick et al. 2016).

To put this story together, the growing accumulation of sequencing data has enlarged the spectrum of genetic variation, and thus, some authors have suggested that each individual genome should be conceived as a unique spectrum of mutational burden. Within this continuum, each individual pathogenic ecology would encompass inherited and *de novo* variants: inherited common variants segregating in the population, inherited rare variants of recent origin in a familiar lineage, combinations of novel emerging rare variants from each parent and *de novo* mutations. This novel trend suggests that historical categories of human diseases can be placed along a single disease continuum. Understanding health status as a continuum breaks down all the practical boundaries of human diseases established due to an incomplete understanding of the mutational load. Therefore, each traditional category only reflects a different phenotypic manifestation arising from the whole

individual genetic burden in conjunction with environmental influences (Lupski et al. 2011). This new debate arising from an incomplete understanding of the role of the different ranges of the whole spectrum of genetic variants (from rare, to low-frequency until common variants and from weak to highly-penetrant effect sizes) showed that we are still lagging behind the profound comprehension of the genetic architecture of complex diseases.

5 Evolution and perspectives of genomic approaches for studying the genetics underlying complex diseases

The emergence of high-throughput sequencing platforms has technically and economically enabled the massive detection of genetic variation, spurring the genetic research of the basis underlying common complex diseases. In this section, I want to briefly provide an historic perspective of the evolution of genetic mapping to facilitate fully comprehension of the success and limitations of large-scale genetic studies. After that, I will proceed to detail the current state art of large-scale genetic analysis for complex diseases. Finally, I will argue the several challenges that are hampering our understanding of the genetics underlying complex diseases and the ultimate translation on clinical-decision making.

5.1 Genetic mapping before the completion of the Human Genome Project (HGP)

The HGP constituted a major landmark that revolutionized biomedical and genetic research. But, what was known before that? Geneticists realized that some traits are inherited according to mathematical Mendel's ratios as consequence of single gene defects, but the vast majority of trait variation resulted from the interplay between several genes and non-genetic factors (Altshuler et al. 2008; Stranger et al. 2011). Afterwards, the discovery of **genetic linkage** (Bateson et al. 1905; Morgan 1910; Morgan 1911) fostered the origin of *genetic mapping*: observing how DNA variation segregates with trait variation without relying on any prior biological guidance enables localizing which genes underlie certain phenotypes (Altshuler et al. 2008; Stranger et al. 2011). The first reports of genetic mapping were linkage analysis formulated by Sturtevant for fruit flies in 1913: crossing parents varying at a Mendelian trait enabled the identification of genetic markers that were segregating with that trait. Later in the 1970s, the emergence of DNA methodologies such as cloning or Sanger sequencing enabled with genetic linkage maps (positional cloning) zooming in the specific causing genes for Mendelian or Monogenic traits (Altshuler et al. 2008). However, until the end of the 20th century, genome-wide linkage analysis in humans had technical impairments such as small family sizes, the impossibility of intervention in parent' crosses and the limited number of genetic markers to trace across individuals (Altshuler et al. 2008). DNA polymorphisms in the form of Restriction fragment length polymorphisms (RFLPs) were described by Jeffreys in 1979 in beta-globin gene cluster (Jeffreys 1979) and he revealed that they commonly occur in the genome. Botstein and colleagues realized in 1980s that this kind of genetic inter-individual variation, was a potential source of marker *loci* (Botstein et al. 1980)*;* they outlined the seed of human genetic linkage

maps which are the basis for mapping inherited diseases (Botstein et al. 1980; Hofker et al. 2014). These findings were crucial for the identification of the *HTT* gene for Huntington's disease in 1983 (Gusella et al. 1983) or the *CFTR* gene in cystic fibrosis (Kerem et al. 1989; Riordan et al. 1989b; Rommens et al. 1989). Therefore, a large fraction of Mendelian diseases were mapped by linkage analysis followed by positional cloning at the mid-1990s (Hofker et al. 2014). Geneticists were tempted to recycle all the lessons from Mendelian disease genes to study complex diseases. However, hatching links between genotype and complex phenotypes was slow and arduous because these disorders do not obey Mendelian inheritance patterns and account for substantial environmental contributions. Linkage analysis or candidate-gene association studies brought sparse successful results (Manolio et al. 2009). **Linkage analysis** takes advantage of those shared DNA segments inherited from common ancestors in order to identify genetic variants strongly segregating with a phenotype and thus, largely contributing to the molecular pathology (Billings and Florez 2010; Torres et al. 2013). This approach is convenient for Mendelian diseases: (1) disease-causing variants are rare and the disease allele segregates within the same chromosomal region within each family and (2) Mendelian diseases are defined by highly penetrant variants which results in co-segregation with disease status (Hirschhorn and Daly 2005; Billings and Florez 2010). Genome-wide linkage analysis succeeded in mapping genes underlying rare monogenic forms for DM such as the maturity-onset diabetes of the young (MODY) (Fajans et al. 2001; Vaxillaire and Froguel 2006) or for the identification of the Major Histocompatibility Complex (*MHC)* locus for type 1 diabetes (T1D) (Castano and Eisenbarth 1990; Tienari et al. 1992). However, in accordance with the CDCV hypothesis, complex diseases are based on allelic variants characterized by a high frequency in the population (MAF ≥ 5% of the population) (Manolio et al. 2009) and susceptibility to disease spread across a large number of genetic variants (Manolio et al. 2009; Billings and Florez 2010). One of the few successes achieved through linkage analysis in complex diseases is the identification of evidences of linkage from the *TCF7L2 locus* for T2D (Grant et al. 2006). Later, this evidence was validated through association studies (Groves et al. 2006). **Genetic association methods** were an alternative approach to disentangle the genetics beneath complex diseases. These first association studies were only able to interrogate specific candidate genes, and therefore, this approach was guided and limited by prior biological knowledge. The main criticisms to this approach were the limited knowledge to provide plausible functional genes and variants to test or the dubious novelty of the results generated. Moreover, a main downside was the lack of replicability across association studies resulting in false positives because of population structure, heterogeneity of the phenotype or low prior odds of association. Additionally, there is also a publication bias towards positive results in which some authors overlooked failed replications that might inflated the estimates of replicability (Tabor et al. 2002; Marigorta and Navarro 2013). Nonetheless, this approach was able to identify *PPARG* and *KCNJ11* as novel candidate genes for T2D, which harbour missense variants associated with T2D (Altshuler et al. 2000; Gloyn et al. 2003).

In summary, before the HGP, disease causing genes and mutations strongly segregating across families were identified for a vast number of monogenic diseases. For complex diseases, linkage analyses and candidate gene approaches achieved a humble success.

5.2 The GWAS era

The HGP encouraged several big projects seeking for the understanding of the human genetics, which rapidly led to the connection with human health (Hood and Rowen 2013; Hofker et al. 2014). The International HapMap Project fostered the progress of array technologies, which resulted in **affordable DNA arrays** allowing for the first time systematic and genome-wide interrogation of the role of common genetic variation in susceptibility to human diseases (Hofker et al. 2014; Price et al. 2015). **Genome-Wide Association Studies (GWAS)** represents one of the most relevant methodological advances in human genetic research and is a direct outcome from the completion of the human genome sequence and the parallel technological development of genotyping arrays (Hofker et al. 2014; Price et al. 2015). Through the interrogation of hundreds of thousands of SNPs, GWAS look for statistically significant differences in allele frequencies in large cohorts of thousands of cases and controls at each polymorphic site, providing associations between genetic variants and disease susceptibility (Manolio et al. 2009; Price et al. 2015). As a brief example, if one specific allele from one SNP is far more represented in cases than in controls, this variant will show statistical association with that disease, and this specific allele will be considered a risk allele (see Figure 5, left) (Hofker et al. 2014). The genes that are located nearby this associated variant can provide novel hypothesis on the pathophysiology of the studied disease. GWAS is a hypothesis-free approach agnostic to any prior biological guidance (Manolio et al. 2009; Visscher et al. 2012). GWAS is grounded in the LD principle whereby a real association from a disease-causing variant is achieved through genotyped variants in LD with the first one (see Figure 5, right).

The degree of the association signal will be determined on the strength of the LD between the causal and the tested or tagSNP, which partially depends on the allele frequencies between them: a disease causal variant exhibiting a rare allele will be poorly correlated (low LD) with a common variant tested, and the resulting p-value from the association will be statistically negligible. Thus, GWAS were conceived to capture association signals from causal variants that should be common in the population because this approach was steered by the CDCV hypothesis (Manolio et al. 2009; Visscher et al. 2012; Hofker et al. 2014). In the next section I am going to detail some practical issues from a statistical point of view that should be addressed in this kind of analysis.

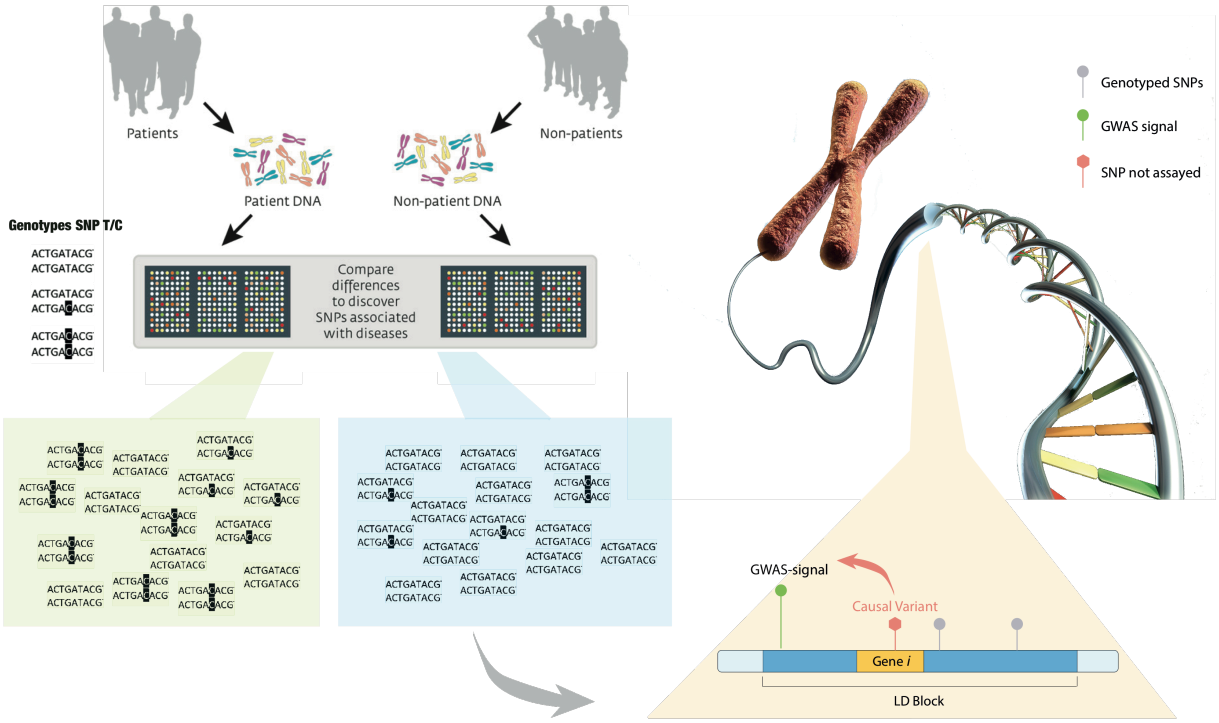


Figure 5. The GWAS approach. At the left, the workflow represents a case-control study in which some healthy controls and patients have been genotyped. To show an example of the allele frequencies comparison between cases and controls, a fraction of the genotypes for a given T/C SNP are represented, underscoring the large representation of the C allele in cases compared to controls. At the right, the underlying rationale of GWAS is based on LD haplotype blocks. Disease-causing variants are captured through a proxy tagSNP in sufficient LD to statistically transmit this association.

5.2.1 Statistics of GWAS

5.2.1.1 Data quality control

In order to produce replicable and robust GWAS results, the quality of the genotyped data should be checked before any further test to ensure the quality of the genotypes at the variant level as well as at the sample level (Anderson et al. 2010; Zeng et al. 2015).

At the variant level, variants accounting for high missing call rates (proportion of individuals without called genotypes for a given SNP) should be removed. Testing statistical differences of the missing rates between cases and controls is also extremely useful to avoid false positive associations. Also, very low frequency alleles are a usual source of genotyping errors, which can incur spurious associations. Thus, SNPs with MAF < 1-2% are usually removed (Anderson et al. 2010; Zeng et al. 2015). Within association analyses, SNPs extremely deviating from HWE should be excluded. However, departure of HWE can occur by a genuine genetic association and thus, solely checking HWE in controls (e.g. p-value < $1x10^{-6}$) is recommended (Zeng et al. 2015). A loose threshold may be incorporated in order to evaluate the whole cohort (e.g. p-value < $1x10^{-20}$) for specific purposes

such as association studies across multiple phenotypes, in which specific filters for a control subgroup cannot be performed.

At the sample level, missing rates are considered by removing those individuals with a high proportion of variants not successfully called (e.g. missing rate > 1-5%). Adjusting for covariates such as sex is of the outmost importance to robustly test for SNP association. Therefore, the given ascertainment of sex information from the genotyped samples should be revised. Chromosome X data allows estimating the sex outcome, which can be compared with the reported sex values (Zeng et al. 2015). Discrepancies between the estimation and the already reported information are indicators for sample swap. Case-control studies are built upon assumptions such as independence among samples. However, apparently independent subjects can entail hidden relationships, which is a common bias in the association tests. Pairwise identify-by-state (IBS) values allow estimating the PIHAT indicator (Zeng et al. 2015), and those individuals exhibiting values above certain threshold (e.g. PIHAT > 0.125-0.185) (de Bakker et al. 2008; Anderson et al. 2010) are removed.

A major source of spurious associations results from population differences between the case and control groups emerging besides the disease status. All the genotypic differences arising from comparing cases and controls from different population origin corresponds to population structure. Allele frequency differences are the result of including distinctive founder populations that are disparately represented in cases and controls. Therefore, carefully evaluating the population origin of cases and controls is of the outmost importance, which should end with the removal of individuals of divergent ancestry (Anderson et al. 2010). In addition, most usual association tests adequately integrates population substructure information in order to correct for population differences (Zeng et al. 2015). Common techniques for identifying, and subsequently removing subjects showing notable differences in ancestry, are principal component analysis (PCA), based on the genetic correlation among individuals. Alternatively, multidimensional scaling (MDS) identifies meaningful dimensions on the basis of genetic distance as IBS (Anderson et al. 2010; Zeng et al. 2015). Furthermore, the inclusion of genotypes from HapMap populations allows clustering study samples. Afterwards, those outliers showing > 3-4 standard deviation from the mean of 2-4 main component vectors are removed (Anderson et al. 2010).

5.2.1.2 Association tests

Single-variant comparison of allele/genotype frequencies between cases and controls are the core of GWAS. Each variant is tested with the null hypothesis of no association assuming a genetic model for the disease risk (Balding 2006).

GWAS for complex diseases recurrently employs the additive genetic model to test for association, which is considered to capture the largest fraction of trait variance for complex diseases (Balding 2006). Thus, every additional copy of the minor allele linearly increases (or decreases) the risk of the disease. However, alternative models such as the recessive or dominant genetic model should not

be completely discarded when modelling how genetic variation contributes to susceptibility to disease (Zeng et al. 2015). The additive genetic model can be tested using the Cochran-Armitage trend test, which corresponds to the score test in a logistic regression (Zeng et al. 2015). The statistics of the Cochran-Armitage trend test are conceived to test a null hypothesis of a zero slope after linearly fitting the estimates of the three genotypic risks (Balding 2006). However, case-control studies are better addressed using a logistic regression. Thus, p_{ij} disease risk for an *i* individual and a *j* genetic variant is formulated as:

$$logit(p_{ij}) = \log\left(p_{ij}(1 - p_{ij})\right) = \beta_o + \beta_1 G_{ij}$$

$ß_0 = ß_1 = 0$ corresponds to the null hypothesis of lack of dependence. Under null hypothesis, logistic regression according any different asymptotically equivalent tests like likelihood ratio, score or Wald test have a chi-squared distribution with one degree of freedom (d.f.) (Zeng et al. 2015). In addition, logistic regression allows accommodating covariates such sex, age and importantly, adjusting for population structure by adding principal or multidimensional scaling components (Balding 2006; de Bakker et al. 2008). Moreover, the exponential function of the regression coefficient $ß_1$ in a logistic regression corresponds to the odds ratio (OR) (Szumilas 2010). An OR describes the odds that a certain outcome will occur (i.e. developing a disease phenotype) given a particular exposure (i.e. a genotype). Therefore, the OR shows if the increased dosage of a particular allele confers risk for a certain disease, and it also allows comparing the magnitude of different risk alleles (Szumilas 2010):

$$OR = 1\ Exposure\ does\ not\ affect\ odds\ of\ disease$$

$$OR > 1\ Exposure\ associated\ with\ higher\ odds\ of\ disease$$

$$OR < 1\ Exposure\ associated\ with\ lower\ odds\ of\ disease$$

Moreover, for quantitative phenotypes, linear regression, variance analysis or t-tests are available choices. For longitudinal studies, the method to rely on is survival analysis, by models such as Cox proportional hazards regression (Zeng et al. 2015). Moreover, more sophisticated Bayesian methods have been developed although they are more computationally intensive. In addition, when the genotypes were not experimentally called (either by sequencing or genotyping) but predicted (see following sections), genotype uncertainty should be taken into account (Balding 2006; de Bakker et al. 2008). Predicted genotypes are represented as probabilities for each of the three genotypes. The subsequent uncertainty of the allele dosages can be incorporated by logistic and linear regression models, which will be reflected in the standard error of the beta coefficient (de Bakker et al. 2008).

5.2.1.3 Minimizing spurious associations

The GWAS approach suffers from a dramatic multiple comparison issue, which results in the inflation of Type I error if no specific action is taken. For a single SNP, the traditional significance level is $\alpha = 0.05$. Thus, if a genotyping array should at least have 500K independent genetic variants, 25K false

positive associations would be expected using α = 0.05 as a threshold. To handle this, one adjustment is the Bonferroni correction, which sets significance level by dividing α by the number of tests performed. GWAS were thought to only encompass 1 M independent tests, which implied that GWAS significance was fixed to $5x10^{-8}$ but some authors argued that this significance threshold was conservative (Balding 2006). However, due to the large fraction of rare and low-frequency variants addressed by whole-genome sequencing (WGS), some studies discussed that more stringent thresholds should be applied, but this is a current unsolved question ($3x10^{-8}$ for MAF ≥ 1%, $2x10^{-8}$ for MAF ≥ 0.5% and $1x10^{-8}$ for MAF ≥ 0.1% at LD r2 < 0.8 in European populations) (Fadista et al. 2016). Another approach to correct for multiple testing is False Discovery Rate (FDR), which controls the expected proportion of false positive associations (Zeng et al. 2015). FDR allows the researcher to tolerate a certain proportion of discoveries (rejected null hypothesis) that are false.

To minimize the occurrence of spurious associations, statistically significant SNPs must be also replicated in independent cohorts. This replication samples should be equivalent to the discovery cohort and the association analysis have to be applied identically as in the original study to ensure consistency (Price et al. 2015; Zeng et al. 2015). There are different criteria for interpreting positive replications (Studies et al. 2007), such as looking for the same direction in the effect sizes. Nevertheless, non-replication can result from hidden population structure in both original and replication study or just a consequence of insufficient power due to a small sample size of the replication dataset (Zeng et al. 2015).

Another important quality control measure is **genomic control,** which is summarized by the lambda λ statistic and measures the extent of the false positive rate**.** The λ factor is calculated from the median of the chi-squared test for the observed values divided by the expected median of the chi-squared distribution, which for one degree of freedom test is ~0.456 (de Bakker et al. 2008; Hinrichs et al. 2009). A λ > 1 is an evidence of a systematic bias or the action of population stratification. Thus, the chi-squared statistics of the genetic markers should be corrected through dividing them by the lambda estimator (Hinrichs et al. 2009). In addition, the Quantile-Quantile (Q-Q) plot representation is a useful tool to identify deviations of the observed distribution from the expected null (see Figure 6, left). This representation compares in a scatter plot the –log of the observed and the expected p-values. However, genuine association signals will deviate from the expected null distribution but they should only represent a small fraction of the observed p-values. Thus, removing known associations of the Q-Q-plot is recommendable to observe if the null distribution can be recovered (de Bakker et al. 2008; Zeng et al. 2015). Of note, this genomic control measure λ is inflated with the increase in the sample size in the presence of polygenic inheritance, even without the action of a confounding bias (Yang et al. 2011). During the development of a previous meta-analysis for schizophrenia, the LD score regression method was developed to discern inflated test statistics from confounding bias and polygenic inheritance (links per marker X^2 summary statistics and linkage disequilibrium). Thus, if inflation were driven by polygenic inheritance, the X^2-statistics

would be linearly proportional to the LD Score: higher LD Scores have higher probabilities of capturing by LD a risk-conferring variant, which tend to also have elevated test statistics (Schizophrenia Working Group of the Psychiatric Genomics 2014; Bulik-Sullivan et al. 2015).

Finally, the p-values of a GWAS are commonly represented by a Manhattan plot that is based on scattering the p-values in the $-\log_{10}$ scale (y-axis) and the physical position of each SNP across every single chromosome (x-axis). The $-\log_{10}$ scale facilitates highlighting smaller p-values, which have the higher potential of being associated with a disease (see Figure 6, right)

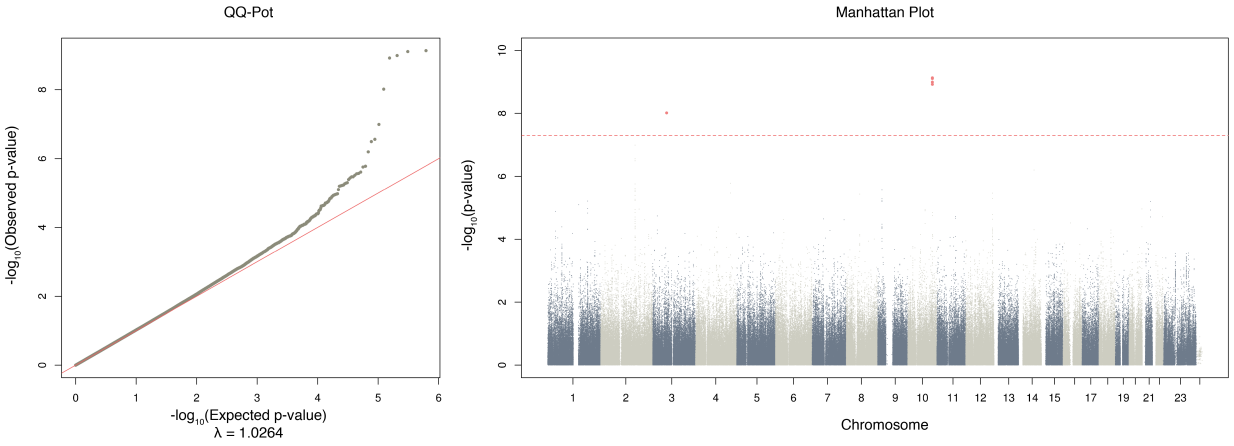


Figure 6. Q-Q and Manhattan plots of a GWAS analysis. Q-Q plots shows the expected $-\log_{10}$ p-values under the null hypothesis (x-axis) respect to the observed $-\log_{10}$ p-values (y-axis). The λ is the measure of genomic inflation and is calculated by the observed median $\mathcal{X}^2$ test statistic divided by the expected median $\mathcal{X}^2$ test statistic under the null hypothesis. In the manhattan plots, the chromosomal position is represented in the x-axis and in the y- axis, the statistical significance ($-\log_{10}$ p-value) of the association test. The red line shows genome-wide significance level (p-value $\leq 5\text{x}10^{-8}$).

5.2.1.4 Meta-analysis

Single GWAS are sometimes underpowered to capture weak effect sizes attributed to common variation, or association with low-frequency or rare variants, which requires increasing as much as possible the sample size. The large number of independent studies carried on simultaneously for a same disease brings the opportunity to combine these datasets via meta-analysis (de Bakker et al. 2008; Zeng et al. 2015). There are different methods for GWAS meta-analysis but the simplest approach is a **p-value meta-analysis**. P-values can be combined with the Fisher's method (Begum et al. 2012; Evangelou and Ioannidis 2013):

$$X^2 = -2\sum_{i=1}^{k} \log\left(p_i\right)$$

where a X^2 follows a chi-squared distribution, p_i corresponds to the p-value of the *i*th study and *k to* the number of studies. The downsides of this approach are that the overall estimates of the effect

sizes cannot be computed, between-dataset heterogeneity is not appropriately addressed and there is a lack of consensus for optimal weighting. Alternatively, a closely related approach to the Fisher method is z-scores (de Bakker et al. 2008; Willer et al. 2010; Begum et al. 2012; Evangelou and Ioannidis 2013):

$$Z = \frac{\sum_i Z_i w_i}{\sqrt{\sum_i {w_i}^2}}; where\ w_i = \sqrt{N_i}\ and\ Z_i = \phi^{-1}(1 - \frac{P_i}{2})$$

This approach takes into account the direction of effects and shows a rather straightforward selection of weights through the sample size values.

The second main approach is based on combining effect-sizes estimates such as via **fixed effects meta-analysis,** which is commonly used. The most accepted method is the inverse variance weighting of the effect size estimates (ß coefficients) by the standard errors (Willer et al. 2010; Begum et al. 2012; Evangelou and Ioannidis 2013):

$$Z = \frac{\langle \beta \rangle}{\langle SE \rangle};\ \langle \beta \rangle = \frac{\sum_i \beta_i w_i}{\sum_i w_i}\ ;\ \langle SE \rangle = \sqrt{\frac{1}{\sum_i w_i}};\ w_i = \frac{1}{{SE_i}^2}$$

This approach is characterized by considering equivalent effect-sizes and standard errors across cohorts, which in a scenario where a substantial amount of heterogeneity between cohorts is present, the results can be biased. In those cases, instead of fixed effects, **random effects meta-analysis** can be used to combine effect sizes. The random model does not assume the same mean effect across studies and it is able to estimate the degree of heterogeneity, which is incorporated into the weight of each study.

Finally, in order to generate robust results, heterogeneity should be minimized. Many metrics have been developed to test heterogeneity but the most widely applied statistic is I^2 statistic, which measures heterogeneity as the proportion of the total variation between studies not attributable to the sample error. Depending on the I^2 values (0-100%), different categories have been established: 0-25% represents ignorable heterogeneity, 25%-50% answers for low heterogeneity, 50-75% corresponds to moderate heterogeneity and 75-100% means high heterogeneity (Evangelou and Ioannidis 2013; Zeng et al. 2015).

5.2.2 Progress in the understanding of complex diseases through GWAS

The first successful GWAS addressed the genetics of age-related macular degeneration (AMD) using ~100K SNPs through 96 cases and 50 healthy controls (Klein et al. 2005). Nowadays, this is much more than a straitened sample size and a poor genomic coverage. However, Klein, R.J. and colleagues were able to identify an intronic common variant strongly associated (p-value = 4.1×10^{-8}) and showing a 7.4-folds increase in disease risk for homozygous individuals for the risk allele (Klein et al. 2005). However, what we know as the GWAS era began with the publication of the Wellcome

Trust Case Control Consortium (WTCCC) study in 2007 in Nature. This major study evaluated the association of ~500K SNPs to seven diseases across 17,000 individuals (14,000 cases and 3,000 shared controls) and identified 24 statistically associated *loci* from which 12 corresponded to previously known regions (Wellcome Trust Case Control 2007; Visscher et al. 2012; Hofker et al. 2014). The field has undergone an explosive progress, which is clearly illustrated with 2,510 studies and 24,065 SNP-trait associations documented to date (http://www.ebi.ac.uk/gwas/home) (Welter et al. 2014).

Despite a large number of identified associations, GWAS signals are characterized by relatively small contributions to disease susceptibility, implying odds ratio of 1.1 and in very concrete occasions, above 1.3 (Hofker et al. 2014; Price et al. 2015). This observation is in line with a common disease whose genetic architecture is articulated across a large number of variants contributing with small effects-sizes. Sample size, allele frequency and effect sizes strongly determine the final statistical power to detect a novel association (Visscher et al. 2012; Price et al. 2015). In rare occasions, a single small GWAS can succeed in unravelling a large fraction of heritability. This scenario would only be successful in complex traits for which most of the associated genetic factors show large effect sizes such as in AMD (Price et al. 2015). Nonetheless, in order to identify common variation of weaker effects, strong initiatives based on the collaboration and data sharing between several groups studying the same specific disease have been critical (Price et al. 2015). The GWAS community integrated this premise and it has been characterized by the creation of several international consortia for several complex diseases.

5.3 Genotype Imputation: a new lease of life for GWAS

Collaborative approaches based on pooling samples and combining summary statistics results from several studies such as meta-analysis were fostered by genotype imputation techniques. **Genotype imputation** was introduced in 2007 (Wellcome Trust Case Control 2007; Marchini and Howie 2010) allowing that markers not directly genotyped for the study individuals can be replaced by genotype predictions. Therefore, genotype imputation allows increasing the number of variants to test for association beyond the initial limited fraction of markers genotyped (Marchini and Howie 2010).

This statistical process begins with a target cohort of samples genotyped at a limited subset of polymorphic sites and a reference panel of individuals typed at a dense set of SNPs or directly sequenced (see Figure 7). Variants at the target typed array have been also interrogated in the reference panel, which allows conceptually accommodating the genotyping backbone of the target individuals within the reference haplotypes. Then, stretches of haplotypes shared between the target and the reference panel are estimated. Finally, by the larger allelic correlation structure in the reference panel, the genotypes for all those missing markers in the target array can be predicted (Huang et al. 2015; Price et al. 2015) (see Figure 7).

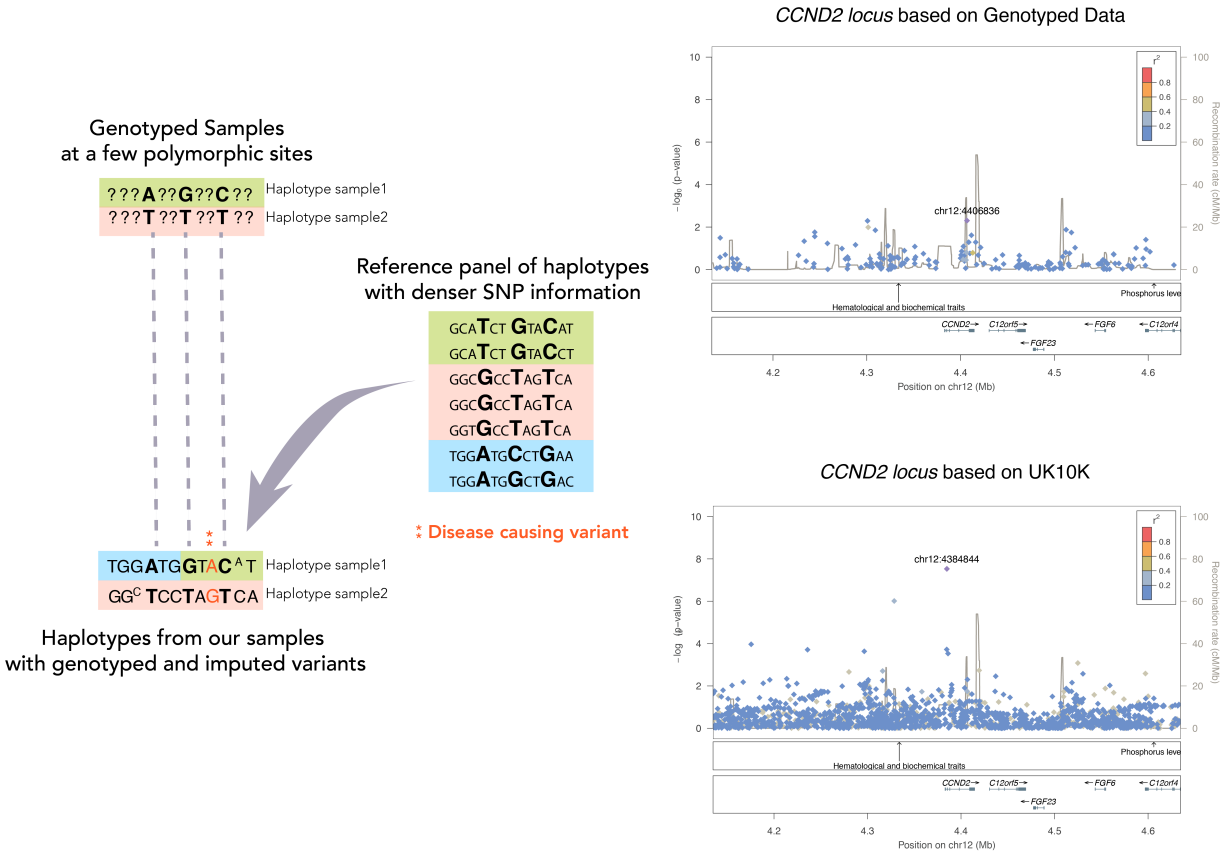


Figure 7. Genotype imputation technique. A study sample set has been typed at a few SNPs while a reference catalogue of haplotypes has been typed or sequenced providing high dense SNP information. By integrating the genotyping backbone of the target samples (top left) with the reference panel (left figure, at the middle), stretches of haplotypes shared between both datasets are estimated. Finally, untyped variants will be predicted by selecting those dense haplotypes blocks from the reference panel closely matching each target individual haplotype. To illustrate the benefits of imputation, we made use of the locuszoom representation for the *CCND2 loci*, showing at the Beagle-axis the $-\log_{10}$ of the association p-value derived from the the logistic regression. The colours highlight the R-squared with the index SNP (in purple). Two locuszooms are represented, showing the poor coverage from the genotyped data and the impossibility of replicating a true T2D association in the *CCND2* gene. In contrast, performing genotype imputation with the UK10K reference panel increased the coverage and captured this true intronic low-frequency association for T2D in the *CCND2* gene.

5.3.1 Advances in Genotype Imputation calculations

In order to find compatible haplotypes for the markers genotyped in the target individuals, the haplotypes for the latter ones should be estimated from the genotypes. This **"phasing"** process takes the genotypes of a set of genetic variants (i.e. the genetic makeup, the genetic composition for each polymorphic site considering the two possible alleles) and estimates the haplotypes (i.e. ordered sequence of alleles from multiple polymorphic sites that are inherited together). Besides being the core engine of imputation methodologies, phasing is also relevant for the study of genetic diversity (International HapMap et al. 2007) or the identification of selection (Sabeti et al. 2007).

A revolutionary change in genotype imputation has been the introduction of **"pre-phasing" prior to imputation** (Howie et al. 2009). Initially, genotype imputation was performed directly on genotypes: these methods identified those typed variants in the target samples and also present in the reference panel, phased them and looked for almost perfect matches between the resulting target haplotypes and the haplotypes in the reference panel (Scheet and Stephens 2006; Marchini et al. 2007). According to this rationale, if there was a match between some haplotypes from the reference panel and the haplotypes built on the basis of the backbone genotypes typed in the target samples, the reference haplotypes had also to match the genotype content of the unknown fraction of genetic variants that had to be estimated. This approach was computationally demanding and was replaced by the pre-phasing step followed by genotype imputation, which is now widely embraced by the whole community. However, despite the reduced computational complexity derived of only imputing alleles, there is a slight decrease in accuracy using this approach (Browning and Browning 2016). Thus, it has been critical generating robust methods for phasing, which have led to a rapid progress.

Since the beginning, most of the existing methods have exploited Hidden Markov Models (HMMs) to iteratively estimate an individual haplotype driven by a set of SNP genotypes on the basis of haplotypes of other individuals. The **most accurate method, until some recent publications in this 2016, has been the SHAPEIT2 algorithm** (Delaneau et al. 2013) that integrated features from the previous SHAPEIT1 method (Delaneau et al. 2012) and the IMPUTE2 phasing (Howie et al. 2009) approach. SHAPEIT2 also allowed multithreading (Open Multi-processing, OpenMP, parallelism framework) for a more efficient use of the computational resources and was able to exploit long stretches of haplotypes shared by samples or **long-range phasing (LRP)**.

However, with the advent of **biobanks and huge datasets**, sample sizes have increased above 10,000 individuals, which can **dramatically impact the computational time in HMM-based phasing methods such as SHAPEIT2**. Other methods have been applied on large datasets (~60,000 individuals) like HAPI-UR (Williams et al. 2012) but the accuracy was reduced in comparison with SHAPEIT2. Recent solutions addressed this challenge such as SHAPEIT3 (O'Connell et al. 2016) and EAGLE (Loh et al. 2016), that in both cases have been applied in the UK Biobank data. SHAPEIT3 recognized the HMM limitations and provided some improvements with

respect to SHAPEIT2 such as benefiting from the increased local similarity between groups of haplotypes due to a recent shared ancestry (O'Connell et al. 2016). EAGLE collected the ideas behind LRP that integrated with conventional HMM-based methods, which retained the accuracy of HMM-based approaches but with a notable increase in computational efficiency (Loh et al. 2016).

Regarding the imputation process, there are several methods available but in any case, they have benefited from parallel computing that can enhance the execution by using multithreading or by just simply imputing different genomic regions independently in a cluster infrastructure. Most widely accepted methods are based on HMM such as IMPUTE2 and the recent updates from the minimac3 and the Beagle 4.1. A recent work presenting the novel and improved minimac3 showed that for huge reference panels involving dozens of thousands of individuals, minimac3 was twice as fast as Beagle 4.1 and 30 times faster than IMPUTE2 and the memory usage was reduced by 72% and 97%, respectively (Das et al. 2016). Regarding the imputation quality, all methods have similar performances, but minimac3 slightly outperformed Beagle 4.1 and IMPUTE2 at the range of $0.0004\% \leq MAF \leq 0.5\%$ (Browning and Browning 2016; Das et al. 2016).

5.3.2 Application of genotype imputation

The application of genotype imputation triggered three main benefits (1) facilitating **meta-analysis**, (2) increasing the statistical power for **GWAS discovery** and (3) improving **fine-mapping**, which are illustrated with the locuszooms before and after genotype imputation in Figure 7. First of all, when performing meta-analysis, genotype imputation allows homogenizing the SNP coverage across different cohorts, as some variants might have been typed in one cohort but not in the other. Moreover, genotype imputation is able to substantially increase the number of available variants for association testing, specially using sequence-based reference panels, which will increase the pace rate of GWAS discovery. In the example shown in Figure 7, the *CCND2* locus, which is known to be associated with T2D through a low-frequency variant, can only be identified by means of genotype imputation. Finally, this increase in the number of genetic variants for association testing is also translated in a much higher genomic resolution. Thus, for those reported *loci* encompassing multiple potential causal variants, genotype imputation facilitates pinpointing the most plausible disease-causing variant or the underlying biological mechanism (Marchini and Howie 2010; Price et al. 2015; Browning and Browning 2016).

Nonetheless, genotype imputation is a statistical prediction, which always goes hand in hand with a **certain error rate**. The accuracy associated to the prediction of common variants is very high. However, there is a rapid decline of the imputation accuracy in the vicinity of the rare a low allele frequency range (Huang et al. 2015; Price et al. 2015). The accuracy of the genotype prediction is tied to several parameters such as the coverage and the quality of the genotyped array of the target samples. In addition, the quality of the phased genotypes into haplotypes, how limited is the

representation in the reference panel of haplotypes carrying rare alleles or differences in LD patterns are other parameters influencing the quality of the imputation (Pistis et al. 2015).

Inadequacy of reference panels is one of the major limitations for accurately impute genotypes, such as small sample sizes like in HapMap that results in notable errors even for common variants (de Bakker et al. 2008) . In addition, imputation accuracy has been suggested to improve by selecting a reference panel closely matching the ancestry of the study population (de Bakker et al. 2008; Huang and Tseng 2014). Nevertheless, in studies with no clear reference match, most experts recommended using a cosmopolitan reference panel (Huang and Tseng 2014). Howie, B. and colleagues demonstrated that larger and diverse reference collections facilitate identifying shared stretches of haplotypes (Howie et al. 2011).

The first published genome-wide imputation analysis relied on the HapMap2 and (afterwards the HapMap3) reference panel, which included 60 CEU individuals typed at 2.1 M markers (International HapMap et al. 2007). The advent of large-scale sequencing and the success of genotype imputation encouraged the creation of several WGS projects to boost variant coverage and imputation quality across the whole spectrum of allele frequency. The first large-scale sequencing project was the *1000 Genomes Project* (1000G), which was conceived to study and to identify human genetic variants showing frequencies of at least 1% in the population but also to provide accurate haplotype information on any type of DNA polymorphism across multiple populations (The 1000 Genomes Project Consortium et al. 2010; Birney and Soranzo 2015). This project ran from 2008 until 2015, divided in 4 stages, a pilot phase and three phases in the main project (although the second phase of the main project was devoted to the technological development). While the pilot phase only identified 14.8 M variants in 179 individuals from four populations, the phase1 of the main project provided 37.9 M variants in 1,092 individuals in 14 populations and in the final Phase 3, 84.4 M variants were catalogued by sequencing 2,504 individuals from 26 populations (http://www.1000genomes.org/data). In line with the progress in terms of sample size and population diversity, ancestry-matching reference panels have been created, samples sizes from reference data are increasing from a few to tens of thousands of individuals and a major focus is also placed in reaching high sequencing depth (Browning and Browning 2016) (see Table 1). An illustrative example is the UK10K project, which sought for a precise characterization of rare and low-frequency variants in the UK population. This data has been used to study the contribution of variants with lower allele frequencies to multiple biomedical relevant and disease conditions but it also has become one of the most relevant resources for genotype imputation (UK10K Consortium et al. 2015). The UK10K consortium assembled whole-genome sequences of ~4,000 British volunteers, exhaustively surveying genetic variation down to 0.1% MAF in the British population. Another example focused on enhancing the characterization of rare variants is the Genome of the Netherlands (GoNL) Project (Genome of the Netherlands 2014). GoNL sequenced 769 Dutch individuals of 250 families at ~13x depth, resulting in 20.5 M SNPs and 1.2 M INDELs, and

extensively capturing structural variation and *de novo* mutations (Genome of the Netherlands 2014). In-house reference panels composed by sequencing a subset of samples from the target cohort of study has been proved to be a valid approach to create genetically similar data to the study samples (Huang and Tseng 2014). Finally, these independent efforts encouraged the creation of a unified reference panel across populations, the Haplotype Reference Consortium (HRC, http://www.haplotype-reference-consortium.org/home). The HRC is the largest panel for imputation and encompasses other sequencing projects such as the 1000G-Phase3, GoNL, UK10K, SardiNIA among others, and additional ~30,000 samples of European ancestry. In the following sections some of these projects will be revised to explain some of the challenges and limitations of genotype imputation.

Table 1. Overview of publicly available reference panels. For each cohort, the number of individuals, the sequencing depth and the ancestry of the population are described.

Cohort	*N* Samples	Depth	Ancestry	Accessibility
HapMap3	1,084	Genotyped	Multi-Ethnic	No restrictions
UK10K	3,781	6.5x	UK-European	EGA
GoNL	748	12x	Dutch-European	EGA
1000G-Phase3	2,504	4x/Exome	Multi-Ethnic	No restrictions
Singapore Sequencing Malay Project (SSMP)	100	30x	South-East Asian	No restrictions
GoT2D	2,974	4x/Exome	Europeans	EGA
Haplotype-Reference Consortium (HRC)	38,821	Diverse	Europeans	HRC imputation sever

5.4 Stress tests for the GWAS statistical rationale

5.4.1 Empowering the interrogation of low-frequency and rare variants

Most of the past genotype imputation based GWAS discoveries were articulated on the basis of the HapMap reference scaffold (Huang et al. 2015) focusing on common variation, whereby the role of

non-SNV polymorphisms or low-frequency and rare variation was ignored. The successive projects developed with the advent of NGS technologies that I detailed previously illustrate the big opportunity for enlarging the landscape of human genetic variation (Hood and Rowen 2013; Price et al. 2015). For instance, the recent Phase 3 release from the 1000G expanded the initial thousand individuals to 2,504 sequenced individuals from 26 geographic locations, yielding in ~88 M variants (84.7 M SNPs, 3.6 M INDELs and 60K SV) (The 1000 Genomes Project Consortium et al. 2015). This improvement served to demonstrate that **haplotypes identified by GWAS were enriched with common SVs** by more than three folds, urging exploring a larger spectrum of genetic variation on disease susceptibility (Sudmant et al. 2015). One of the main limitations of genotype imputation for GWAS approaches was ascertaining low-frequency and rare variants and different strategies were suggested to break these constraints (Huang et al. 2015; Kim et al. 2015; Price et al. 2015). One strategy is based on increasing the **sample size of the reference panel** by combining as many sequencing data as possible, even from diverse populations, such as the extensions that the 1000G underwent or the rationale beneath HRC (Kim et al. 2015; Price et al. 2015).

Alternative strategies are based on **genetic studies on isolates**, in which rare variants can rise high frequencies because of founder effects or genetic drift (Zavattari et al. 2000). For instance, a dramatic increase in T2D prevalence in the small and isolated Greenlandic population fostered an association mapping study of four T2D-related traits. A novel common and missense variant (OR = 10.3, p-value = 1.6×10^{-24}) showed strong association with T2D by terminating a long isoform of *TBC1D4* causing very specific phenotypes (Moltke et al. 2014). This example is a proof of concept of the opportunity brought by GWAS outside traditional large homogenous populations.

However, isolated populations only offer a gain in statistical power for a limited number of rare variants that drift to higher allele frequencies. In order to study the full spectrum of rare variation, **large-scale studies** of several thousands of individuals in several populations are necessary. This vast sample sizes can only be reached by collaborative efforts among several centres which yielded to the creation of biobanks (Price et al. 2015). Large population biobanks integrate genome-wide genetic information with large amounts of phenotypic information, lifestyle, diet and other environmental exposures (Price et al. 2015). An example of that is the UK Biobank, a large and prospective study comprising 500K individuals. The UK Biobank still keeps on the collection of genotypic and phenotypic information, involving questionnaires, physical measures and sample assays for a longitudinal follow-up for different health-related outcomes (Price et al. 2015; Sudlow et al. 2015). Additionally, analysing such a huge British sample size from the UK Biobank through a population specific reference panel such as the UK10K Consortium is expected to provide very accurate results. Actually, the **UK10K project has demonstrated to be an effective solution when ascertaining rare and low frequency imputed variants in UK but also Italian populations** (Huang et al. 2015; UK10K Consortium et al. 2015).

Finally, sequencing data itself constitutes a unique opportunity to disentangle the role of low-frequency and rare variation in complex diseases. WGS of large cohorts as in a GWAS approach is prohibitively expensive. Alternative approaches consist in targeted gene sequencing, whole-exome sequencing (WES) of the 1-2% of the genome coding for proteins, low-depth WGS, rare-variant genotyping arrays or extreme phenotype sampling, which is based on sampling at the extreme of the trait distribution (Lee et al. 2014; Pistis et al. 2015). In addition, association analysis of low-frequency and rare variants are still underpowered and novel methods have been developed. These methods are based on **aggregating the association signal** from multiple variants into **biologically relevant units such as genes**, rather than testing single-variant effects (Lee et al. 2014). There are different methods but they can be broader categorized as burden or variance-component test (Pistis et al. 2015). A burden approach is based on aggregating carriers of rare variants within a gene and comparing their phenotype or disease susceptibility with the fraction of non-carriers. This approach is limited by the consideration that rare alleles contribute in the same direction. The second wave of rare-variant association tests considered a distribution of the genetic effects, such as SKAT tests, which is able to modulate prioritization and weighting strategies (Lee et al. 2014; Pistis et al. 2015). Meta-analysis has also been accommodated for rare-variant association studies on the basis of score summary statistics per individual variant and a matrix summarizing LD correlation patterns between markers. This strategy has been implemented in rareMETAL or skatMETA packages (Lee et al. 2014; Pistis et al. 2015).

5.4.2 Genetic inequalities: the X-chromosome exclusion

When breaking down the large number of SNP-trait associations reported in the NHGRI GWAS catalogue, there is an obvious **underrepresentation** of the X-chromosome. Actually, only a third part (33%, 242 out of 743) of the GWAS publications included the X-chromosome in their analysis as denoted for the period ranging from 2010 to 2011 (Wise et al. 2013; Konig et al. 2014; Kukurba et al. 2016). In addition, although X-chromosome comprises 5% of the human genome content (Wise et al. 2013; Tukiainen et al. 2014), encompasses 1,500 genes and is comparable in size with the chromosome 7 (Tukiainen et al. 2014), the NHGRI GWAS catalog (Welter et al. 2014) only reported 55 SNP-trait associations (p-value ≤ $5x10^{-8}$) while chromosome 7 accounts for ~280. Gathering up all these observations the question of why X-chromosome data remains underutilized has risen. A consensus suggestion is that the need of specific analytical methods for processing and interpreting X-chromosome variation impaired the analysis of the X-chromosome in GWAS publications. Furthermore, the feeling that the myriad of autosomal associations was sufficient to achieve a high-profile publication has also been reported behind this phenomenon (Wise et al. 2013; Kukurba et al. 2016). In addition, large-scale functional genomics have also excluded the X-chromosome from their analysis (Kukurba et al. 2016).

What make X-chromosome special in such a manner that discourages researchers to include it in their analysis? There is an asymmetry of the genetic dose between females and males (females have two copies of the X-chromosome while males, one). However, allele dosages are balanced through an **inactivation process** of the X-chromosome at the early stage of development in females. Therefore, X-inactivation varies throughout the body and tissues but of note, 15% of the *loci* completely escape this X-inactivation and for another 10%, this process is variable. Therefore, we have a high heterogeneity that challenges analytical methods. Furthermore, less proportion of data compared to autosomes under-power X-chromosome GWAS to detect variants with modest effect-sizes (Tukiainen et al. 2014; Kukurba et al. 2016). Therefore, these particularities should be included in the imputation process or the association analyses but the majority of the software required for any of these steps such as SHAPEIT2, IMPUTE2 or SNPTEST have developed specific workflows to cope with that.

5.4.3 Poor disease understanding keeps clinical translation out of the picture

The deployment of the four P healthcare (Predictive, Preventive, Personalized and Participatory) results from the convergence of systems medicine, big data and patient involvement. Human genomics can contribute to the first three Ps through successive steps beginning with the identification of the genetic association until unravelling the underlying disease causal mechanism. The correct ascertainment of disease-causal variants would empower risk prediction models to better illustrate disease susceptibility and to provide personalized and more effective disease prevention strategies. For instance, sodium-glucose transporter 2 (SGLT2) inhibitors have been consolidated as a novel class of oral anti-diabetic agents (Nauck 2014) that yield to reduced hyperglycaemia and in some cases, to lower cardiovascular events (Zinman et al. 2015). Indeed, these agents reproduce the physiology of familiar renal glucosuria, which is caused by a loss-of-function mutation in *SLC5A2*, the gene coding for SGLT2 (Santer et al. 2003). Moreover, drug efficacy is tied in some cases to some rare and common genetic variants and the translation of GWAS signals to molecular mechanisms is crucial to identify novel “druggable” components and pathogenic key pathways (Hofker et al. 2014; Paul et al. 2014).

Despite the undeniable success of GWAS in the identification of novel associations with disease susceptibility, the impact of these findings into the clinical practice is still minimal (Hofker et al. 2014). There are two main arguments behind this arduous translation. First, despite the build-up of novel GWAS associations, **a large fraction of the estimated heritability still remains unexplained** for the majority of complex diseases. Therefore, common SNPs are not the unique genetic answer behind complex diseases, which gave way to the involvement of rare variation with stronger effects, epistatic mechanisms and unknown interactions between gene and environment (Manolio et al. 2009; Hofker et al. 2014; Price et al. 2015). In line with that, extensive genotyping, improved genotype imputation methods and sequencing data is crucial to capture all those variants falling

outside linkage correlation patterns tied to tag SNPs (Hofker et al. 2014; Paul et al. 2014; Price et al. 2015). In second place, the **functional interpretation of disease-associated variants is limited**, which results in a poor understanding of the molecular mechanisms triggering most complex diseases. A key step is the identification of real genetic disease causal variants, which can be difficult due to a masking effect of LD correlations. A GWAS discovery focuses our attention on a broad region with a huge number of correlated SNPs. In order to unveil which is the true disease-causing variant the recurrent and standard practice is fine-mapping. **Fine-mapping** relies on performing a dense interrogation of complete associated *loci*. Fine-mapping approaches benefit from using a wide range of ethnicities to span the number of LD patterns, narrowing the number of candidate SNPs (Paul et al. 2014; Price et al. 2015). In parallel, extensive efforts have been dedicated to elucidate the **downstream effect of associated signals**. However, the main focus has been directed towards protein coding altering variants but this approach has not been extremely resourceful, as the core of the underlying biology from most GWAS signals (above 90%) corresponds to **perturbations in gene regulation** (Ward and Kellis 2012d; Hofker et al. 2014; Paul et al. 2014; Price et al. 2015). Our still limited understanding of gene regulation impairs the direct interpretation of the functional effects driven by the association signals. Moving up the complexity ladder, gene regulation is highly tissue/cell-specific, tied to the developmental stage and to external stimulus or environmental factors (Paul et al. 2014; Price et al. 2015). In addition, trait-associated signals perturbing gene regulatory elements may affect the transcriptional output from a distal gene, that would be difficult to pick up (Paul et al. 2014). The need of a systematic interpretation of noncoding disease associated signals has led to the emergence of large-scale projects providing reference genome **functional annotation maps**. The observation of enrichment of regulatory biochemical signatures in GWAS *loci* can guide hypothesis of the undergoing regulatory disease mechanism. The Encyclopedia of DNA Elements Project Consortium (ENCODE) (Encode Project Consortium 2012) has released functional maps of chromatin states, transcriptor factor binding sites and gene expression for several and mostly derived cell lines. The NIH Roadmap Epigenomics Mapping Consortium (Roadmap) (Bernstein et al. 2010) or the BLUEPRINT Consortium (Adams et al. 2012) have focused on the construction of epigenome maps mainly based on primary tissues (Ward and Kellis 2012d; Paul et al. 2014). These public functional maps in conjunction with the development of computational resources empowered researchers to prioritize non-coding variants (Ward and Kellis 2012d; Flannick and Florez 2016). However, choosing the suitable cellular system for the annotation of GWAS *loci* is not trivial. Primary tissues are more direct and real tissue representatives while cultured cell lines retain main characteristics of primary tissues. However, in cultured cell lines, chromatin structure and DNA methylation perturbations or even chromosomal rearrangements are frequent, which can mislead functional interpretation of GWAS discoveries (Paul et al. 2014; Price et al. 2015). Of the outmost importance is selecting the most informative annotation mark. Authors recommended using open chromatin marks which are general hallmarks for most regulatory elements but they lack specificity;

ChIP-seq (chromatin immunoprecipitation sequencing) of histone modifications informing about active promoters and enhancers but imprecise broader peaks impairs clear elucidation of candidate functional variants (Paul et al. 2014). A common approach to unravel the role of genetic variation on gene expression is **expression quantitative trait locus (eQTL) analysis. eQTL analysis** addresses the association between genetic variants and variation in the expression levels of mRNA (Figure 8) (Morley et al. 2004; Grundberg et al. 2012; Westra et al. 2013) and they have been highly used to prioritize functional regulatory variants (Ward and Kellis 2012d; Paul et al. 2014).

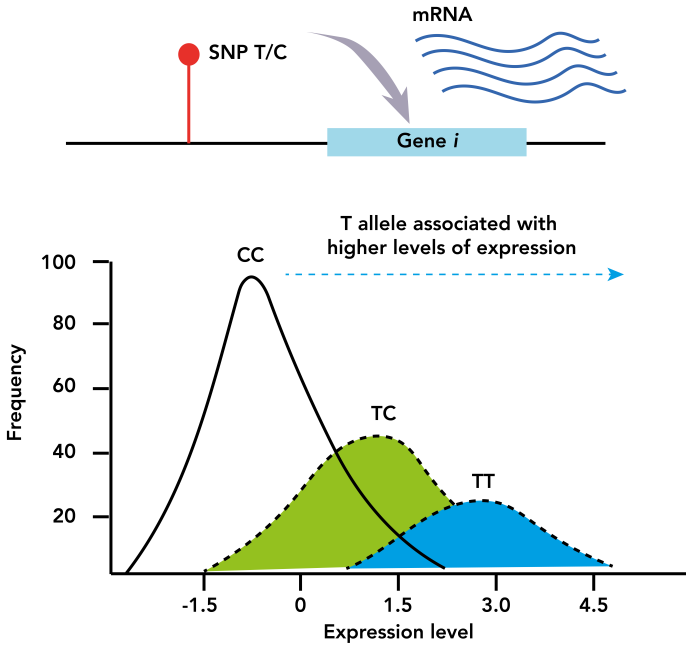


Figure 8. Example of a typical eQTL. At the top of the figure, a candidate eQTL for differentiated levels of expression of the gene *i*. At the bottom, a representation of how the different genotypes of this candidate SNP are associated with different distributions of expression values for the gene *i*. This example illustrates that the T allele is associated with higher expression levels of the gene *i* (Adapted from (Nica and Dermitzakis 2013)).

The Genotype-Tissue Expression (GTEx) project was conceived to provide a data resource to enable research to study the relationship between genetic variation and gene expression in multiple human tissues (GTEx Consortium 2013). The pilot phase resulted in the analysis of RNA sequencing 1,641 samples across 43 tissues from 175 individuals that represents the most comprehensive project of gene expression across diverse human tissues (Hofker et al. 2014).

Therefore, multiple tools are available to guide the researcher in the functional interpretation of genetic variation. Later, regulatory functional candidate variants should be experimental assayed to prove molecular function and causality through several experimental assays comprising luciferase reporter assays, gel-shift, or allele-specific chromatin assays (Paul et al. 2014). These strategies for the functional translation of non-coding GWAS associations are urgently required to push forward disease understanding and translation to clinics.

6 Computational sciences and their involvement in genomic research

The advent of NGS technologies stacked more and more sequenced genomes resulting in Terabytes and Petabytes of data requiring for efficient and accurate software solutions. Therefore, this huge amount of data cannot be solely analysed by individual workstations, which stressed the need of incorporating parallelization techniques and High-Performance Computing (HPC) into genetic research HPC infrastructures, large-scale storage and a large and dynamic ecosystem of analytical tools. Typical analysis such as genotype imputation and GWAS encompasses sequential executions involving multiple software tools, which are referred as a "workflow" (Spjuth et al. 2015).

HPC environments consist of clusters, grids or clouds with batch systems for scheduling jobs. Recently, a major focus is placed on cloud computing that relies on shared resources delivered on-demand through Internet. Cloud computing commonly benefits from virtualization, which enables building computing environments independent of physical infrastructure answering the actual computational needs of the users (Schadt et al. 2010; Spjuth et al. 2015). This scenario opened the possibility to package entire analysis in virtual machine images (VMI), or taking profit of the parallelism of distributed environments. Thus, GWAS approaches are being translated to parallel computing, especially with the increase in the number of variants to test resulting from genotype imputation or WGS (Spjuth et al. 2015).

First of all, one major outcome from the integration of computational science in biological research was just engineering or redesigning bioinformatics applications to accommodate parallel computing (Ocana and de Oliveira 2015). Parallel computing is a strategy that enables running multiple executions or instructions simultaneously. Therefore, a program split in independent parts can use a single computer with multiple processors or a network of interconnected computers to run each part in parallel. Examples of interfaces that foster parallel computing are: OpenMP (Open Multi-Processing), for multi-threading on a single shared memory infrastructure and MPI (Message Passing Interface), a communication protocol for multi-processing applications executed in different computing nodes of a cluster (i.e. not shared memory) (Schadt et al. 2010; Yang et al. 2014). For instance, some genotype imputation algorithms are based on multi-threading, such as Beagle (Browning and Browning 2016) or minimac.

Second most genetic analysis can involve large parallelism within the different analytical steps until eventually gathering all the output generated into single final results. This preamble facilitates understanding the rationale behind MapReduce approaches. MapReduce splits a problem into multiple sub-questions in a 'map' step to afterwards, performing a 'reduce' step in which collects and integrates the output of each small question into a single answer (Schadt et al. 2010). Technically, this programming model and implementation for the analysis of large datasets distributes the computational load on multiple connected computing nodes (Schonherr et al. 2012). Thus, time-intensive imputation analyses based on huge reference panels can exploit frameworks based on

MapReduce techniques. For instance, the free service of the Michigan Imputation Server has been developed within the CloudGene technology, a free platform that is built on the usability of MapReduce applications through the Hadoop framework (Schonherr et al. 2012; Spjuth et al. 2015). CloudGene makes use of a user-friendly graphic interface in order to minimize the installation and maintenance of MapReduce on cluster systems, or the data management in a distributed system (Schonherr et al. 2012; Spjuth et al. 2015). Thus, large bioinformatics efforts have been direct to automatizing and assembling analysis pipelines to enhance wide adoption and reproducibility.

7 Data-sharing: pushing forward the pace of GWAS discovery and the molecular understanding of complex diseases

The HGP popularized the strengths of data-sharing initiatives by making available data through user-friendly databases such as GenBank or the UCSC Genome Browser. This shift of the community to free the access of data arose from the concept of 'democratization of data'; giving access to third party researchers, outside from big consortia, is critical to exploit, with additional expertise, public available data in order to improve the understanding of disease biology (Hood and Rowen 2013). Several initiatives to encourage data sharing emerged, such as the creation of centralized repositories for GWAS data, such as the database of Genotypes and Phenotypes (dbGaP) (Tryka et al. 2014) and the European Genome-phenome Archive (EGA) (Lappalainen et al. 2015). The underlying goal is maximizing the scientific outcomes resulting from public funded resources through the application of novel analytical methodologies. In addition, sharing GWAS data allows assembling more powerful case-control studies, by increasing the sample size for exploring modest and weak SNP-trait associations or to achieve sufficient power to test rare variants for association (Johnson et al. 2013). The potential of this approach has been underscored by the advent of genotype imputation techniques, allowing homogenization of genomic coverage. Furthermore, with the availability of novel sequence-based reference panels such as the 1000G or the UK10K project, genomic resolution can be increased by orders of magnitude. Thus, genetic variation that was ignored in the initial study can be ascertained, yielding to new GWAS discoveries without requiring substantial expenditures. Actually, the effect of this GWAS sharing initiatives on enhancing novel research has been evaluated through the several publications resulting from secondary uses of dbGaP data. PubMed reported 924 publications driven by secondary use of dbGaP data and 25% of these studies were published in journals with an impact factor greater than 10 (Paltoo et al. 2014). Therefore, secondary research involving dbGaP can yield to significant achievements in a wide range of fields, such as unknown associations between the Human Leukocyte Antigen (HLA) *locus* and Parkinson's disease (Hamza et al. 2010). Furthermore, the combination of several dbGaP GWAS datasets allowed one of the largest alcohol dependence GWAS leading to novel associated *loci* (Gelernter et al. 2014).

Going back to the limitations of complex diseases, in order to reach a more comprehensive overview of their genetic architecture, increasingly larger cohorts must be interrogated. In addition, in order to

achieve a mechanistic insight from the candidate associations and genes resulting from these studies, a huge amount of resources and a variety of approaches should be undertaken. Besides secondary research studies, approaches such as integrated portals have been conceptualized in order to link genetic experts with experimentalists, and physicians (Flannick and Florez 2016). These portals will respond to all these needs and they should aggregate, harmonize and analyse as much genomic data sets and phenotypic information as possible. The value of this approach is that it would empower genetic association studies by linking data contributors with a variety of researchers in a comprehensive way for this broad community. Therefore, robust and efficient pre-computed analysis will provide extensive genetic information to pursuit a specific hypothesis from a non-necessarily genetic expert user. However, on-demand on-line analysis will always be required. Therefore, these settings should reserve some space to extend their original workflows in order to span the number of possible biological questions to ask, which constitutes a computational conceptual challenge (Flannick and Florez 2016). The Type 2 Diabetes Knowledge Portal (T2D Portal) corresponds to one of these portals, comprising the efforts of more than 100 investigators seeking for a rapid and intuitive access to genetic analysis of hundreds of thousands of samples (The American Diabetes Association 2015) .

To summarize, collaborative and data-sharing approaches are mandatory to commit with our desire to push forward our understanding of the pathophysiology of complex diseases such as Type 2 Diabetes.

8. Type 2 Diabetes: a paradigm of the genetic research in complex diseases

Diabetes Mellitus (DM) is a complex chronic metabolic disease characterized by **high levels of blood sugar** (hyperglycaemia) driven by a depletion of insulin secretion or defective insulin function (Tallapragada et al. 2015). The increasing incidence of DM is disturbing: 415 M of affected population worldwide in 2015. This prevalence has been spurred by the rapid rise in obesity and life-style changes such as the reduced physical activity. Additionally, current estimations stated that DM will be the 7th leading cause of death by 2030 and diabetic patients will rise up to 642 M of individuals in 2040 (International Diabetes Federation 2015). Particularly worrisome are **193 M of population that remain undiagnosed**, which place them at higher risk of developing DM related complications. DM is accompanied by a high rate of morbidity and mortality due to the chronic elevated glucose blood levels on the vasculature, which can result in a progressive loss of vision, renal failure, peripheral and autonomic neuropathy and macrovascular complications (i.e. stroke) (Forbes and Cooper 2013). The long-term support required for these patients has risen the public expenditure 5% to 20% (International Diabetes Federation 2015).

8.1 Type 2 diabetes pathophysiology

90% of the DM cases are suffering from T2D, which are diagnosed on the basis of glucose blood levels after fasting (Fasting plasma glucose test, FPG) or 2h after a glucose challenge (2hGlu, Oral glucose tolerance test, OGTT), or on haemoglobin A1c (HbA1c) that provides a 3-month average estimation of blood glucose (International Expert 2009; Mohlke and Boehnke 2015). The hallmark feature of T2D pathophysiology is an hyperglycaemia driven by a progressive insulin resistance in liver, muscle and adipose tissue and a depletion of the pancreatic β-cell function resulting in hampered insulin secretion (Cornell 2015). Several authors have reported strong evidences for a genetic component for T2D risk, the depletion in insulin secretion and insulin action (Poulsen et al. 1999; Poulsen et al. 2005) and this reduced insulin sensitivity has been shown to co-occur with obesity, adverse lipid concentrations, hypertension and an exacerbated inflammatory state (National Cholesterol Education Program Expert Panel on Detection and Treatment of High Blood Cholesterol in 2002). From a systemic point of view, these molecular defects in the T2D pathophysiology involve at least seven organs and tissues (pancreas, liver, skeletal muscle, adipose tissue, brain, gastrointestinal tract and kidney) (Defronzo 2009), which are represented in Figure 9.

Loss of β-cell function has been genetically associated with the impairment of the pancreatic development, insulin secretion and storage (Grant et al. 2009). The decrease of the **β-cell function in pancreas** has been demonstrated to be age-related (Chang and Halter 2003), which is consistent with the co-evolution of T2D prevalence along with aging (Centers for Disease Control and Prevention (CDC) 2014). Insulin resistance promotes biosynthesis and release of insulin that can lead to an "exhaustion" of the β-cell in the process of adaptation to the large insulin demands (Kahn 2001). Moreover, major T2D risks factors such as obesity and physical inactivity (Hu 2011) are associated with insulin resistance and can consequently led to β-cell failure in the long-term (Defronzo 2009; Hu 2011). Excess of rapidly absorbable carbohydrates from diet increases insulin and blood glucose levels (Hu 2011) while fat deposits in liver and muscle fosters insulin resistance in these tissues (Defronzo 2009). Furthermore, glucotoxicity (chronic exposure to high glucose levels) hampers β-cell function and insulin secretion (Poitout and Robertson 2002) while lipotoxicity (elevated concentrations of plasma free fatty acids -FFA-) impairs insulin secretion and results in the depletion of β-cells (Carpentier et al. 2000; Kashyap et al. 2003).

T2D patients show an **overproduction of glucose in liver**, which has become resistant to the repressive effects of insulin (Defronzo 2009). Elevated levels of glucagon (produced in pancreatic α-cells) also increases the hepatic glucose production (Defronzo 2009). The main mechanism for the uptake of exogenous glucose is the **insulin-stimulated transport of glucose into skeletal muscle (Huang and Czech 2007)**. T2D patients exhibit a decrease in the glucose transport favouring hyperglycaemia because an insensitivity to the effects of insulin (Cusi et al. 2000).

In the adipose tissue, insulin resistance in adipocytes leads to increased circulating FFAs, which fosters gluconeogenesis, hepatic and muscle insulin resistance and impairs insulin secretion (Bays et al. 2004). In addition, defective adipose tissue results in an excess of inflammatory and atherogenic cytokines, which can stimulate insulin resistance, and is unable to adequately secrete insulin-sensitizing adipocytokines (Bays et al. 2004).

T2D cases also **exhibit insulin resistance in the brain**, inhibiting the effect of insulin in satiety, increasing the food intake (Pagotto 2009).

In the gastrointestinal tract, T2D patients were reported to exhibit a lowered release of glucagon-like peptide (GLP-1) and reduced sensitivity to glucose-dependent insulinotropic polypeptide (GIP), which reduces insulin secretion, increases glucagon secretion and consequently, enlarges the liver glucose release (Nauck et al. 2004).

Finally, the increased capacity of the **kidney** to reabsorb glucose in patients with T2D worsens the hyperglycaemia by increasing glucose circulatory levels (DeFronzo et al. 2013). Also, it has been suggested that for T2D patients, renal gluconeogenesis is exacerbated (Meyer et al. 1998).

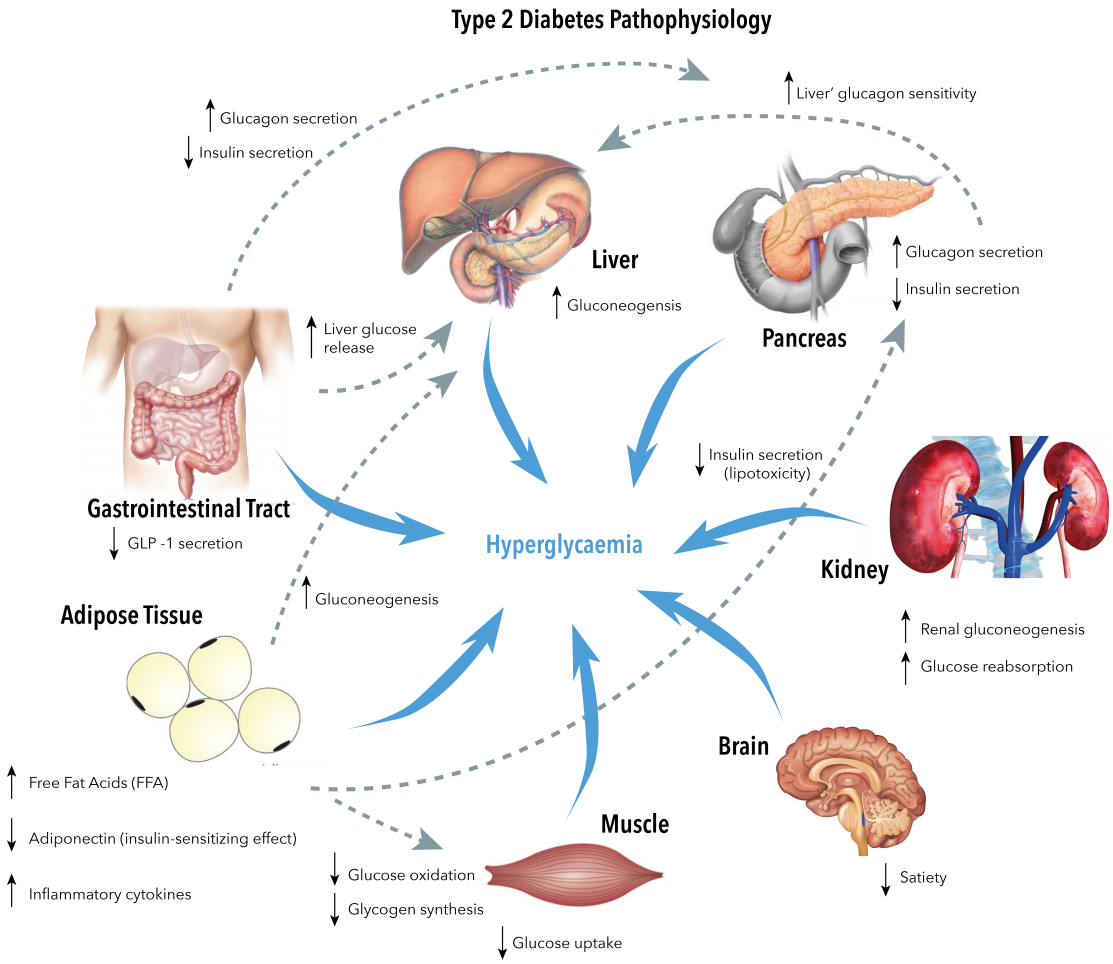


Figure 9. The pathophysiology of T2D across multiple tissues. Adapted from (Defronzo 2009)

The pace of disease progression as well as the pattern of long-term complications is heterogeneous across patients, which implies more individual tailoring in order to meet the particular risk factors encountered in each case (Tallapragada et al. 2015). Additionally, several studies directed to understand the progress in the T2D pathophysiology indicated that T2D is more like a jigsaw of several metabolic disorders determined by distinctive intermediate traits (body mass index, waist-hip-ratio, hypertension…) (1998; Zoungas et al. 2009; Ismail-Beigi et al. 2010). However, most of the pharmacological treatments are directed to lower the hyperglycaemia, ignoring the risk to all the several complications related to T2D and there is no single therapy able to target all the effects and organs involved in the T2D pathogenesis (Srinivasan and Florez 2015).

8.2 Other forms of Diabetes Mellitus

The other major form of DM is **Type 1 diabetes (T1D)** that represents 5-10% of DM cases (Atkinson et al. 2014) and it is driven by autoimmune destruction of ß-cells manifested through the detection of autoantibodies for pancreatic ß-cells (Naylor et al. 2011; Onengut-Gumuscu et al. 2015). The pathophysiology of T1D is much more known, more accurate diagnosis and prognosis are available and the treatment is insulin administration. Besides these two polygenic conditions, there are also rare monogenic forms of DM characterized by highly penetrant genetic defects in single genes, which lead to pancreatic ß-cell dysfunction and hyperglycaemia (Naylor et al. 2011; Tallapragada et al. 2015). Monogenic forms of T2D account for 2-5% of diabetes cases and a large number of patients are incorrectly diagnosed for T1D or T2D, which can result in unnecessary use of insulin or the inefficient ascertainment of at-risk family members (Naylor et al. 2011). This results in ineffective treatments and makes impossible the identification of at-risk relatives (Naylor et al. 2011; Tallapragada et al. 2015). The **maturity-onset diabetes of the young (MODY)** is characterized by non-ketotic diabetes with an early onset set between 6 months and 35 years of age (Naylor et al. 2011; Flannick et al. 2016). MODY does not imply the intervention of autoimmune antibodies or insulin resistance; hyperglycaemia is due to a decrease in ß-cell mass. Nevertheless, 14 MODY subtypes are described and their responsible genes identified, which represent 80-90% of the diagnosed cases. Another rare DM forms are **Neonatal Diabetes Mellitus (NDM)**, which manifest during the first few weeks of the newborn (diagnosed before 6 months). Increased blood sugar levels in these patients can be either transient or permanent NDM (TNDM and PNDM) as well as syndromic cases of NDM. NDM is characterized by low birth weight and ß-cell dysfunction (Naylor et al. 2011; Tallapragada et al. 2015; Flannick et al. 2016). TNDM recedes at 18 weeks of age but these patients are in high risk of developing diabetes as adults while PNDM can involve either isolated hyperglycaemia or other extra-pancreatic defects, requiring life-long treatment (Tallapragada et al. 2015). Finally, other rare forms include mitochondrial diabetes mellitus, multiorgan syndromes such as Wolcott-Rallison and Wolfram syndromes. Some of these monogenic forms have been successfully addressed due to the strong correlation between genetics and disease manifestation, which led to better diagnosis and effective treatments. As previously said, for 90% of DM cases that

corresponds to T2D, treatments are inadequate and unable to cope with late-stage complications (Flannick et al. 2016).

8.3 The progress in the understanding of the genetic architecture of type 2 diabetes

On the cutting-edge of the study of the genetics of complex multifactorial polygenic diseases, type 2 diabetes (T2D), the **most prevalent form of DM**, always have had an outstanding place (Billings and Florez 2010). In contrast to monogenic forms of diabetes mellitus showing clear inheritance patterns, T2D results from the interplay between different genetic factors and environmental factors. The estimated heritability of T2D ranges from 30-70% depending on diagnostic criteria but also on age (Poulsen et al. 1999; Almgren et al. 2011; Willemsen et al. 2015). Therefore, **genetics is a key player** in the T2D pathophysiology with a more critical role in the first stages of development of the disease according to twin and family studies (Poulsen et al. 1999; Almgren et al. 2011). The genetic architecture of T2D has been thought to follow the CDCV model, and therefore, mostly based on the contribution of common variants showing modest and small effect sizes. Although more than 100 robust T2D associated variants have been linked to disease susceptibility (Fuchsberger et al. 2016), **less than 10% of the T2D heritability** can been explained by the known associated variants (Manolio et al. 2009; Billings and Florez 2010; Hofker et al. 2014).

Next, I will review the successive findings in the genetics of T2D that this last decade has witnessed which are represented in Figure 10.

8.3.1 Common Variants

Genome-Wide Association Studies have been the gold standard for the identification of the majority of known genetic factors contributing to T2D risk (mostly falling at the range of common variants), and have demonstrated the polygenic nature of T2D (Tallapragada et al. 2015). This success explained the poor performance of familiar linkage analysis and candidate gene studies (Bonnefond and Froguel 2015) performed previously. The pre-GWAS era led to the identification of *PPARG* (Altshuler et al. 2000) and *KCNJ11-ABCC8* (Gloyn et al. 2003) by candidate gene studies and *TCF7L2* through linkage analysis (Grant et al. 2006). After that, there was a succession of several waves of GWAS. At the beginning, studies of a few thousand individuals resulted in dozen novel *loci* (Scott et al. 2007; Sladek et al. 2007; Steinthorsdottir et al. 2007; Wellcome Trust Case Control 2007; Zeggini et al. 2007). These evidences validated the GWAS approach and suggested that common variants would not show large effect sizes. Therefore, in order to **identify common variants of weaker effects**, data-sharing and collaborative initiatives were translated into large meta-analysis. The Diabetes Genetics Replication and Meta-analysis (DIAGRAM) consortium answered this demand and it was able to assemble 10,000 individuals, including 4,500 T2D cases (Zeggini et al. 2008). This kind of initiatives has been constant in the following years.

This success encouraged performing GWAS and meta-analysis **across several populations**, and there was also a new wave of GWAS based on **T2D-related metabolic quantitative traits** in non-diabetic individuals. This approach led to the identification of novel T2D *loci* in Europeans such as *MTRN1B* (Bouatia-Naji et al. 2009; Lyssenko et al. 2009) or *PROX1, ADCY5, DGKB* (Dupuis et al. 2010) that were primarily associated with fasting glucose (FG). However, the reduced overlap between genes associated with T2D and those influencing normal glycaemic traits suggested that the knowledge of the physiology of metabolic traits would not be translated in T2D risk. A recent study has reported how the combination of known FG SNPs can predict the incidence of impaired fasting glycaemia over a 9-year follow-up while they were able to predict T2D incidence (Vaxillaire et al. 2014). Moreover, most of the T2D *loci* reported were mainly involved in the ß-cell function, with a smaller contribution from insulin resistance-related genes.

A new round of GWAS meta-analysis using a **custom genotyping array** called Metabochip, built on the basis of a backbone of nominally associated variants with 23 cardio-metabolic traits and associated diseases as well as to fine-map well established disease and trait associated *loci* (Voight et al. 2012). This inexpensive array enabled researches to genotype a much larger number of samples to boost statistical power for *loci* discovery. The DIAGRAM consortium enlarged the sample size of previous meta-analysis by genotyping through the Metabochip ~150K individuals (with ~38K cases) that led to the identification of ten further *loci* (Morris et al. 2012).

The first **non-European-based GWAS** for T2D was performed in 2008 and led to the identification of the *KCNQ1* gene based on variants common in East Asians (MAF ~3 %) (Unoki et al. 2008; Yasuda et al. 2008). In 2010, a further large meta-analysis through individuals of European ancestry was unable to replicate those signals found in East Asian populations but they provided an independent signal for the same gene (Voight et al. 2010). Since 2012, many studies addressed T2D risk in many different ethnicities, which resulted in novel T2D *loci* driven by risk alleles that showed differentiated allele frequencies across populations. For instance, the *SLC16A11-SLC16A13* locus was identified in both Japanese (Hara et al. 2014) and Latino populations (Sigma Type 2 Diabetes Consortium et al. 2014c) or studies in South Asians led to the identification of *TMEM163* (Tabassum et al. 2013) and *SGCG* (Saxena et al. 2013). Another example is a study in the Greenlandic population based on using the Metabochip array that found a novel association in the *TBC1D4* gene with an OR=10.3 under a recessive inheritance model (Moltke et al. 2014). These studies underscored the statistical power from studies based on diverse, founder and historically isolated populations for the identification of novel risk *loci*.

In order to strengthen the identification of T2D *loci* whose risk alleles are shared across populations, **meta-analysis across several ancestries** were performed. For instance, a meta-analysis based on European, African-American, Hispanic-Latino and Asian studies allowed the discovery of the *BCL2 locus* (Saxena et al. 2012). In addition, a trans-ethnic meta-analysis assembling more than 110K

individuals, including 26,500 cases, led to the identification of seven new *loci* (DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium et al. 2014). These meta-analyses have extensively benefited from the emergence of genotype imputation methods, especially in this situation that required the array homogenization of cohorts from different ancestry.

Another approach that led to the discovery of novel T2D *loci* was taking into account the relationship of T2D risk with body mass index (BMI), which was translated in evaluating T2D in lean compared with obese individuals, resulting in the identification of the *LAMA1* gene (Perry et al. 2012).

8.3.2 Low-frequency and Rare Variants

The advent and the subsequent drop of NGS technologies allowed seeking for novel discoveries driven by **low-frequency and rare variants** in well-established T2D genes but also they provided new T2D *loci.* These studies were based on using **WES, WGS,** dense **genotyping arrays** based on **coding variation**, genotyping arrays with **genotype imputation** based on **sequence-based reference panels** and **gene region-based tests** (Mohlke and Boehnke 2015).

With respect to novel T2D *loci* driven by low-frequency variants, the combination of WGS and genotype imputation led to the identification of a rare missense variant (rs78408340, MAF ~ 0.006) in the *PAM* gene (in conjunction with an independent common signal at the same gene), an intronic low-frequency variant at the *CCND2* gene showing a substantial protective effect-size (rs76895963, OR ~0.5) and a rare frameshift in the MODY (chr13:g.27396636delT, MAF ~ 0.002) *PDX1* gene (Steinthorsdottir et al. 2014).

Regarding the contribution of low-frequency and rare variants in well-established T2D *loci*, a low-frequency missense variant in the *HNF1A* gene was identified through WES in Mexican and Latino individuals, showing a five-fold increased risk of T2D (Sigma Type 2 Diabetes Consortium et al. 2014a). **Targeted gene sequencing studies** and **gene-based tests** aggregating the contribution of multiple exonic variants were a more successful approach. For instance, 36 rare variants identified by exome sequencing in the *MTNR1B* gene were tested through cell assays, which allowed capturing very rare loss-of-function variants showing more than five-fold increased risk of T2D (MAF < 0.001, OR ~ 5.67) in the *MTNR1B* gene (Bonnefond et al. 2012). Furthermore, 49 non-synonymous variants were identified by *PPARG* sequencing but only nine rare variants that showed reduced activity in a novel adipocyte differentiation assay conferred a substantial increase in T2D risk (OR=7.22) (Majithia et al. 2014). In addition, a large targeted gene study that integrated *SLC30A8* variants identified by exome sequencing or genotyping data in 150K individuals across five ancestry groups resulted in 12 predicted protein-truncating rare variants with a 3-fold reduced risk of T2D (Flannick et al. 2014).

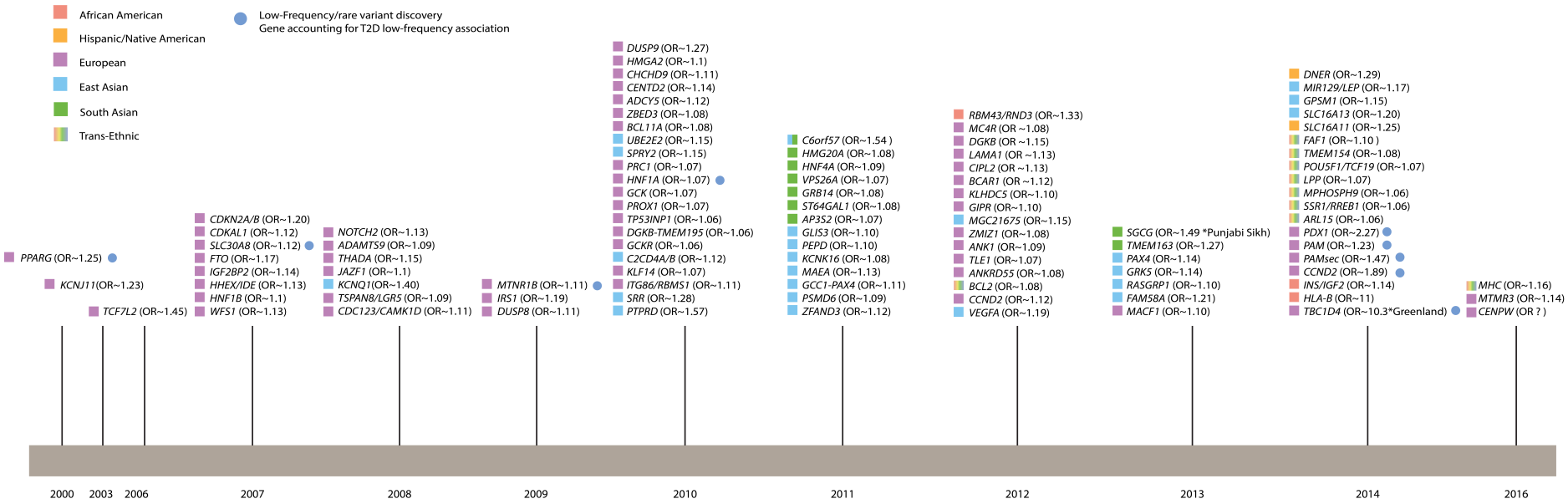


Figure 10. Progress in the understanding of the genetic basis of type 2 diabetes. The discovery of novel T2D associated genes is shown in a timeline representation. For each *locus,* we provided the OR of the association signal reported in the discovery. We highlighted with coloured squares the ethnicity of the discovery study as detailed in the legend. Moreover, blue circles were used to underscore those *loci* that have been discovered or independently replicated by low-frequency and rare variants.

These latter studies demonstrated a **broader allelic series for T2D risk** through the identification of low-frequency and rare variants in known T2D genes. Additionally, this hypothesis has been addressed in other studies. For instance, by sequencing seven MODY genes in ~4,000 individuals from general population, individuals diagnosed with MODY or exhibiting extreme T2D phenotypes were reported to show an excess of low-frequency predicted deleterious mutations (Flannick et al. 2013). Moreover, in the latest state of the art T2D large-scale genetic study, rare-variant associations were aggregated across 81 genes attributed to monogenic forms of DM to perform gene-set analysis. A particular gene-set accounting for variants with MAF < 0.01 from 13 genes underlying MODY or NDM and reported in the OMIM database, showed compelling enrichment for T2D associations (Fuchsberger et al. 2016).

These studies suggested that rare**,** low-frequency and common variants with modest and small effect sizes that contribute to the T2D risk coexists with also highly penetrant rare alleles for monogenic forms od DM. Within this **full spectrum of genetic variants**, common T2D and rare monogenic forms of DM are just two extremes in a disease continuum. On the basis of this rationale, addressing the risk of DM in a unified model can provide new insights for both common and rare forms of DM. In the context of T2D, association studies according to sub-phenotypes similar to rare monogenic DM disorders can improve the statistical power for the discovery of novel T2D *loci* (Flannick et al. 2016).

To round off this thorough review of the progress in the understanding of the genetic architecture of Type 2 diabetes I would like to summarize the last results from the ultimate T2D large-scale genetic study. This study exploited (1a) WGS in 2,657 individuals from the GoT2D data that was also used to (1b) impute sequence-based variants into GWAS data from 44,414 individuals, (2) WES from 12,940 individuals from the T2D-GENES consortium and (3) exome-array data from DIAGRAM (Fuchsberger et al. 2016). This study aimed to also shed light to this debate of the role of rare variants with large effects in T2D heritability against a model of trait variation mainly based on a vast number of common variants of weaker effects. Indeed, this controversial debate draws from the classical discussion between Mendelians and Biometricians. The authors reported that the HapMap coverage was able to capture all common-variant associations identified by WGS. Moreover, WGS and WES did not show a major involvement of low-frequency variants in T2D risk but they also claimed that sequencing studies were still limited in terms of sample size, limiting the statistical power to capture rare and low-frequency associated variants. Nonetheless, simulations along with empirical data suggested that the role of variants at the lower allele frequency ranges have a much-reduced role in T2D heritability compared to common variants (Fuchsberger et al. 2016).

In contrast to this latter study that was based in large consortia generating amounts of new, diverse and statistically powered data, this thesis opted for an alternative approach which consisted on the development and implementation of novel analytical and computational methodologies, in order to better exploit the available large-scale individual-level T2D GWAS data.

Hypothesis and Objectives

The main goal of this thesis was to achieve a more comprehensive understanding of the genetic and the molecular basis of Type 2 diabetes (T2D). This goal has been articulated by these hypotheses:

Hypothesis 1: A substantial fraction of missing heritability for Type 2 diabetes is still hidden in publicly available GWAS data that could be discovered by applying novel analytical methodologies.

Hypothesis 2: There is a large spectrum of unexplored genetic variation (low-frequency and rare variation) that contributes at some extent to the aetiology of Type 2 diabetes.

These hypotheses have been translated into three main objectives:

Objective 1: To standardize the use of genotype imputation with novel sequence-base reference panels for accurate ascertainment of low-frequency and rare variation through the development of specific protocols and guidelines.

Objective 2: To accelerate and to automatize large-scale genetic analysis through the implementation of computationally efficient pipelines.

Objective 3: To combine all the previous developments to re-analyse all the publicly available Type 2 diabetes GWAS data from European ancestry through genotype imputation with novel sequence-based reference panels and novel functional annotation maps.

Methods

The methods section has been split in two major sections. The **first block** corresponds to all the resources and methodologies followed to (1) computationally package a QC protocol for genotyped data and (2) the several approaches used to articulate some practical guidelines for genotype imputation with sequence-based reference panels for GWAS and meta-analysis approaches. The **second main block** contains all the resources, techniques and approaches undertaken for the re-analysis of all publicly available T2D GWAS datasets as well as all the methods followed for the downstream analysis.

1. Computational and analytical frameworks for imputation based GWAS

1.1 Automatizing and packaging GWAS analytical workflows

This work was conceived to **computationally automatize a quality control (QC) protocol for genotyping array data**. The different steps included have been extensively reported in previously studies (Anderson et al. 2010) and they are widely accepted by the community. Thus, common QC practices have been packaged in this automatized pipeline and I strongly recommend reading the work from (Anderson et al. 2010) for specific details of the QC. All the commands and even some scripts such as the one for filtering highly related individuals were obtained from that pipeline. As explained by the authors, most of the analysis described below rely on **PLINK (v1.9 version)** (Chang et al. 2015), that has been wrapped-up with **R-scripting** (also used to provide graphical reports) and **UNIX commands** for some filtering and data management tasks. This pipeline exploits **SLURM** (Simple Linux Utility for Resource Management) (Jette et al. 2002) and **LSF** (Load Sharing Facility) (Zhou et al. 1993) queueing systems for the execution of the different tasks in a cluster environment. I will briefly review some of the steps already detailed in the aforementioned previous work.

a) **Variant Based Filtering:** to check excess of **missing genotypes, deviance of HWE, statistical differences in missing data rates and lower allele frequencies**, we used PLINK. These flags allow retrieving the corresponding per-maker values for each of these parameters.

```
--missing
--hardy
--test-missing
--freq
```

These analyses are independent executions. Then, according to the cut-offs established by the user, the workflow filters defective markers failing any of these parameters.

b) Sample Based Filtering

In order to check **sex discordance and high missing data rates**, the following flags from the PLINK software are used in sequential executions:

```
--check-sex

--mind 0.02
```

For the identification of highly related individuals, an **identity by state (IBS) pair-wise comparison matrix** for all the individuals is generated. Afterwards, the degree of recent shared ancestry from each pair of target individuals is computed, the **identity by descent (IBD)** estimate that allows identifying highly related individuals. To lower the computational complexity of these calculations, only independent SNPs should be included but also because IBS calculations are LD sensitive. This process is achieved by doing a **"pruning" of the data** so that any pair of variants has an r^2 value to a given threshold. To successfully do that, strong LD regions such as the MHC regions have been discarded. The fraction of independent set of variants is extracted from the study samples.

```
### PRUNNING calculation

plink –-noweb –-bfile file –-exclude range high_LD_regions.txt –-indep 50 5 0.2
--out output_list_prunned_snps
```

This fraction of independent variants is **merged with a population reference data** because the same IBD information will facilitate clustering the study population with the different ancestries enclosed in the reference data (i.e. HapMap data). To merge the target and the reference datasets, those variants with pairs of alleles A/T or C/G should be discarded. Then, we identified a set of common variants in both datasets allowing merging the datasets (performed in UNIX).

```
plink –-noweb –-bfile file –-bmerge reference.bed reference.bim reference.fam
–-extract common_variants_no_AT_CG –out output_merged
```

Because of strand discrepancies, this step can fail and it will end the execution despite prior removal of A/T or C/G SNPs. As a master R-script controls the whole QC process, until this previous execution, this part of the pipeline is run as a single job in a SLURM or LSF queueing system. The master R-script launches a second job in order to check the success of this step. If this step fails, problematic variants are flipped. If the error is persistent, these variants are removed by launching a third job in the computing cluster that must ensure merging both sets of data.

To generate the pair-wise IBS/IBD matrix, we used the **parallel computing capabilities of the PLINK software** (specially improved in the 1.9 release) that allows to split this job in several independent tasks. The decision to divide in more or less parts is done by the master R-script in terms of the sample size. For instance, the first out of four tasks in which this job has been split is executed by:

```
plink --noweb --bfile file --genome --parallel 1 4 --out genome_list_chunk_1
```

Once this step has been achieved for each sub task, all the outputs are concatenated in a single file. Through a script obtained from the published QC protocol of Anderson, C.A. and colleagues (Anderson et al. 2010), highly related individuals are filtered. From a pair of individuals **exhibiting an IBD proportion > 0.125 (third-degree relatives)**, the individual showing the lowest call rate (previously computed when assessing the excess of missing data rate) is discarded.

To evaluate the population structure underlying study samples, we used multidimensional-scale analysis with PLINK, which has been coded to **generate 7 new components (PC)**. The IBD matrix previously calculated is required in the clustering analysis. After that, an R-script **discards individuals showing more that 4 standard deviations within the target samples according to the first four new components**. A plot based on four PCs represents reference populations and the target cohort to facilitate a manual inspection from the user of this analysis.

```
plink --noweb --bfile file --read-genome --cluster --mds-plot 7
--out novel_components_from_MDS
```

A last round of variant based filtering is performed before performing association tests using a Fisher test model and logistic regression, which is adjusted by the new components already calculated.

```
#BASIC ASSOC

plink --noweb --bfile file --assoc --out basic_association_test

#LOGISTIC REGRESSION

plink --noweb --bfile file --logistic --covar new_PCs –covar-name
PC1,PC2,PC3,PC4,PC5,PC6,PC7 –ci 0.95 --hide-covar --out logistic_association_test
```

The pipeline **(1) summarizes all the statistics calculated per marker and per individual**, (2) seeks for the different variants and subjects excluded, (3) **lists top-hits variants and regions**, (4) provides **manhattan-plots and Q-Q-plots** for each association test and (5) **Q-Q-plots for the HWE p-values**.

The QC ends with a report of the whole execution summarizing the different decisions undertaken with respect to the exclusion of variants and samples in each step.

This workflow encompasses a set of R-packages (*gap*, *getopt*, *optparse* and *IRanges)* that are installed according to the resources of the cluster infrastructure if necessary. A master R-script looks for the full disposal of all the required R-libraries, initializes the main execution and manages the different jobs of tasks executed using different levels of parallelism and also following the appropriate queueing system.

1.2 Fostering guidelines for accurate genotype imputation of common and low-frequency variants for GWAS and sequence-based reference panels

1.2.1 Experimental data and pre imputation quality filtering

The data considered for this section was obtained from the Wellcome Trust Case Control Consortium (WTCCC) (2007) data through the European Genotype Archive (EGA, https://www.ebi.ac.uk/ega/studies/EGAS00000000028) (Barrett et al. 2009c). The genotyping data and the subjects included in the following tests were filtered according to the guidelines provided by the WTCCC, whose criteria of exclusion are in line with standard quality filters for GWAS (Anderson et al. 2010). We used **two cohorts of population controls from the WTCCC2 stage,** the 1958 British Birth cohort (~3,000 samples, 58C) and the National Blood Samples from the UK Blood Service Control Group (~3,000 samples, NBS). Both sets were genotyped by Affymetrix v6.0 and Illumina 1.2M chips. After applying the quality filtering criteria, 2,706 and 2,699 subjects from the Affymetrix and Illumina data, respectively, were available for the 58C samples, leaving an intersection of 2,509 individuals genotyped by both platforms. For the NBS samples, the final effective subjects were 2,674 and 2,501 for the Affymetrix and Illumina datasets, respectively, but only 2,163 were common in both chips. After variant quality filtering and excluding all the variants with minor allele frequency (MAF) below 0.01, 717,556 and 892,516 variants were remained for 58C Affymetrix and Illumina platforms, respectively while for the NBS samples, 714,771 and 893,369 markers were included for Affymetrix and Illumina, respectively.

1.2.2 Genotype imputation with IMPUTE2

We used the two-step genotype imputation approach based on **SHAPEIT2** (Delaneau et al. 2012) to pre-phase the study genotypes into full haplotypes in order to ameliorate the computational burden, and **IMPUTE2**[2] (Howie et al. 2009), which was currently reviewed as one of the best methods for genotype imputation (Marchini and Howie 2010). We excluded all C/G and A/T SNPs from the genotype imputation step to avoid strand orientation issues between the reference panel and the genotyped data. Phasing and Imputation were executed using the following commands:

```
shapeit --input-gen input.gen input.sample --input-map
genetic_map_chr_1_combined_b37.txt.gz --output-max out.haps out.sample
--thread 16 --effective-size 20000 --output-log out.log
```

[2] For "Results" sections, *Preventing the occurrence of spurious association from errors in genotyping* and *Exploring the impact of genotype imputation in meta-analysis approaches,* IMPUTE2 2.2.2 instead of IMPUTE2 2.3.0 release and SHAPEIT1 instead of the SHAPEIT2 release, were used.

```
impute2 -use_prephased_g -m genetic_map_chr_1_combined_b37.txt.gz
-h ALL.chr1.integrated_phase1_v3.20101123.snps_indels_svs.genotypes.nosing.haplotypes.gz
-l ALL.chr1.integrated_phase1_v3.20101123.snps_indels_svs.genotypes.nosing.legend.gz
-known_haps_g output.haps -int 244000001 245000000 -exclude_snps_g list_snps_AT_CG
-impute_excluded -Ne 20000 -o output_impute -o_gz
```

We used the GTOOL software (http://www.well.ox.ac.uk/~cfreeman/software/gwas/gtool.html, version 0.7.5) to homogenize strand annotation by merging the imputed results obtained from each set of genotyped data. To test for association under an additive model we used the following command-line and the SNPTEST tool (we did not adjust for any covariate in this comparison of imputed data across genotyping arrays from the same healthy individuals)

```
snptest_v2.5 -data merged_controls_cases.gen merged_controls_cases.sample
-frequentist 1 -method em -pheno bin1 -hwe –o controls_cases_snptest.out
–log controls_cases_snptest.log
```

All these steps and many more were implemented in the **GUIDANCE software** (Sanchez et al. 2016), an application that works on top of the COMP Superscalar (COMPSs) framework (Tejedor et al. 2012; Lordan et al. 2014), to phase, impute genotypes and perform association testing in high-performance computing environments (Tejedor et al. 2012).

We used three sequence-based reference panels to perform genotype imputation: 1000G-Phase1[3] (March 2012 and June 2014), the 1000G-Phase3 and the UK10K (UK10K Consortium et al. 2015) (https://ega-archive.org/studies/EGAS00001000713) reference panels.

1.2.3 Fixing appropriate quality thresholds across genotyping platforms

We evaluated genotype imputation for each reference panel considering 2,509 58C individuals that were genotyped by both independent genotyping platforms.

Four scenarios were considered: (a) fraction of variants originally genotyped (GT) by both Illumina (IL) and Affymetrix (Affy) platforms **(GT both)**; (b) Variants genotyped by Affy, but not present in IL array **(Affy GT)**; (c) Variants genotyped by Illumina, but not present in the Affy array **(IL GT)**; and (d) Variants not typed in IL nor in the Affy arrays, and therefore, **imputed from IL and Affy datasets** (d). This last scenario corresponded to the largest fraction of variants.

[3] For the "Results" sections, *Preventing the occurrence of spurious association from errors in genotyping* and *Exploring the impact of genotype imputation in meta-analysis approaches*, we used the 1000G-Phase1 release from March 2012.

As the individuals typed (and imputed) using Affy and IL SNPs as backbones were the same, **we expected no statistical differences when comparing the allele and genotype frequencies in any of the variants**. We addressed the accuracy of genotype imputation as the percentage of false positive (FP) associations, defined as variants showing genome-wide significant differences (p-value $\leq 5 \times 10^{-8}$) when comparing the two datasets derived from IL and Affy datasets.

The quality of the imputed variants was also evaluated using the **allelic dosage R^2 coefficient** between the genotype dosages estimated when imputing using Affy or IL as backbone. The Affy GT and IL GT SNPs (SNPs genotyped by one platform, but imputed on the other dataset) were also used to evaluate the **correspondence between the allelic dosage R^2 scores and the IMPUTE2-*info* scores** associated to the imputed genotypes. The linear model, between the allelic dosage R^2 and the IMPUTE2-*info*, was used to set an *info* score threshold of 0.7, which reached an allelic dosage R^2 of 0.5. This cut-off was uniform across all reference panels and platforms.

1.2.4 Combining imputed variants from each reference panel

For the 58C-imputed results from Affymetrix and Illumina, we chose the genotypes from the reference panel that **showed higher accuracy, estimated from the IMPUTE2-*info* score**. For this final combined set of variants we also assessed the differences in accuracy and coverage. We also filtered out those variants showing MAF < 0.001 and HWE p-value $\leq 1 \times 10^{-8}$.

1.2.5 Preventing the occurrence of spurious association from errors in genotyping

This strategy was based on the ***–pgs* option from IMPUTE2**, which *imputes* genotyped variants even though they have already been genotyped, based on the surrounding SNPs. We then compared **the imputed results against the real genotyped results**. Several metrics were evaluated in order to disclose which of them was the one that had the greatest discriminator power (See Figure 11). The best metric was:

diffBeta = abs (1 - beta);

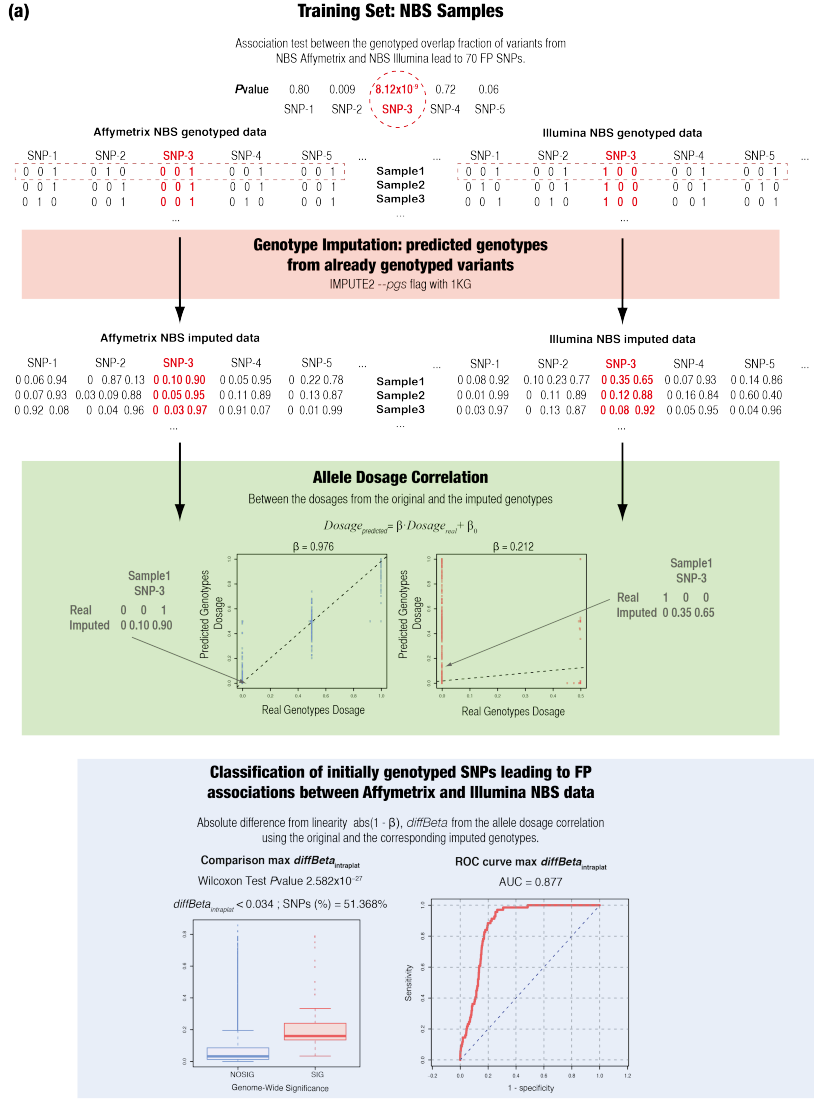
beta is the slope of the allele dosage correlation between the real and the *imputed* genotypes (termed ***diffBeta***). Note that we expect a perfect correlation and a slope of 1 when there is complete agreement between the *imputed* and the real results.

Using the 58C cohort, we eliminated all the SNPs with *diffBeta* higher than 0.10, as this showed to eliminate 87.1% of biased SNPs, as trained in the NBS cohort. After eliminating these SNPs, we phased the genotypes again, and we imputed the genotypes again. We then evaluated the coverage and the percentage of FP using different *info* score thresholds. We then look for the IMPUTE2-*info* filter allowing to at least reducing in a 90% the ratio of FP of the original pipeline (applying only a IMPUTE2-*info* cut-off of 0.7). In order to test this strategy in a dataset relying on a different array, we

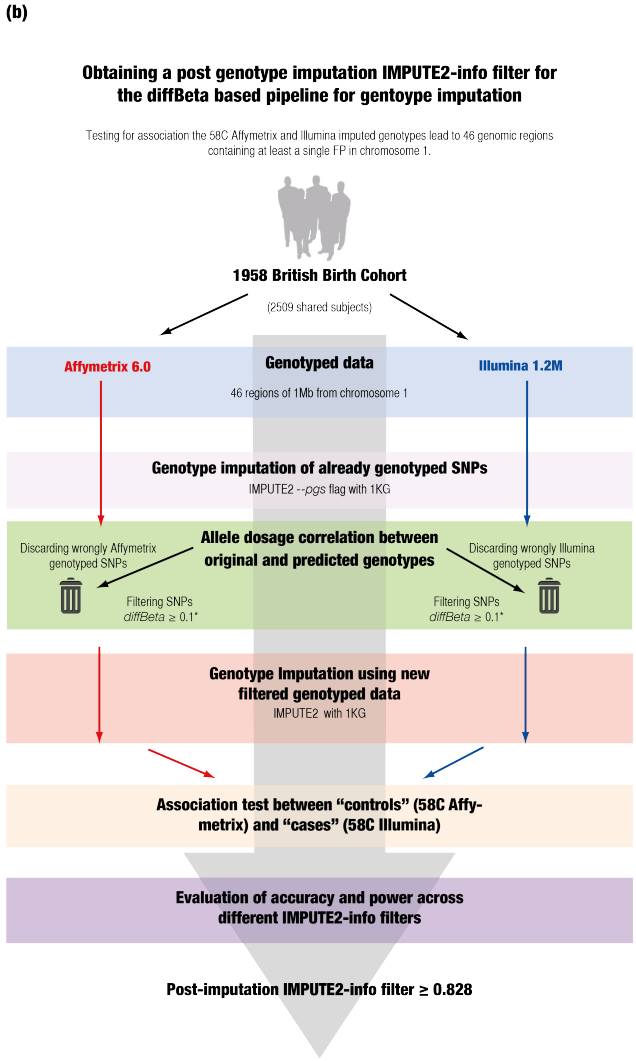
compared a fraction of the 58C samples that were genotyped by both the Affymetrix 500K arrays and Illumina 1.2M arrays.

1.2.6 Exploring the impact of genotype imputation in meta-analysis approaches

We **compared the allele frequencies of genotyped and imputed variants** between the **58C and NBS control samples using Affy and IL** as the starting genotyping platform. We compared NBS (Affy) vs 58C (Affy) and (b) NBS (IL) against 58C (IL) datasets (see Figure 12). The steps of genotype imputation and testing for association were performed for both sets as previously described. We then meta-analysed the association results using METAL (Willer et al. 2010) with the sample-size and the inverse variance fixed effect methods, separately, and we analysed the degree of heterogeneity using the I^2 score.



diffBeta ≥ 0.1 discards 87.1% of badly genotyped SNPs



*Parameter trained using the NBS cohort from the WTCCC stage2
(a) leaving 12.9% of badly genotyped variants

Figure 11. Filtering inaccurately genotyped variants before genotype imputation**.** (a) Training algorithm for the identification of wrongly genotyped variants in genomic regions containing at least one FP association (comparison between NBS Affy and IL results). In red, FP associations arising from biased genotyping between platforms. We used the IMPUTE2 –pgs option. The metric with best performance from comparing real against imputed genotypes was the *diffBeta*. A box-plot shows *diffBeta* values for Affy and IL SNPs from well genotyped (blue) and wrongly genotyped SNPs. The Receiving Operating Characteristic (ROC) curve shows the discriminatory power of the *diffBeta*. We set a threshold of 0.1 (87% of FP discarded). (b) Validation of the *diffBeta*-filtering pipeline in 58C imputed results. We used 46 genomic regions from chromosome 1 containing at least one FP association (imputed or genotyped SNPs). We applied the same procedure based on the QC with the *diffBeta* metric.

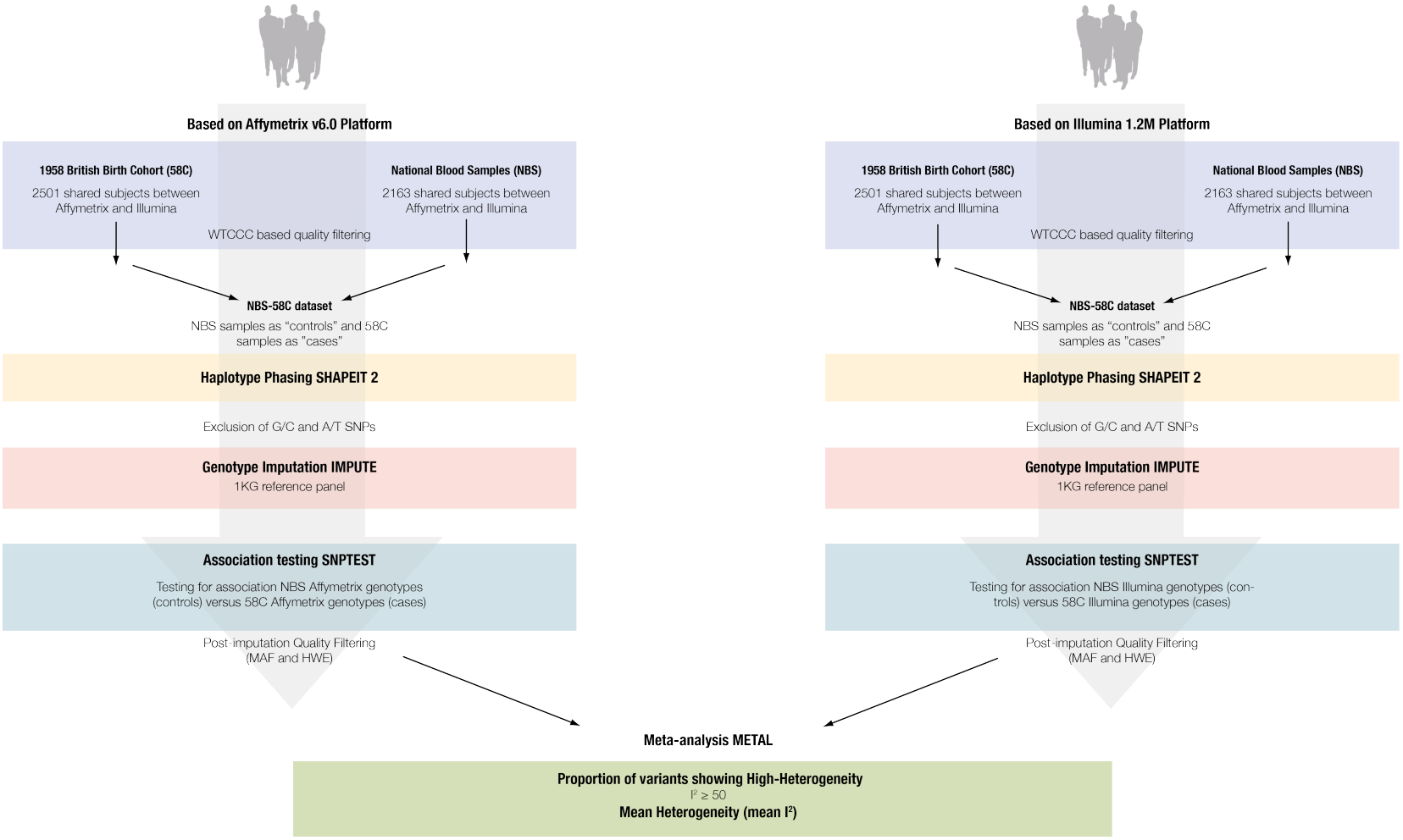

Based on Affymetrix v6.0 Platform
1958 British Birth Cohort (58C)
2501 shared subjects between Affymetrix and Illumina
National Blood Samples (NBS)
2163 shared subjects between Affymetrix and Illumina
WTCCC based quality filtering
NBS-58C dataset
NBS samples as "controls" and 58C samples as "cases"
Haplotype Phasing SHAPEIT 2
Exclusion of G/C and A/T SNPs
Genotype Imputation IMPUTE
1KG reference panel
Association testing SNPTEST
Testing for association NBS Affymetrix genotypes (controls) versus 58C Affymetrix genotypes (cases)
Post-imputation Quality Filtering (MAF and HWE)
Based on Illumina 1.2M Platform
1958 British Birth Cohort (58C)
2501 shared subjects between Affymetrix and Illumina
National Blood Samples (NBS)
2163 shared subjects between Affymetrix and Illumina
WTCCC based quality filtering
NBS-58C dataset
NBS samples as "controls" and 58C samples as "cases"
Haplotype Phasing SHAPEIT 2
Exclusion of G/C and A/T SNPs
Genotype Imputation IMPUTE
1KG reference panel
Association testing SNPTEST
Testing for association NBS Illumina genotypes (controls) versus 58C Illumina genotypes (cases)
Post-imputation Quality Filtering (MAF and HWE)
Meta-analysis METAL
Proportion of variants showing High-Heterogeneity
I² ≥ 50
Mean Heterogeneity (mean I²)

Figure 12. Strategy to evaluate the effect of post-imputation quality filters on the power of GWA-meta-analysis. We obtained two independent GWAS results by comparing NBS (controls) against 58C (cases) samples (NBSvs58C) that have been genotyped by both Affy and IL array platforms. We followed the two-step protocol for genotype imputation: pre-phasing genotypes with SHAPEIT before imputing nearby variants. We tested for association NBS and 58C groups for the IL and Affy datasets, independently. We performed meta-analysis between both studies (NBSvs58C Affy and NBSvs58C IL) and I^2 heterogeneity was compared across IMPUTE2-*info* filters.

2. Novel insights of the genetic architecture of T2D: crossing the boundaries of common variants

2.1 Details of independent discovery GWAS datasets

We collected all publicly **genetic individual-level data for Type 2 diabetes (T2D) case/control studies from 5 independent datasets available** in the dbGaP (http://www.ncbi.nlm.nih.gov/gap) and EGA (https://www.ebi.ac.uk/ega/home) public repositories, comprising a total of 13,201 cases and 59,656 controls. Each dataset was independently harmonized and quality controlled before performing genotype imputation and association testing.

2.1.1 Cohort: NuGENE NORTHWESTERN

dbGaP Study Accession: phs000237.v1.p1
Ethnicity: European (USA)
T2D cases after QC: 527
Controls after QC: 601

Type 2 diabetes case selection criteria:

Neither group should have T1D diagnosis codes (ICD-9 250.x1 or 250.x3)

1) Identification of patients who already have a T2D diagnosis:

    a) Include patients with Type 2 Diabetes diagnosis based on the list of codes for the International Statistical Classification of Diseases and Related Health Problems, the ICD9 code (excluding those with ketoacidosis codes).

    b) Exclude patients (currently) treated only with insulin AND have never been on a type 2 diabetes medication, and: diagnosed with T1D, or even if not diagnosed with T1D, diagnosed with T2D on < 2 dates in an encounter or problem list.

2) Identification of patients who do not yet have a T2D diagnosis: Include patients with HbA1c lab value ≥ 6.5%, FG > 125 mg/dl or random glucose > 200 mg/dl AND prescribed one of the medications (or combinations thereof) sulfonylureas, meglitinides, biguanides, thiazoldinediones, alpha-glycosidase inhibitors, DPPIV inhibitor and injectable.

Control selection criteria:

    a) Have had at least 2 clinic visits (face-to-face outpatient clinic encounters).

    b) Have not been assigned an ICD9 code for diabetes (type 1 or type 2) or any diabetes-related condition.

c) Have not been prescribed insulin or Pramlintide, or any medications for diabetes treatment, or diabetic supplies such as those for medication administration or glucose monitoring.

d) Do not have a reported (random or fasting) blood glucose ≥ 110 mg/dl and have had at least 1 glucose measurement.

e) Do not have a reported HbA1c ≥ 6.0%.

f) Do not have a reported family history of diabetes (type 1 or type 2).

2.1.2 Cohort: FUSION

dbGaP Study Accession: phs000100.v4.p1
Ethnicity: European (Finland)
T2D cases after QC: 901
Controls after QC: 772

Type 2 diabetes case selection criteria:

a) 644 FUSION and 275 Finrisk 2002 T2D cases as defined by WHO 1999 criteria of FPG ≥ 7.0 mmol/l or 2-hr plasma glucose ≥ 11.1 mmol/l, by report of diabetes medication use, or based on medical record review.

b) FUSION cases with known or probable T1D among their first-degree relatives were excluded.

c) The 644 FUSION cases each reported at least one T2D sibling.

d) The Finrisk cases came from a Finnish population-based risk factor survey.

Control selection criteria:

a) 331 FUSION and 456 Finrisk 2002 normal glucose tolerance (NGT) controls as defined by WHO 1999 criteria of fasting glucose < 6.1 mmol/l and 2-hr glucose < 7.8 mmol/l.

b) FUSION controls include 119 subjects from Vantaa, Finland, who were NGT at ages 65 and 70 years, and 212 NGT spouses of FUSION subjects. The controls were approximately frequency matched to the cases by age, sex, and birth province.

2.1.3 Cohort: GENEVA Genes and Environment Initiatives in Type 2 Diabetes (Nurses' Health Study/Health Professionals Follow-up Study) GENEVA NHS/HPFS

dbGaP Study Accession: phs000091.v2.p1
Ethnicity: European (USA)
T2D cases after QC: 2614
Controls after QC: 3061

Type 2 diabetes case selection criteria:

Through 1996 follow-up, criteria for confirmed T2D included one of the following:

a) One or more classic symptoms (excessive thirst, polyuria, weight loss, hunger, pruritus, or coma) plus FPG ≥ 140 mg/dl (7.8 mmol/L) and/or random plasma glucose ≥ 200 mg/dl (11.1 mmol/L) and/or plasma glucose 2 hr after an OGTT ≥ 200 mg/dl; or

b) At least two elevated plasma glucose levels (as described above) on different occasions in the absence of symptoms; or

c) Treatment with hypoglycaemic medication (insulin or oral hypoglycaemic agent).

In response to the current American Diabetes Association (ADA) diagnostic criteria (FPG cut point ≥ 126 mg/dl [7.0 mmol/L]), we revised the Supplementary Diabetes Questionnaire for participants reporting a new diagnosis of diabetes on the 1998 or later questionnaires. This revised supplementary questionnaire ascertains the level of elevation in fasting plasma glucose and will enable us to determine which participants had fasting plasma glucose ≥ 140 mg/dl (the earlier diagnostic cut point) and which had a fasting plasma glucose ≥ 126 (the current diagnostic cut point). The criteria for confirmed T2D during the 1998–2000 follow-up cycle and later cycles remain the same, except for the elevated FPG criterion for which the cut point was changed from 140 mg/dl to 126 mg/dl. The revised supplementary questionnaire enables us to classify cases in categories of glucose elevation and determine the proportion diagnosed in each category (e.g. FPG 126–139 versus ≥ 140 mg/dl) allowing us to conduct sensitivity analyses with exclusion of participants that meet the ADA criteria and not the National Diabetes Data Group (NDDG) criteria.

Control selection criteria: No diabetes mellitus.

2.1.4 Cohort: Wellcome Trust Case Control Consortium (WTCCC)

EGA Study ID: EGAS00000000005 (EGAS00000000001 + EGAS00000000002 + EGAS00000000009)
Ethnicity: European (UK)
T2D cases after QC: 1894
Controls after QC: 2917

T2D case selection criteria:

The T2D cases were selected from UK Caucasian subjects who form part of the Diabetes UK Warren 2 repository. In each case, the diagnosis of diabetes was based on either current prescribed treatment with sulphonylureas, biguanides, other oral agents and/or insulin or, in the case of individuals treated with diet alone, historical or contemporary laboratory evidence of hyperglycaemia

(as defined by the WHO). Other forms of diabetes (for example, MODY, mitochondrial diabetes, and T1D) were excluded by standard clinical criteria based on personal and family history. Criteria for excluding autoimmune diabetes included absence of first-degree relatives with T1D, an interval of ≥1 years between diagnosis and institution of regular insulin therapy and negative testing for antibodies to glutamic acid decarboxylase. Cases were limited to those who reported that all four grandparents had exclusively British and/or Irish origin, by both self-reported ethnicity and place of birth. All were diagnosed between age 25 and 75. Approximately 30% were explicitly recruited as part of multiplex sibships (Wiltshire et al. 2001) and ~25% were offspring in parent-offspring 'trios' or 'duos' (that is, families comprising only one parent complemented by additional sibs) (Frayling et al. 1999). The remainders were recruited as isolated cases but these cases were (compared to population-based cases) of relatively early onset and had a high proportion of T2D parents and/or siblings (Groves et al. 2006). Cases were ascertained across the UK but were centralized on the main collection centres (Exeter, London, Newcastle, Norwich,Oxford). Selection of the samples typed in WTCCC from the larger collections was based primarily on DNA availability and success in passing Diabetes and Inflammation Laboratory (DIL)/Wellcome Trust Sanger Institute DNA quality control.

Control selection criteria:

a) The 1958 Birth Cohort (also known as the National Child Development Study) includes all births in England, Wales and Scotland, during one week in 1958. From an original sample of over 17,000 births, survivors were followed up at ages 7, 11, 16, 23, 33 and 42 years (http://www.cls.ioe.ac.uk/studies.asp?section=000100020003). In a biomedical examination at 44-45 years (Strachan et al. 2007) (http://www.b58cgene.sgul.ac.uk/followup.php), 9,377 cohort members were visited at home providing 7,692 blood samples with consent for future Epstein-Barr virus-transformed cell lines. DNA samples extracted from 1,500 cell lines of self-reported white ethnicity and representative of gender and each geographical region were selected for use as controls.
b) The second set of common controls was made up of 1,500 individuals selected from a sample of blood donors recruited as part of the current project. WTCCC in collaboration with the UK Blood Services (NHSBT in England, SNBTS in Scotland and WBS in Wales) set up a UK national repository of de-identified samples of DNA and viable mononuclear cells from 3,622 consenting blood donors, age range 18-69 years (ethical approval 05/Q0106/74). A set of 1,564 samples was selected from the 3622 samples recruited based on sex and geographical region (to reproduce the distribution of the samples of the 1958 Birth Cohort) for use as common controls in the WTCCC study. DNA was extracted as described below with a yield of 3054 ± 1207 μg (mean ± 1 s.d.).

2.1.5 Cohort: Resource for Genetic Epidemiology Research on Adult Health and Aging (GERA)

dbGaP Study Accession: phs000674.v1.p1
Ethnicity: European (USA)
T2D cases after QC: 6995
Controls after QC: 49845

Inclusion criteria:

a) Eligible for the Research Program on Genes, Environment, and Health (RPGEH) survey

   i) ≥ 18 years of age at time of survey mailing (2007).

   ii) Kaiser Permanente Northern California Region enrollee for at least 2 years prior to survey.

b) Consented to contribute biospecimen to RPGEH and returned saliva sample by cut-off date for GERA genotyping.

c) All available samples from minorities were included, plus Non-Hispanic Whites selected at random to reach 110,266 participants with extracted DNA whose samples were submitted for genotyping.

d) Successfully genotyped (DQC ≥ 0.82; call rate ≥ 0.97) from extracted DNA.

e) Consented explicitly to have data deposited in NIH-maintained database.

Exclusion criteria:

1) Subject requested withdrawal from study after DNA extraction and genotyping.

2) Validity of link between biospecimen and study participant questionable because of genotype-phenotype discordance, e.g. gender.

A participant was coded as a patient for T2D if he/she had at least two diagnoses within this disease category that had to be recorded on separate days. Diagnoses were obtained from patient encounters at Kaiser Permanente Northern California facilities from January 1, 1995 to March 15, 2013. The March 2013 ICD9-CM diagnoses used for the Type 2 Diabetes category were:

a) 250.00 Diabetes mellitus without mention of complication, type II or unspecified type, not stated as uncontrolled.
b) 250.02 Diabetes mellitus without mention of complication, type II or unspecified type, uncontrolled.

c) 250.10 Diabetes with ketoacidosis, type II or unspecified type, not stated as uncontrolled.
d) 250.12 Diabetes with ketoacidosis, type II or unspecified type, uncontrolled.
e) 250.20 Diabetes with hyperosmolarity, type II or unspecified type, not stated as uncontrolled.
f) 250.22 Diabetes with hyperosmolarity, type II or unspecified type, uncontrolled.
g) 250.30 Diabetes with other coma, type II or unspecified type, not stated as uncontrolled.
h) 250.32 Diabetes with other coma, type II or unspecified type, uncontrolled.
i) 250.40 Diabetes with renal manifestations, type II or unspecified type, not stated as uncontrolled.
j) 250.42 Diabetes with renal manifestations, type II or unspecified type, uncontrolled.
k) 250.50 Diabetes with ophthalmic manifestations, type II or unspecified type, not stated as uncontrolled.
l) 250.52 Diabetes with ophthalmic manifestations, type II or unspecified type, uncontrolled.
m) 250.60 Diabetes with neurological manifestations, type II or unspecified type, not stated as uncontrolled.
n) 250.62 Diabetes with neurological manifestations, type II or unspecified type, uncontrolled.
o) 250.70 Diabetes with peripheral circulatory disorders, type II or unspecified type, not stated as uncontrolled.
p) 250.72 Diabetes with peripheral circulatory disorders, type II or unspecified type, uncontrolled.
q) 250.80 Diabetes with other specified manifestations, type II or unspecified type, not stated as uncontrolled.
r) 250.82 Diabetes with other specified manifestations, type II or unspecified type, uncontrolled.
s) 250.90 Diabetes with unspecified complication, type II or unspecified type, not stated as uncontrolled.
t) 250.92 Diabetes with unspecified complication, type II or unspecified type, uncontrolled.

The rest of subjects no coded as T2D patients were considered as controls.

2.1.6 DIAGRAM Trans-Ethnic meta-analysis.

We used the summary statistics for the **trans-ethnic T2D GWAS meta-analysis** (DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium et al. 2014) from the DIAGRAM consortium, which comprises the following ancestry-specific meta-analyses: the DIAGRAM Consortium (12,171 cases and 56,862 controls, European ancestry); the AGEN-T2D Consortium (6,952 cases and 11,865 controls, East Asian ancestry); the SAT2D Consortium (5,561 cases and 14,458 controls, South Asian ancestry); and the MAT2D Consortium (1,804 cases and 779 controls, Mexican and Mexican American ancestry). Each individual study undertook sample and SNP QC, and the genomic resolution was increased up to 2.5 million autosomal SNPs thanks to genotype imputation using reference panels from Phase II/III HapMap. QCed SNPs with MAF>1%, (except MAF>5% in

the Mexican and Mexican American ancestry GWAS due to smaller sample size) were tested for association with T2D under an additive model adjusted for several study specific covariates. Association summary statistics were combined via fixed-effects according to the ancestry group, and the results of each ancestry-specific meta-analysis were combined thanks to a fixed effects inverse-variance weighted meta-analysis, comprising a total sample size of 26,488 cases and 83,964 controls.

2.1.7 Type 2 Diabetes Knowledge Portal (T2D Portal)

The T2D Portal (http://www.type2diabetesgenetics.org/) is a central repository for obtaining summary statistics from large genetic association studies of T2D, including projects based on **WES and exome arrays for low-frequency data** and **SNP arrays covering common variation (GWAS)**. Besides this, T2D Portal has also included the results from GWAS meta-analysis of 24 other traits.

In our study we used the summary statistics from WES analysis and exome chip analysis. First, the summary statistics of 16,857 individual exome sequences were derived from the integration of multiple projects such as T2D-GENES, GoT2D and SIGMA. This dataset comprises individuals from 5 ethnic groups (African-American, East Asian, South Asian, European and Hispanic) (Sigma Type 2 Diabetes Consortium et al. 2014a; Fuchsberger et al. 2016). Additionally, we also used the summary statistics from exome chip analysis of 75,670 individuals from European ancestry. This dataset has integrated the efforts from the DIAGRAM consortium, the GoT2D project and the T2D-GENES project (Fuchsberger et al. 2016). This data was accessed on June 2016.

2.2 Summary of replication datasets

2.2.1 InterAct

The **InterAct consortium** (Langenberg et al. 2014) entails a case-cohort study that aroused from the existing large cohort 'EPIC' study. The EPIC study comprises 350,000 participants from 10 European countries and a lot of effort was put in standardizing lifestyle and dietary information. After a follow-up of 8 years, T2D has been diagnosed to 12,403 and InterAct has also defined a cohort of 16,154 controls free of diabetes at baseline. From the SNP and sample QCed data, we extracted the male samples corresponding to 6,763 individuals, which were re-analysed using genotype imputation with the UK10K reference panel. Association with T2D has been evaluated using an additive logistic model with SNPTEST v2.5.2 adjusted by age and body-mass index.

2.2.2 Slim Initiative in Genomic Medicine for the Americas (SIGMA) T2D Genetics Consortium

The **SIGMA consortium** GWAS dataset comprised of 8,214 individuals (3,848 T2D cases and 4,366 controls), consisting of four independent cohorts of Mexican or individuals with Latin American ancestry: The Diabetes in Mexico Study (DMS), Mexico City Diabetes Study (MCDS), Multiethnic Cohort (MEC) and UNAM/INCMNSZ Diabetes (UIDS) cohorts (Sigma Type 2 Diabetes Consortium et

al. 2014c). The genotyping of the study participants using the Illumina OMNI2.5 array was described previously (Sigma Type 2 Diabetes Consortium et al. 2014c). These cohorts, after the SNP and sample QC, were imputed using the UK10K reference panel and the association with T2D was tested under an additive logistic model only considering male samples with SNPTEST v2.5.2 adjusted by age and body-mass index.

2.2.3 Danish cohort

The **Danish** replication data consisted of **five sample sets**: 1) Inter99, a population-based randomized controlled trial (CT00289237, ClinicalTrials.gov) investigating the effects of lifestyle intervention on cardiovascular disease (Jorgensen et al. 2003); 2) Health2006 cohort, a population-based epidemiological study of general health, diabetes and cardiovascular disease (Thuesen et al. 2014); 3) ADDITION-DK screening cohort, 4) Vejle Biobank diabetes case-control study; and 5) clinical type 2 diabetes cases ascertained at Steno Diabetes Center.

All individuals were of Danish nationality. Written informed consent was obtained from all participants. The studies were approved by the local Scientific Ethics Committees and were performed in accordance with the principles of the Declaration of Helsinki II.

T2D was defined according to WHO 1999 criteria. Control individuals had FPG < 6.1 mmol/L (all study groups) and furthermore 2 hr plasma glucose during an oral glucose tolerance test < 7.8 mmol/l (study group 1). For the case-control analysis, we relied on **three definitions of controls**: a) Any subject with fasting plasma glucose <6.1 mmol/L; b) any subject with fasting plasma glucose < 6.1 and older 55 years (which corresponds do average age at onset); c) any subject older than 55 and oral glucose tolerant test below 7.8 mmol/l (study group 1).

The **Kaplan-Meier method** was used to **plot cumulative incidence of T2D** against **time of follow-up in the Inter99 cohorts**, which were followed for **11 years** on average. Cox proportional hazards regression models were used to address the risk of incident T2D. Individuals with self-reported diabetes at the baseline examination and individuals present in the Danish National Diabetes Registry before the baseline examination were excluded from the present analyses of incident T2D. The analyses for the follow-up study were restricted to male individuals younger than 45 years old, which will reach 55 years old after 11 years of follow-up.

2.3 *De-Novo* of Danish samples

Genotyping of Danish samples was performed by KASPar SNP Genotyping System (LGC Genomics, Hoddeson, UK). Ten selected samples from the 1000G Project (Coriell) were genotyped together with the study samples to estimate mismatch between genotyping and sequencing. All genotypes (5 heterozygous and 5 homozygous for reference allele) were concordant. Furthermore, 1,602 study

samples were genotyped in duplicate and no mismatches were observed. Moreover, general call rate was 98% and genotype distribution was in accordance with HWE.

2.4 Quality control for genotyped data

All genotyped datasets underwent the same 3-**step QC protocol** using PLINK (Purcell et al. 2007), including 2 stages of SNP removal and an intermediate stage of sample exclusion.

The exclusion criteria for genetic markers consisted on a) proportion of missingness ≥ 0.05, b) HWE p-value ≤ $1x10^{-6}$ for controls and HWE p-value ≤ $1x10^{-20}$ for all the cohort, c) differences in proportion of missingness between cases and controls p-value ≤ $1x10^{-6}$ and d) MAF < 0.01. Only for the GERA cohort we considered a MAF of 0.001 as exclusion criteria because of the large sample size of this dataset.

For sample quality control we considered the following exclusion criteria: a) gender discordance, b) subject relatedness (pairs with $\widehat{\pi}$ ≥ 0.125 from which we removed the individual with the highest proportion of missingness), c) variant call rates ≥ 0.02 and d) population structure showing more than 4 standard deviations within the distribution of the study population according to the first seven principal components.

2.5 Genotype phasing, genotype imputation and association analysis

To efficiently perform genome-wide imputation and association testing we have used **GUIDANCE**, an integrated framework developed in our group, which allows phasing into haplotypes, imputing genotypes and performing association testing in a single execution with optional user intervention (under review) (Sanchez et al. 2016). This application performs the currently extended GWAS strategy making an efficient use of computational resources without requiring specific expertise in parallel computing. As input, we provided quality controlled genotyping array data in any of the commonly used file formats (PLINK and IMPUTE2 (Howie et al. 2009; Howie et al. 2012) formats), phenotype with covariate information and the reference panels. We performed a **two-stage imputation procedure**, which consisted in pre-phasing the genotypes into whole chromosome haplotypes followed by imputation itself. The pre-phasing was performed using SHAPEIT2 (Delaneau et al. 2013), IMPUTE2 for genotype imputation and the SNPTEST tool (https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html) for association testing. We **integrated the association results** from **1000G-Phase1** (June, 2014) (The 1000 Genomes Project Consortium et al. 2012) and **UK10K** (UK10K Consortium et al. 2015) reference panels by choosing, for each variant, the reference panel that provided the best IMPUTE2-*info score*. Notice that for 1000G-based genotype imputation in chromosome X (chrX) we had to rely on the "v3.macGT1" release (August, 2012). Imputation for each reference panel was obtained separately, applying stringent quality filtering criteria (including variants with IMPUTE2-*info* score ≥ 0.7 MAF ≥ 0.001, HWE controls > $1x10^{-6}$). Association testing was performed considering an additive logistic

regression using SNPTEST, the 7 derived principal components, sex, age and BMI, except for WTCCC, where age and BMI was not available (Supplementary Material 1). For chrX, we restricted the analysis to non-pseudoautosomal (non-PAR) regions. We stratified the analysis by sex, which fits the additive logistic regression considering independent effects for males (coded as 0/1) and females (coded as 0; ½; 1) and baseline terms independent for males and females. Thus, we accounted for hemizygosity in the chrX for males, while for females we followed an autosomal model.

For the GERA cohort, due to the large computational burden that comprises the whole genotype imputation process in such a large sample size, we randomly split this cohort in two homogeneous subsets of ~30,000 individuals each, in order to specially minimize the memory requirements.

2.6 70KforT2D meta-analysis and inclusion of publicy available summary statistics

We meta-analyzed the different sets from the **70KforT2D dataset** with METAL (Willer et al. 2010), using the inverse variance fixed effects meta-analysis.

For the **meta-analysis with the DIAGRAM Trans-Ethnic study**, we excluded from the whole 70KforT2D datasets those cohorts that overlapped with the DIAGRAM data. Therefore, we meta-analysed the GERA and NuGENE cohorts (7,522 cases and 50,446 controls) from the 70KforT2D analysis with the Trans-Ethnic summary statistics results. As standard errors were not provided for the DIAGRAM Trans-Ethnic meta-analysis, we performed a sample size based meta-analysis, which converts the direction of the effect and the p-value into a Z-score. In addition, we also performed an inverse variance fixed effects meta-analysis to estimate the final effect sizes. This approach required the estimation of the beta and standard errors from the summary statistics (p-value and odds ratio).

For the **meta-analysis with the T2D Portal** we only included from the 70KforT2D cohort the NuGENE, GENEVA and GERA cohorts (8,136 cases and 53,507 controls) to avoid overlapping samples. Like in the previous scenario, standard errors were not provided for the T2D Portal and we used a sample-size based meta-analysis with METAL. However, to estimate effect sizes we also estimated the standard errors from the p-values odds ratios (OR), and we performed inverse-variance fixed effects meta-analysis.

For the **replication of the chromosome Xq23 variant** in the Danish cohorts, we used a **meta-analysis method that accounts for overlapping samples (MAOS)** (Lin and Sullivan 2009), as there was sample overlap between the **follow-up results** and the **case-control analysis**.

2.7 Pathway and enrichment analysis

In order to provide biological hypothesis from our GWAS results, we used **DEPICT (Data-driven Expression Prioritized Integration for Complex Traits)** (Pers et al. 2015) to prioritize the likely causal genes at associated *loci*, to highlight **enriched pathways** based on genes in associated loci, and to identify **tissues/cell types** where genes from associated *loci* are highly expressed. DEPICT

relies on publicly available gene sets (including molecular pathways) and leverages gene expression data from 77,840 gene expression arrays to perform gene prioritization and gene set enrichment based on predicted gene function and so-called reconstituted gene sets. A reconstituted gene set contains a membership probability for each gene and conversely each gene is functionally characterized by its membership probabilities across 14,461 reconstituted gene sets. As input to DEPICT we used all summary statistics from autosomal variants with p-value<$1x10^{-5}$ in the 70KforT2D meta-analysis. We used an updated version of DEPICT which handled 1000G-Phase1 integrated haplotypes (The 1000 Genomes Project Consortium 2010; The 1000 Genomes Project Consortium et al. 2012) (June, 2014, www.broadinstitute.org/depict). DEPICT was run using 3,412 associated SNPs with p-value<$1x10^{-5}$. From these, we identified independent SNPs (PLINK clumping parameters: --clump-p1 5e-8 --clump-p2 1e-5 --clump-r2 0.6 --clump-kb 250). LD r^2>0.5 distance cut-off was used to define *locus* boundaries (note that this *locus* definition is different than the *locus* definition user elsewhere in the text) yielding 70 autosomal *loci* comprising 119 genes. DEPICT was run using default settings, that is using 500 permutations for bias adjustment, 50 replications for false discovery rate estimation, normalized expression data from 77,840 Affymetrix microarrays for gene set reconstitution (see reference (Pers et al. 2015) for details), 14,461 reconstituted gene sets for gene set enrichment analysis, and testing 209 tissue/cell types assembled from 37,427 Affymetrix U133 Plus 2.0 Array samples for enrichment in tissue/cell type expression. From DEPICT we identified 103 reconstituted gene sets that are significantly enriched (FDR<5%) for genes found among the 70 trait associated *loci*. After the gene set enrichment analysis, we omitted reconstituted gene sets in which genes in the original gene set were not nominally enriched (Wilcoxon rank-sum test). By design, genes in the original gene set are expected to be enriched in the reconstituted gene set; lack of enrichment therefore complicates interpretation of the reconstituted gene set because the label of the reconstituted gene set will be inaccurate. Using this procedure, the following reconstituted gene set were removed from the results (Wilcoxon rank-sum P-values in parentheses): MP:0004247 gene set (P=0.73), GO:0070491 gene set (P=0.14), MP:0004086 gene set (P=0.173264735083), MP:0005491 gene set (P=0.54), GO:0005159 gene set (P=0.04), MP:0005666 gene set (P=0.05), ENSG00000128641 gene set (P=0.02), MP:0006344 gene set (P=0.42), MP:0004188 gene set (P=0.22), MP:0002189 gene set (P=0.02), MP:0000003 gene set (P=0.0845155407131), ENSG00000116604 gene set (P=0.13), GO:0005158 gene set (P=0.07), MP:0001715 gene set (P=0.014). The post-analysis filtering step left us with 89 significantly enriched reconstituted gene sets. The Affinity Propagation tool (Frey and Dueck 2007) was used to cluster related reconstituted gene sets (script to produce the network diagram can be downloaded from https://github.com/perslab/DEPICT).

2.8 Definition of 99% credible sets of GWAS significant loci

We defined the 99% credible sets of variants, which represent all the variants that have, in aggregate, **99% probability of containing the causal variant** driving the association with T2D.

For each genome-wide significant region, we constructed the 99% credible of variants considering 1 Mb downstream and upstream from the top SNP using our 70KforT2D meta-analysis based on imputed data (Northwestern NUgene**,** GERA, FUSION, GENEVA and WTCCC datasets, comprising 12,231 cases and 57,196 controls). We computed the R^2 values from all the variants within this 2 Mb region with respect to the top SNP and we selected variants showing an $R^2 > 0.1$ with the leading SNP in each region.

Credible sets of variants are analogous **to confidence intervals** as we assume that the credible set for each associated region contains, with 99% probability, the true causal SNP if this has been genotyped or imputed (Wellcome Trust Case Control et al. 2012; Morris 2014). The credible set construction allows to provide for each variant placed within a certain associated *locus* a **posterior probability of being the causal one** (Wellcome Trust Case Control et al. 2012). We estimated the **approximate Bayes' factor (*ABF*)** for each variant that can be calculated as:

$$ABF = \frac{\sqrt{1-r}}{e^{(-r\cdot\frac{z^2}{2})}}$$

where:

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

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

The β and the SE (standard error) are the estimated effect size and the corresponding standard error resulting from testing for association under a logistic regression model. The posterior probability for each variant was obtained as:

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

where ABF_i corresponds to the approximate Bayes' factor for the marker *i,* and *T* represents the sum of all the *ABF* values from the candidate variants enclosed in the interval being evaluated. This calculation assumes that the prior of the β corresponds to a Gaussian with mean 0 and variance 0.04, which is also the same prior commonly employed by SNPTEST (Marchini et al. 2007), the program used for calculating single-variant associations.

Finally, we ranked variants according to the *ABF* (in decreasing order) and from this ordered list we calculated the **cumulative posterior probability**. We included variants in the 99% credible set of each region until the cumulative posterior probability of association exceeded 0.99.

2.9 Conditional analysis of putative candidate regions

For the *EHMT2* region, since this region was less than 1 Mb away from the ***HLA* region where T1D and T2D associations** have been described, we performed a series of **conditional analyses** to exclude that this association was independent of both *loci*, and also to discard that this association is driven by possible contamination of T1D diagnosed as T2D cases. For the conditional analysis framework, we included the lead SNP in our study as a covariate in the logistic regression model, assuming that every residual signal arisen corresponds to a secondary signal independent from the lead SNP (Yang et al. 2012; Morris 2014). We performed the analyses, conditioning on the top variant identified in this study, but also on the top variants previously described for T2D and T1D (Hakonarson et al. 2007; Wellcome Trust Case Control 2007; Barrett et al. 2009a; Cook and Morris 2016). For this purpose, we used the full 70KforT2D resource (Northwestern NUGene, GERA, FUSION, GENEVA and WTCCC cohorts imputed with 1000G-Phase1 and UK10K reference panels). Finally, all the results were meta-analyzed as explained in previous sections.

2.10 Fine-mapping and functional annotation

We used the **Variant Effect Predictor (VEP)** (McLaren et al. 2010) for the functional characterization of the variants of the 99% credible sets. The VEP application determines the effect of variants (SNPs, insertions, deletions, CNVs or structural variants) on genes, transcripts, protein and regulatory regions. We used as input the coordinates of the variants within the 99% credible sets and the corresponding alleles to find out the genes and RefSeq transcripts affected by the variants and the consequence of our variants on the protein sequence.

We used **Combined Annotation Dependent Depletion (CADD) scoring function** to obtain an alternative metric for prioritizing functional, deleterious and disease causal variants (Kircher et al. 2014). This framework integrates multiple annotations in one metric, the C-score. This new metric correlates with allelic diversity, pathogenicity of regulatory effects and highly ranks causal variants within individual genome sequences.

In order to **prioritize functional regulatory variants**, we used the V6 release from the **GTEx data** that provides gene-level expression quantifications and eQTL results based on the annotation with GENCODE v19. This release included 450 genotyped donors, 8,555 RNA-seq samples across 51 tissues and 2 cell lines, which led to the identification of eQTLs across 44 tissues (Carithers and Moore 2015). Moreover, **RNA-seq data from human pancreatic islets** from 89 deceased donors catalogued as eQTLs and exon use (sQTL) was also integrated with the GWAS data to prioritize candidate regulatory variants (Fadista et al. 2014).

2.11 Characterization of indels

For all GWAS significant INDELs within the 99% credible sets, we examined whether these variants were present or absent in the 1000G-Phase1 or UK10K reference panels. We also compared whether the structural variants were present or not in the 1000G-Phase3 reference panel.

Finally considering all these previous analyses, we visually inspected the aligned BAM files of the most relevant INDELs from both projects to discard that they could be alignment artifacts.

2.12 In *silico* functional characterization of X chromosome variant with Roadmap Epigenome data

To evaluate the putative regulatory role of rs146662075 we used the **WashU EpiGenome Browser** (http://epigenomegateway.wustl.edu/browser/ , last access on June 2016) (Zhou et al. 2011; Zhou et al. 2013; Zhou et al. 2015). We used public hubs of data: (1) the Reference human epigenomes from the Roadmap Epigenomics Consortium track hubs and (2) the Roadmap Epigenomics Integrative Analysis Hub. This data was released by the NIH Roadmap Epigenomics Mapping Consortium. By exploiting next-generation sequencing technologies, the consortium was able to map DNA methylation, histone modifications, chromatin accessibility and small RNA transcripts in stem cells and primary *ex vivo* tissues selected as representatives of the normal counterparts of tissues and organ systems predominant in human diseases. The current Release 9 contains for each epigenomic data type across 183 biological samples, named as unconsolidated epigenomes because of the redundancies resulting from the existence of multiple samples from a particular unique cell type or tissue. All this experimental data was processed in order to decrease redundancy as well as to increase data quality and uniformity, leading to 111 consolidated epigenomes (Roadmap Epigenomics et al. 2015). In addition, in the final integrative analyses 16 epigenomes from The Encyclopedia of DNA Elements (ENCODE) (Encode Project Consortium et al. 2007) project have been processed similarly, resulting in a total of 127 consolidated epigenomes.

Since this variant was located in a highly conserved region, surrounded by **several DNAse I hypersensitive sites**, we searched for enhancer marks through the **HaploReg** web server (Ward and Kellis 2012a; Ward and Kellis 2016) in order to assess if the rs146662075 variant in Xq23 was located within an active enhancer. Alongside with H3K4Me1, there were **several H3K27Ac marks** across multiple tissues including Fetal Muscle Leg and Fetal Muscle Trunk. **RNA-seq data** was therefore used to evaluate whether **gene expression** of any of the **closest genes (*AGTR2* and *SLC6A14* genes)** from rs146662075 (fixed scale at 80 RPKM), correlated with the **presence of the H3K27ac enhancer marks** (a more strict marker for active enhancers in contrast with H3K4Me1 marks (Creyghton et al. 2010), also highlighted by the HaploReg search (Ward and Kellis 2012a; Ward and Kellis 2016)) through the WashU Epigenome Browser. For visualizing the H3K27ac marks around rs146662075, we focused on a region of 8 Kb and we used a fixed scale at 40 $-\log_{10}$ Poisson p-value of the counts relative to the expected background count (λ_{local}).

2.13.1 Electrophoretic Mobility Shift Assay

Nuclear extracts from **mouse myoblast C2C12 cells** (ATCC CRL-1772, kind gift from Antonio Zorzano, IRB, Barcelona) were obtained as described elsewhere. Double stranded oligonucleotides containing either the common or rare variants of rs146662075 were labeled using dCTP [α-32P] (Perkin Elmer). Oligonucleotide sequences are as follows (SNP location is underlined): probe-C-F: 5'-gatcTTTGAACAC<u>c</u>GAGGGGAAAAT-3' and R:5'-gatcATTTTCCCCTC<u>g</u>GTGTTCAAA-3' and probe-T- F: 5'- gatcTTTGAACAC<u>t</u>GAGGGGAAAAT-3' and R: 5'-gatcATTTTCCCCTC<u>a</u>GTGTTCAAA -3'**.** Assay specificity was assessed by pre-incubation of nuclear extracts with 50- and 100- fold excess of unlabeled wild-type or mutant probes, followed by electrophoresis on a 5% non-denaturing polyacrylamide gel. Findings were confirmed by repeating binding assays on separate days.

2.13.2 Luciferase assays of AGTR2 variant (rs146662075)

A region of 969 bp surrounding rs146662075 was amplified from human genomic DNA using F: 5'-GCTAGCATATGGAGGTGATTTGT -3' and R: 5'-GGCACTTCCTTCTCTGGTAGA-3' oligonucleotides and cloned into pENTR/D-TOPO (Invitrogen). Allelic variant rs146662075T was introduced by site-directed mutagenesis using the following primers: F: 5'-CCTTTTTTTACTTTGAACACTGAGGGGAAAATCATGCTTGGC -3' and R: 5'-GCCAAGCATGATTTTCCCCTCAGTGTTCAAAGTAAAAAAAGG-3'. Enhancer sequences were shuttled into pGL4.23[luc2/minP] vector (Promega) adapted for Gateway cloning (pGL4.23-GW, **2)** using Gateway LR Clonase II Enzyme mix (Invitrogen). Correct cloning was confirmed both by Sanger sequencing and restriction digestion.

C2C12 (ATCC CRL-1772) and **293T** (ATCC CRL-3216) **cells** were transfected in quadruplicates with 500 ng of pGL4.23-GW enhancer containing vectors and 0.2 ng of Renilla normalizer plasmid. Transfections were carried out in in 24-well plates using Lipofectamine 2000 and Opti –MEM (Thermo Fisher Scientific) following manufacturer's instructions. Luciferase activity was measured 48 h after transfection using Dual-Luciferase Reporter Assay System (Promega). Firefly luciferase activity was normalized to Renilla luciferase activity and results were expressed as a normalized ratio to the empty pGL4.23[luc2/minP] vector backbone. Statistical significance was evaluated through a t-Student's test.

Results

This thesis has sought for a **cost-effective strategy to push forward the discovery of novel *loci* for complex traits by assuming that improved analytical techniques and methodologies can extract novel information from a huge amount of GWAS data**, which already have been analysed and are publicly available.

In the first main block (1), I will detail the amenities of a packaged analytical workflow for **performing systematic Quality Control (QC) protocols for genotyped data**. Afterwards, (2) I will present the **different guidelines and recommendations for integrating imputed data** in GWAS and meta-analytic approaches that I collected after revising multiple scenarios derived from genotype imputation with novel sequence-based reference panels.

In the second main block of results, I will describe how all this experience has been applied to the **analysis of publicly available T2D GWAS data** and which **discoveries of the T2D aetiology** have been attained following this strategy.

1. Implementation of efficient computational and analytical frameworks for imputation based GWAS

The development of this work has allowed the doctorand to contribute to the following two articles:

Bonnelykke K, Sleiman P, Nielsen K, Kreiner-Moller E, Mercader JM, Belgrave D, den, Dekker HT, Husby A, Sevelsted A, Faura-Tellez G … Bonàs-Guarch S … Bisgaard H 2014. A genome-wide association study identifies CDHR3 as a susceptibility locus for early childhood asthma with severe exacerbations. *Nat Genet* **46**: 51-55.

Horikoshi M, Beaumont RN, Day FR, Warrington NM, Kooijman MN, Fernandez-Tajes J, Feenstra B, van Zuydam NR, Gaulton KJ, Grarup N … Bonàs-Guarch S … Freathy RM 2016. Genome-wide associations for birth weight and correlations with adult disease. *Nature* **538**: 248-252.

Contribution of the PhD candidate

- Development and execution of an efficient computational framework for quality control of genotyped data.
- Implementation of accurate practices for quality filtering imputed variants through sequence-based reference panels for a broad deployment of genotype imputation in GWAS scenarios which was applied in both studies.
- Costumed statistical and bioinformatic data analyses.

1.1 Automatization and packaging of GWAS analytical workflows

In order to favour robustness and portability in GWAS and genotype imputation, automatized analytical workflows to perform systematic analysis are required. Most systematic biases in GWAS approaches come from **wrong genotype calling** or study design. These artefacts have the potential of dramatically increasing the number of false positives and negative associations (loss of statistical power), especially when performing genotype imputation. We packaged **widely accepted practices for the QC of genotyped data**, and we integrated them in an **automatized pipeline for a quick and effective identification and removal of defective markers and samples**.

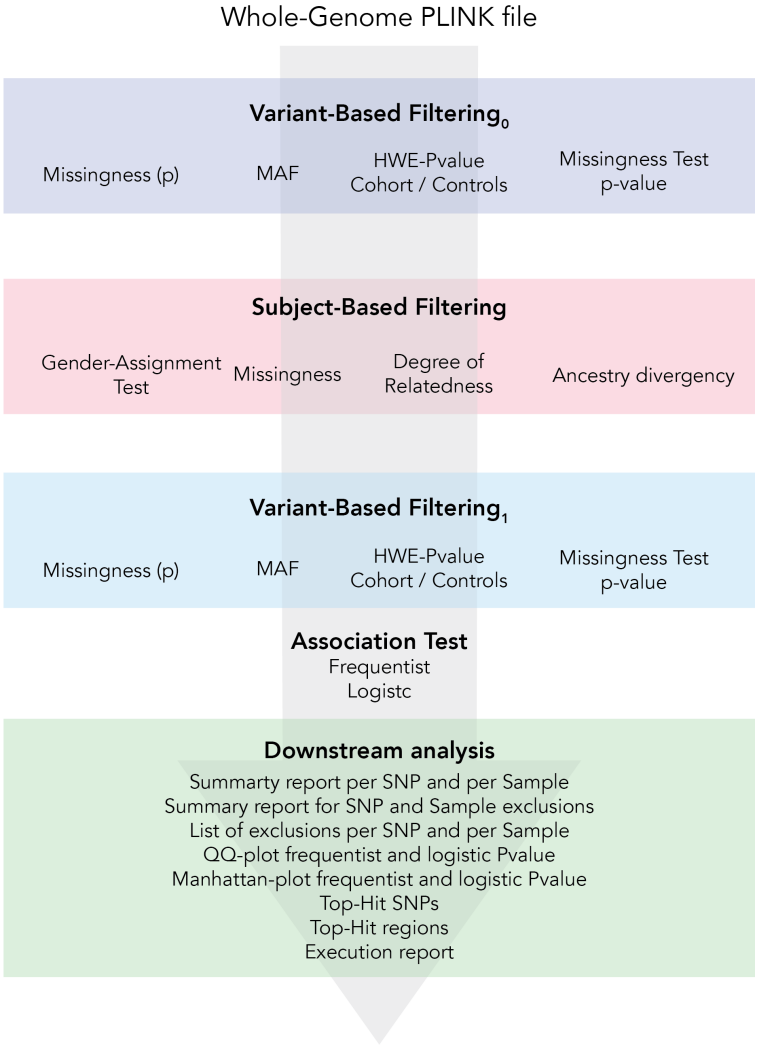


Figure 13. QC Protocol analysis. First, a PLINK format file containing the genotyped data should be provided. This workflow performs three stages of two variant-based filtering and an intermediate stage of sample-based filtering. Different parameters and metrics are evaluated. At the marker level: allele frequency, proportion of missing genotypes, statistical differences in the proportion of missing genotypes between cases and controls and deviance of HWE. At the sample level: assignment of gender information, proportion of missing genotypes, the degree of relatedness and ancestry divergence. Finally, multiple text and graphical reports are provided.

This protocol is based on exploiting the capabilities of the PLINK software (Purcell et al. 2007), UNIX-environment and R-scripting, which also allows performing the graphical steps of the workflow. R scripting also has a master role in controlling all the executions and in performing parallelism computing according to the queuing systems of the cluster infrastructure (see Methods).

This protocol consists in **three main stages**, and ends with multiple reports that summarize the different analyses performed at the marker and at the individual level (see Figure 13). As illustrated in Figure 13, this workflow first performs a per-marker QC focused on the following parameters: (1) Excess of missing genotypes per marker, (2) Significant deviation of HWE (which is evaluated for all the cohort, but also for controls and cases, separately), (2) Significant differences in the proportion of missing genotypes between cases and controls and (4) Low MAF values.

Second, per-individual QC is performed by the identification and removal of: (1) Subjects with discordant sex information, (2) Individuals showing excessive missing data proportion, (3) Highly related individuals and (4) Subjects of divergent ancestry. Finally, a new round of per-marker QC encompassing the same four steps previously explained is computed.

These practices are quite standard and widely accepted by the community (Anderson et al. 2010) and our main contribution is how we computationally articulated these analytical steps. From the computational point of view, the structure of the QC pipeline was conceived as exposed in Figure 14. The full control of the execution lies in a master R-script that performs task management and monitoring. This master R-script initializes the execution by checking the current working directory, the full disposal of the required R-libraries, the cluster machines and the corresponding queuing system in which the main execution is run and also takes into account the settings specified by the user. As represented in Figure 14, the master R-script launches the first block of tasks in a single job and periodically checks if the execution has finished. After that, the master R-script keeps on launching the rest of jobs in a sequential manner after each job has finished successfully. Each job entails different memory requirements and degrees of parallelism, which is tuned by the master R-script according to the sample size. Once the involvement of clustering resources is not necessary, downstream analyses are performed that result in the creation of summary statistics reports, graphical representations and a final report of the whole QC execution.

The user can execute this pipeline for any kind of genotyped data in a cluster environment by following this template:

```
/apps/R/3.0.2/bin/Rscript /path/to/QC/QC_master_pip.R

--out_dir /path/to/output/qc --input /path/to/inputPLINKformatfile/input

--genome hg18 --sex_check YES --gwas yes --pop_ref hapmap3 --hwe_coh 1e-20
```

As shown in the previous command-line, the input data (in PLINK format) is provided by the `--input` flag. The output data generated is collected in the directory specified by the `--out_dir` flag.

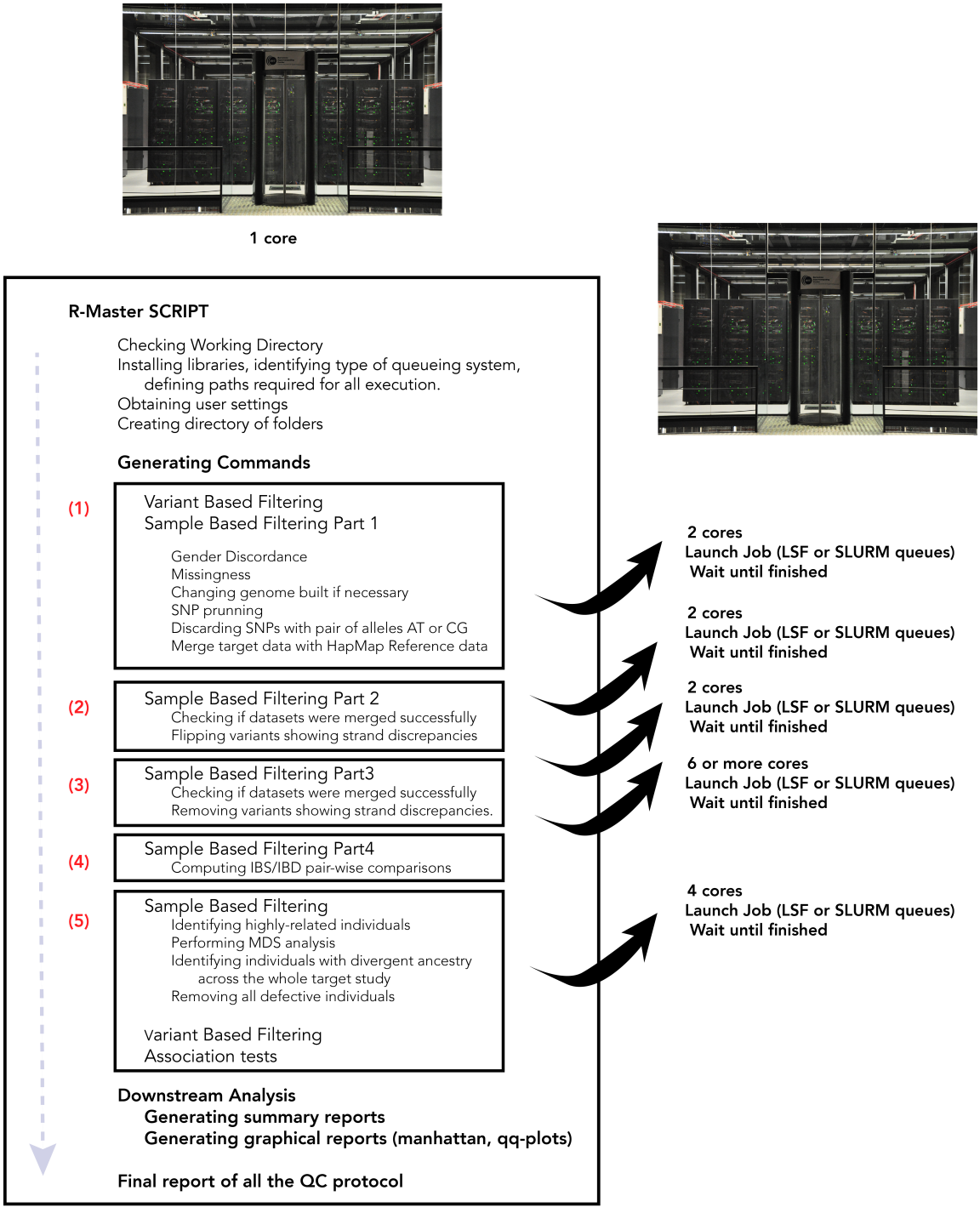


Figure 14. Packaged workflow for a QC protocol for genotyped data. The big black external box represents the execution of the master R-script that initializes the main execution and generates 5 jobs of several sub-tasks (displayed as the inner black squares). Each job is run in a cluster environment (right) in a sequential-manner but each job of tasks may involve different levels of parallelism. Finally, downstream analyses are performed and a final report of the QCed data is generated.

Moreover, besides using the default settings, all the available parameters can be adjusted. For instance, the genome built of the original data (`--genome`) shall be provided in order to get a correspondence with the build from the default HapMap data used for the population clustering analysis. Additionally, if the X-chromosome data is available, discordance in gender assignment may be checked (`--sex_check`). If the input data corresponds to a case-control study, association tests and the stratification of HWE filtering across cases or controls may be performed (`--gwas`). All the available options are detailed with the `--help` information to improve usability.

An example of a QC of 5,828 samples and 741,192 genotyped variants required 2h 51' 14'' to perform the full protocol (83 tasks in this case) within the *MareNostrum III* cluster (Highest parallelism 6 cores from a node of 16 cores/node, Intel SandyBridge, 2.6 GHz with 32 Gb/memory per node). Below, I provided some of the graphical outputs obtained such as the Q-Q-plots for the HWE p-values and for the association test p-values under the logistic regression and the corresponding Manhattan plot (Figure 15). Interestingly, the multidimensional new components generated during clustering analysis are collected in one of the output files and may be used to adjust the logistic regression model used in association testing for population structure. These basic outputs serve as a manual inspection to ensure that any artefact escaped from the filtering steps of the QC protocol at the variant level. Moreover, the clustering plots from Figure 16 are a useful representation to identify any batch effect from persistent differences in population structure within our study. We also used the multidimensional new components to identify any batch effect between cases and controls that can result in spurious associations and loss of statistical power in the association tests (Figure 17).

Finally, a summary report (Figure 18) is generated and details all the filtering steps performed, the parameters used and the number of variants and individuals discarded after each step that facilitates reviewing all the analysis. This kind of computational optimizations provides robustness and replicability in analytical workflows and helps the researcher to specifically focus on the interpretation of the results.

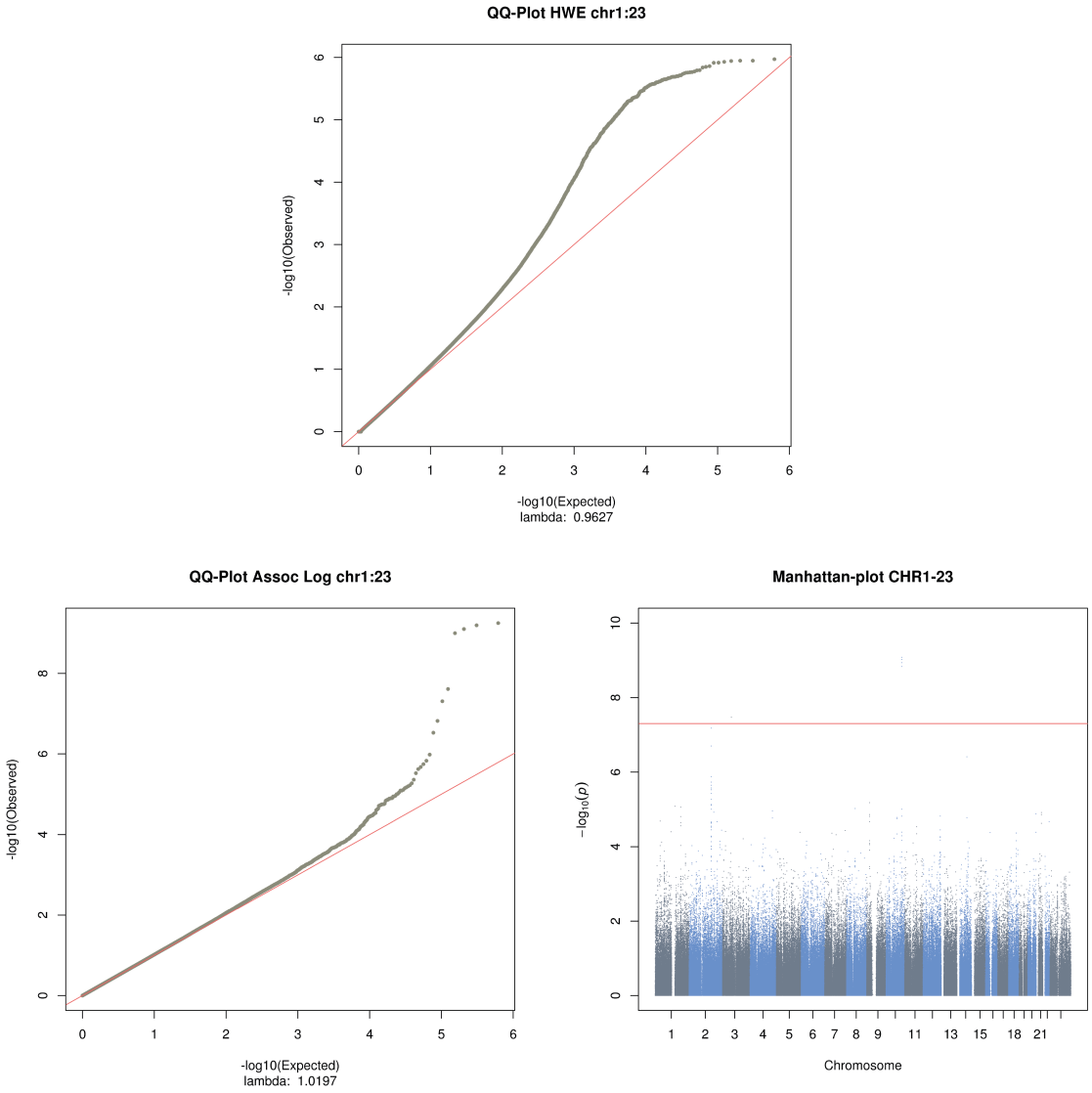


Figure 15. Q-Q-plots for HWE and logistic regression p-values, and the Manhattan plot. The Q-Q-plots represent the expected (under the null hypothesis) and the observed $-\log_{10}$ p-values for HWE deviance in controls and the logistic regression tests in the x and y-axis, respectively. The Manhattan plot (bottom-right) represents each association $-\log_{10}$ p-value under logistic regression across the genome.

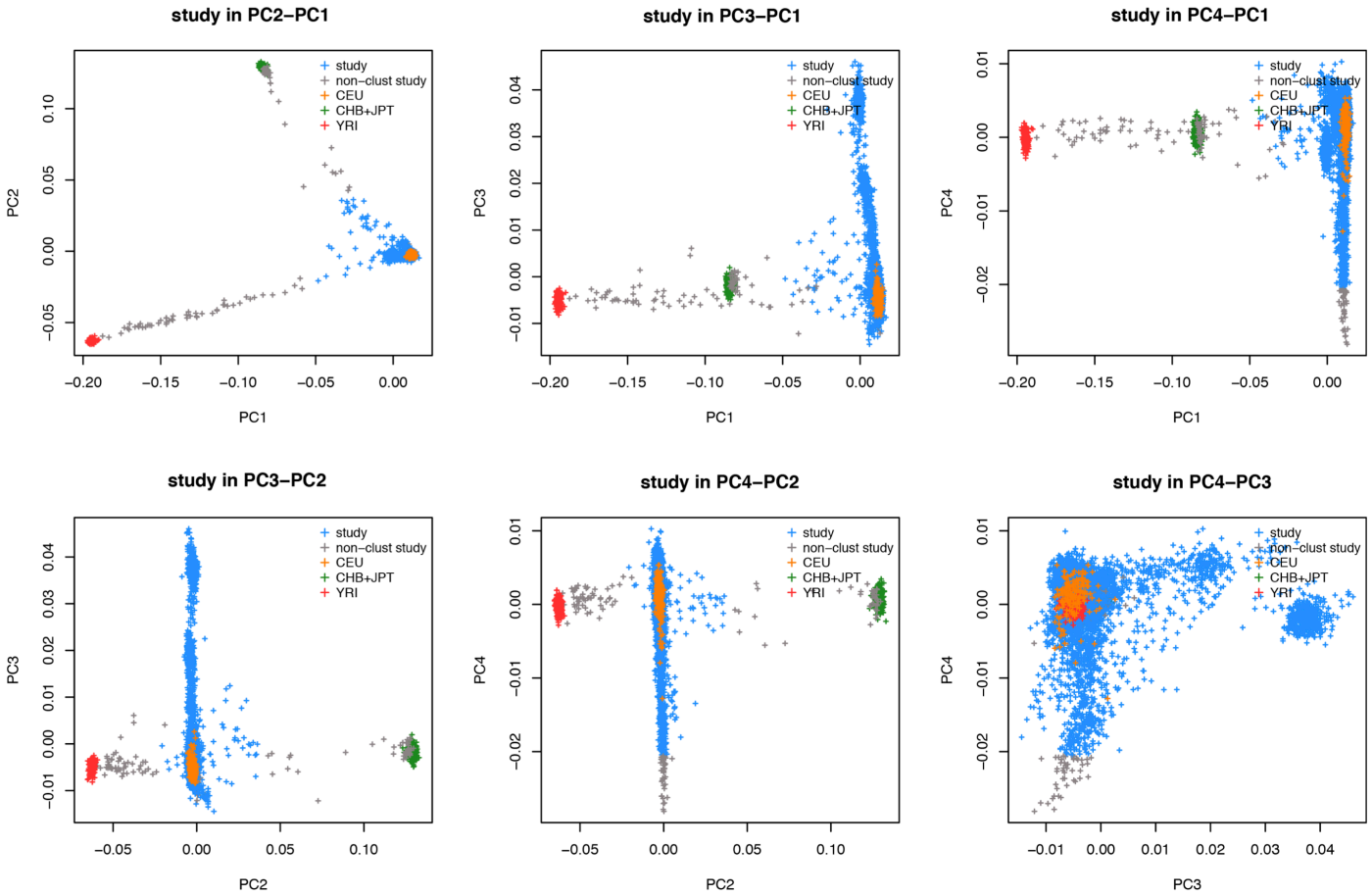


Figure 16. Clustering of a case-control cohort across different population ancestries and according to four multidimensional scaling components. Each dot corresponds to a single individual and y and x-axis represents the components generated in a multidimensional scaling analysis (mds), in a multiple component comparison. Reference populations have been obtained from HapMap3, and coloured accordingly: European Ancestry (CEU) as orange, African from Yoruba population (YRI) as red and Asian Ancestry is represented by two subsets (CHB+JPT). Study samples are coloured in blue and those individuals not clustering within the whole study population (and removed), as grey (see Methods for the exclusion filter).

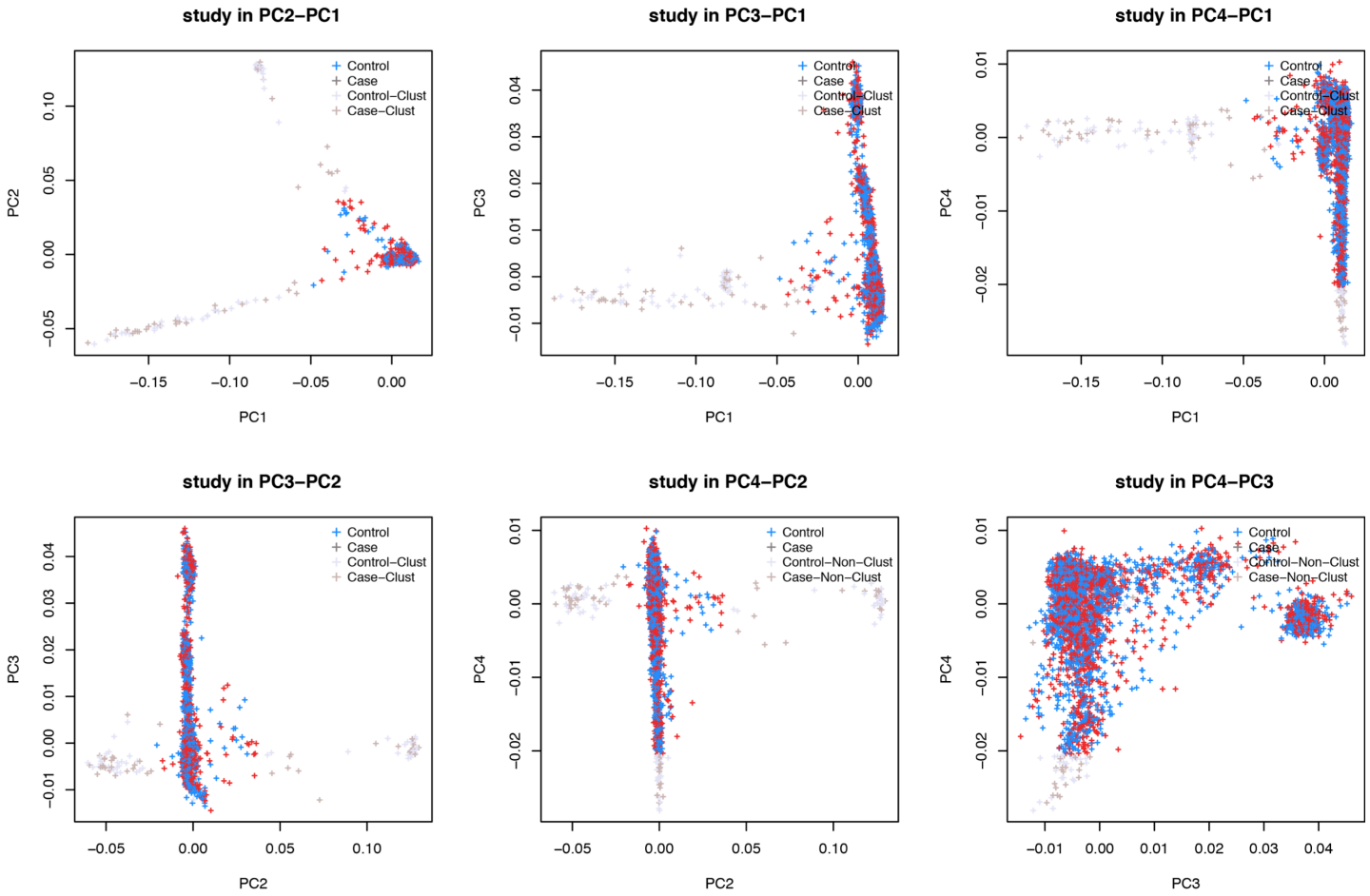


Figure 17. Clustering of a case-control cohort according to four multidimensional scaling components. Each dot corresponds to a single individual and y and x-axis show the components generated in a multidimensional scaling analysis (MDS). Study samples are coloured in blue for controls and red for cases. Those individuals that were excluded (see Methods) are coloured in light red and grey.

```
################################################################################
                    QUALITY-CONTROL ANALYSIS FOR GWAS DATA
################################################################################
Computational Genomics Group - Barcelona Supercomputing Center
Sat Apr 25 14:24:23 2015
Input:
      -SNPs (n):  741192     -Subjects (n):  5828

Cut-Offs used to filter the data:
      -MAF                 = 0.01
      -Missingness per SNP = 0.05
      -Missingness per ind = 0.02
      -Missingnes Pvalue   = 1e-06
      -HWE Cohort Pvalue   = 1e-20
      -HWE Ctrls  Pvalue   = 1e-06
      -HWE Cases  Pvalue   = 0

Stage0: Variant Based Filtering
            -Discarded SNPs by MAF (n): 109392
            -Discarded SNPs by Missingness (n): 16741
            -Discarded SNPs by HWE Cohort (n): 44
            -Discarded SNPs by HWE Ctrls  (n): 1696
            -Discarded SNPs by HWE Cases  (n): 0
            -Discarded SNPs by Test-Missingness (n): 11
DISCARDED SNPs AFTER STAGE0 (n): 119971

Stage1: Subject Based Filtering
            -Discarded subjects by Gender (n): 18
            -Discarded subjects by Missingness (n): 60
            -Discarded subjects by Relatedness (PI_HAT > 0.125) (n): 0
            -Discarded subjects by Population clustering (non-EU) (n): 93
DISCARDED SUBJECTS AFTER STAGE1 (n): 153

Stage2: Variant Based Filtering
            -Discarded SNPs by MAF (n): 2760
            -Discarded SNPs by Missingness (n): 0
            -Discarded SNPs by HWE Cohort (n): 0
            -Discarded SNPs by HWE Ctrls  (n): 8
            -Discarded SNPs by HWE Cases  (n): 0
            -Discarded SNPs by Test-Missingness (n): 1

FINAL SNPs     (n): 618452
FINAL SUBJECTS (n): 5675
Time-finished :  Sat Apr 25 14:24:28 2015
                      //// QC-PROTOCOL FINALLY COMPLETED ////
```

Figure 18. Summary report of a QC analysis. The initial number of variants and markers are provided at the beginning as well as the settings for this analysis. Afterwards, for each stage (Stage0 Variant-Based, Stage1 Sample-Based and Stage 2 Variant-based filtering) and each specific filtering process, the number of variants or samples discarded, are detailed. Finally, the resulting coverage and sample size is presented.

1.2 Fostering guidelines for accurate genotype imputation of common and low-frequency variants for GWAS and sequence-based reference panels

To evaluate genotype imputation quality across the broader spectrum of allele-frequency and diverse GWAS scenarios, we imputed genotypes into the 58C cohort from the WTCCC2 (Barrett et al. 2009c), which entails ~3,000 individuals that were genotyped by both Affymetrix v6.0 (Affy) and Illumina 1.2M (IL) platforms. We performed genotype imputation independently using either Affy or IL genotypes as the backbone. The underlying rationale is that, despite using a different genotyping array, as we were imputing the genotypes into the same subjects, we would not expect differences in allele frequencies if there were no genotyping or imputation bias (Figure 19).

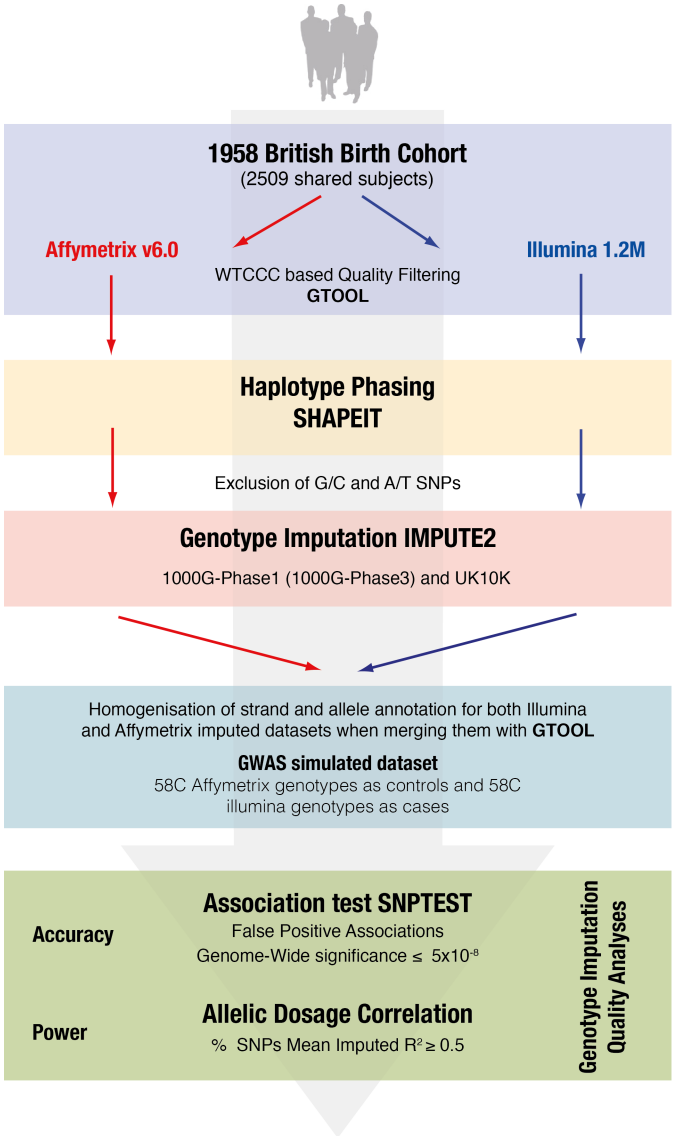


Figure 19. Strategy to evaluate genotype imputation accuracy. The 58C control dataset (~2,509 shared samples) was genotyped by two different platforms (Affy and IL). 58C Affy and IL data were (1) variant and sample quality filtered according to the guidelines specified by the WTCCC study, (2) phased using SHAPEIT2 and (3) genotype imputation was performed with IMPUTE2. Multiple reference panels were used for imputation separately (1000G-Phase1, 1000G-Phase3 and UK10K). Imputed results from Affy ("controls") and IL ("cases") were merged simulating a GWAS scenario. Accuracy was estimated as the number of genome-wide significant associations (FP) from testing for association the Affy-based with the IL-based genotypes. Power was measured as the R-squared correlation coefficient (mean imputed R^2) between the dosages from each set of imputed results.

Of note, our accuracy estimators were based on false positives (FP), which correspond to variants showing genome-wide significance ($5x10^{-8}$) when comparing the results between Affy and IL. Likewise, we would expect a clear correlation between the genotypes imputed from IL and those imputed from Affy, which has been measured with the mean imputed R^2, the R-squared correlation coefficient calculated by comparing the dosages from each set of imputed results. For the analysis of a scenario for which the same platform is used for cases and controls, we evaluated how different post-imputation QCs may affect the I^2 heterogeneity measure in meta-analytic approaches.

1.2.1 Fixing appropriate quality thresholds across genotyping platforms

- *Quality Control filtering for imputed variants*

We first evaluated the performance of the IMPUTE2-*info* quality score. We addressed the performance of loose IMPUTE2-*info* cut-offs such as 0.3 and 0.5. Loose cut-offs are still used, even in recent publications based on genotype imputation with for instance, the unified panel of 1000G-Phase1+UK10K (Wain et al. 2015) (IMPUTE2-*info* cut-off = 0.5) and the 1000G-Phase3+UK10K (Lane et al. 2016) (IMPUTE2-*info* cut-off = 0.1) generated by Huang, J. and colleagues (Huang et al. 2015). By means of the simulated GWAS that allowed us to compare the imputed results from Affy and IL using 1000G-Phase1 as the reference panel, we addressed the rate and the absolute number of FP associations for these still widely used thresholds. We obtained 36,746 (0.31%) and 19,377 (0.18%) genome-wide significant FP associations for a 0.3 and 0.5 IMPUTE2-*info* score, respectively.

Thereafter, we sought for a criterion to consistently filter badly imputed variants based on the IMPUTE2-*info* score. To do that, we focused on the analysis of SNPs imputed on one platform, but genotyped on the other platform (i.e. genotyped by Affy and imputed from IL and vice-versa). This analysis showed us that there is not a one-to-one correlation between the IMPUTE2-*info* scores and the true allele dosage R^2 correlation coefficients. In fact, the IMPUTE2-*info* measure was overestimating the imputed mean R^2 (corresponding to the allele dosage R^2 correlation between the imputed and the true genotyped allele dosages). We therefore identified that an IMPUTE2-*info* score cut-off of 0.7 was needed in order to achieve an allelic dosage R^2 of at least 0.5, which was determined to discern the fraction of well-imputed variants in the original MACH paper (Li et al. 2010). Thus, we recommend using as a general practice for genotype imputation an IMPUTE2–*info* score cut-off above 0.7. This linear relationship between the imputed mean R^2 and the *info* score provided by IMPUTE2 was also observed across different reference panels and genotyping platforms (Figure 20). This threshold led to 7,063 (0.07%) FP with the 1000G-Phase1 reference panel, which implies a decrease of 76.49% and 58.79% in the percentage of FP with respect to the IMPUTE2-*info* cut-offs 0.3 and 0.5, respectively.

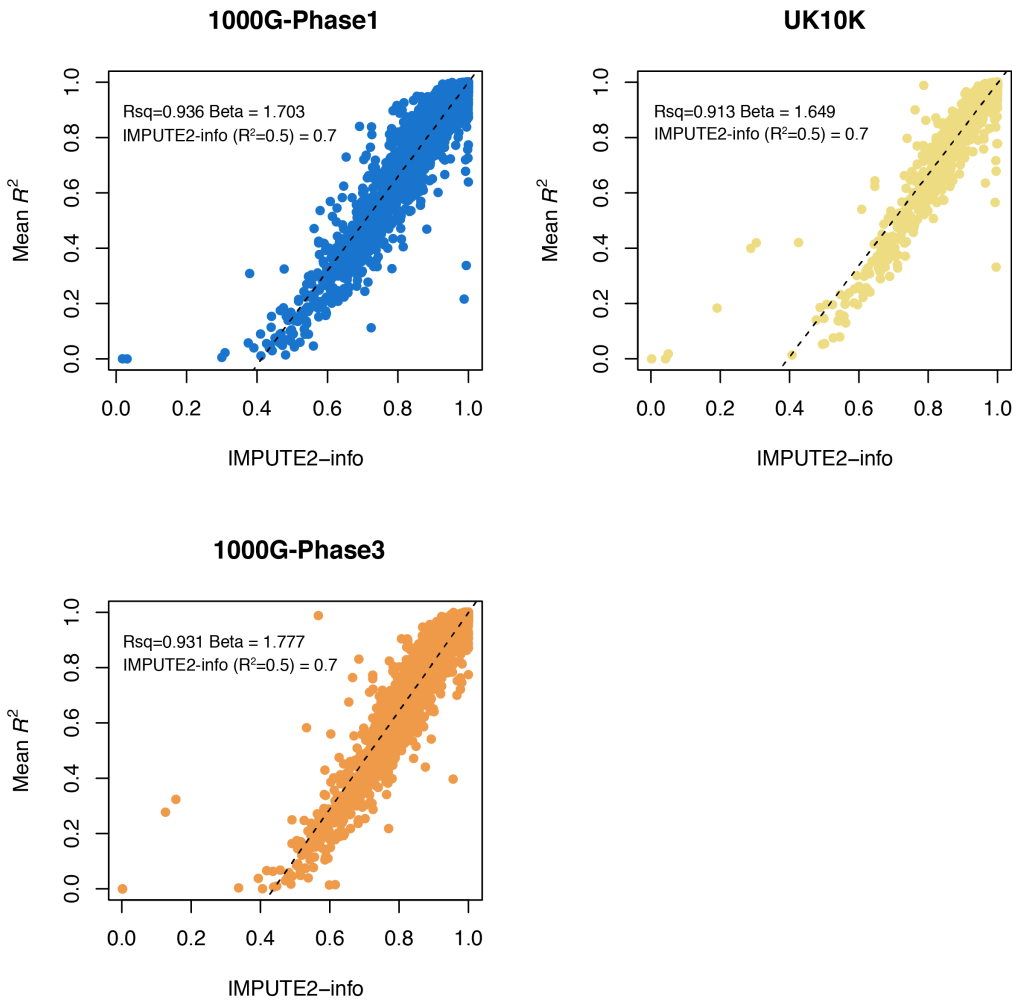


Figure 20. Correlation between IMPUTE2-*info* scores and allele dosage imputed R^2**.** All the SNPs on chromosome 21 that were either genotyped by Affy and imputed by IL or vice-versa, using 1000G-Phase1, UK10K and 1000G-Phase3 as reference panels are represented. The plot shows that IMPUTE2-*info* overestimates the real allele dosage correlation.

- *Seeking for an increased genomic coverage across the allele frequency spectrum*

After identifying an appropriate quality filter to retain well-imputed variants (IMPUTE2-*info* score ≥ 0.7), we evaluated **genotype imputation accuracy across the whole allele spectrum** through 1000G-Phase1 (and lastly, the 1000G-Phase3) and UK10K reference panels.

As shown in Table 2 and Table 3, we exposed how the UK10K reference panel led to the most accurate imputed results at any range of the allele spectrum. Moreover, we realized that 1000G-Phase3 extensively improved the coverage for rare variants with respect to 1000G-Phase1 (67.82% increase in coverage, 6.67% decrease in the FP rate). With respect to imputed INDELs or larger deletions, there are no evidences of notable differences in the % of spurious associations in comparison with the fraction of imputed SNPs (1000G-Phase1 % FP SNP = 0.072, % FP INDELs = 0.074; UK10K % FP SNP = 0.019, % FP INDELs = 0.013, 1000G-Phase3 % FP SNP = 0.058, % FP INDELS = 0.041).

Table 2. Evaluation of Genotype Imputation. Assessment of genotype imputation with three reference panels (1000G-Phase1, UK10K and 1000G-Phase3) and two integrated set of imputed variants (1000G-Phase1-UK10K, 1000G-Phase3-UK10K) keeping, for each variant, the imputed results from the best panel based on IMPUTE2-info score. Four scenarios were considered: (a) Total Overall Variants, (b) IL GT SNPs (variants genotyped by Illumina, not present in the Affymetrix chip and imputed using this last genotyped data), (c) Affy GT SNPs (genotyped by Affymetrix and imputed using the IL data) and (d) Both imputed (imputed variants in both sets of data). We showed the number of variants, false positive signals and, in parenthesis, the ratio of false positive signals obtained

	1000G-Phase1		UK10K		1000G-Phase3		1000G-Phase1-UK10k		1000G-Phase3-UK10k	
	variants n(%)†	FP n(*p*)*	variants n(%)†	FP n(%)*	n(%)†	FP n(%)*	variants n(%)†	FP n(%)*	variants n(%)†	FP n(%)*
Total Variants	9,753,072	7,063 (0.072)	10,343,027	1,928 (0.019)	11,180,215	6,327 (0.057)	11,567,853	7,754 (0.067)	11,965,521	5,884 (0,049)
IL GT SNPs	610,135	573 (0.094)	595,520	290 (0.049)	608,703	604 (0.099)	616,813	463 (0.075)	593,782	461 (0,078)
Affy GT SNPs	353,521	347 (0.098)	342,839	239 (0.070)	351,415	316 (0.090)	354,503	305 (0.086)	341,616	284 (0,083)
Both imputed	8,539,953	6,018 (0.070)	9,156,904	1,276 (0.014)	9,971,865	5,297 (0.053)	10,346,993	6,862 (0.066)	10,788,806	5,031 (0,047)

Number of genetic variants with IMPUTE2-*info* score ≥ 0.7 (n). In parenthesis, for the IL GT and Affy GT SNPs.

*Number of false positive associations for each set of variants considering a genome-wide significance of $5x10^{-8}$. In parenthesis, the proportion of false positive signal

Table 3. Evaluation of genotype imputation across ranges of allele frequency. Assessment of genotype imputation according to two reference panels (1000G-Phase1, 1000G-Phase3, UK10K) and two integrated set of imputed variants (1000G-Phase1-UK10K, 1000G-Phase3-UK10K) keeping, for each variant, the imputed results from the best reference panel based on IMPUTE2-info score. This assessment was done for three ranges of the allele frequency spectrum (Rare, Low-frequency and Common), as well as the overall contribution per panel, also shown in Table 2. We showed the absolute number of variants and false positive signals and, in parenthesis, the ratio of false positive signals obtained.

	All Variants		Rare Variants (0.001 ≤ MAF < 0.01)		Low-Frequency (0.01 ≤ MAF < 0.05)		Common (MAF ≥ 0.05)	
panel	*n* variants†	FP (%)*	*n* variants†	FP (%)*	*n* variants†	FP (%)*	*n* variants†	FP (%)*
1000G-Phase1	9,753,072	7,063 (0.072)	1,374,914	211 (0.015)	2,176,785	1,409 (0.065)	6,201,373	5,443 (0.088)
UK10K	10,343,027	1,928 (0.019)	2,404,171	76 (0.003)	2,206,860	590 (0.027)	5,731,996	1,262 (0.022)
1000G-Phase3	11,180,215	6,327 (0.057)	2,307,409	316 (0.014)	2,319,353	1,116 (0.048)	6,553,453	4,895 (0.075)
1000G-Phase1 UK10K	11,567,853	7,754 (0.067)	2,674,666	1,024 (0.038)	2,429,227	1,787 (0.074)	6,463,960	4,943 (0.076)
1000G-Phase3 UK10K	11,965,521	5,884 (0.049)	3,088,939	742 (0.024)	2,406,422	1,277 (0.053)	6,470,160	3,865 (0.060)

Number of genetic variants with IMPUTE2-*info* score ≥ 0.7 (n). In parenthesis, for the IL GT and Affy GT SNPs.

*Number of false positive associations for each set of variants considering a genome-wide significance of $5x10^{-8}$. In parenthesis, the proportion of false positive signals

Moreover, we generated a set of imputed variants integrating the contributions of the 1000G-Phase1 and UK10K on the basis of maximizing the amount of well-imputed variants by choosing, for each variant, the reference panel that provides the best IMPUTE2-*info* score. This approach led to an increase in the coverage of the imputed results based on 1000G-Phase1 of 18.61% and with respect to the results from using the UK10K reference panel, of 11.84%. However, when combining UK10K and 1000G-Phase1, we cannot achieve the minimal occurrence of spurious associations observed from using UK10K reference panel alone. However, as shown in Table 2, there is a reduction in the ratio of total false positives associations of a 6.90% through combining 1000G-Phase1 and UK10K contributions in comparison with the ratio obtained from using solely the 1000G-Phase1 imputed results. As exposed in Table 3, we clearly saw that this improvement came from variants at the range of common allele frequencies but a large number of artefacts were additionally created for lower allele frequencies when using different reference panels for cases and controls. We performed the same analysis but in this case, by integrating the imputed results from 1000G-Phase3 and UK10K. Interestingly, the occurrence of FP associations was extensively minimized, which led to a reduction of 26.61% in the rate of FP between the two integrated panels. For rare and low-frequency variants, the decrease in the rate of FP was 36.79% and 39.45%, respectively. This new set of results emphasized the notable improvement of the Phase3 release with respect to the Phase1 release of the 1000G, which was particularly relevant at lower allele frequencies.

As a second line of evidence, we evaluated the statistical power of genotype imputation across different ranges of allele frequencies, measured by the percentage of Affy and IL GT SNPs with imputed mean R^2 above 0.5. As shown in Figure 21, the curve for the 1000G-Phase1+UK10K imputed results outperforms 1000G-Phase1 and UK10K, even slightly, at any range of allele frequency, specially benefiting from 1000G-Phase1 for highly common variants. Actually, this integrated dataset was comparable to the 1000G-Phase3 in terms of coverage across the whole spectrum of allele frequency. Thereafter, the most accurate performance was obtained by merging 1000G-Phase3 and the UK10K reference panels, which maximizes the coverage at any range of allele frequency, but particularly at the low-frequency range.

These analyses taught us that by combining the imputed results from several reference panels we are able to extensively gain in genomic resolution at any range of allele frequency, as well as in the degree of genotype imputation accuracy. Of note, the 1000G-Phase3 release brought consistency to our merging strategy by minimizing the occurrence of FP associations.

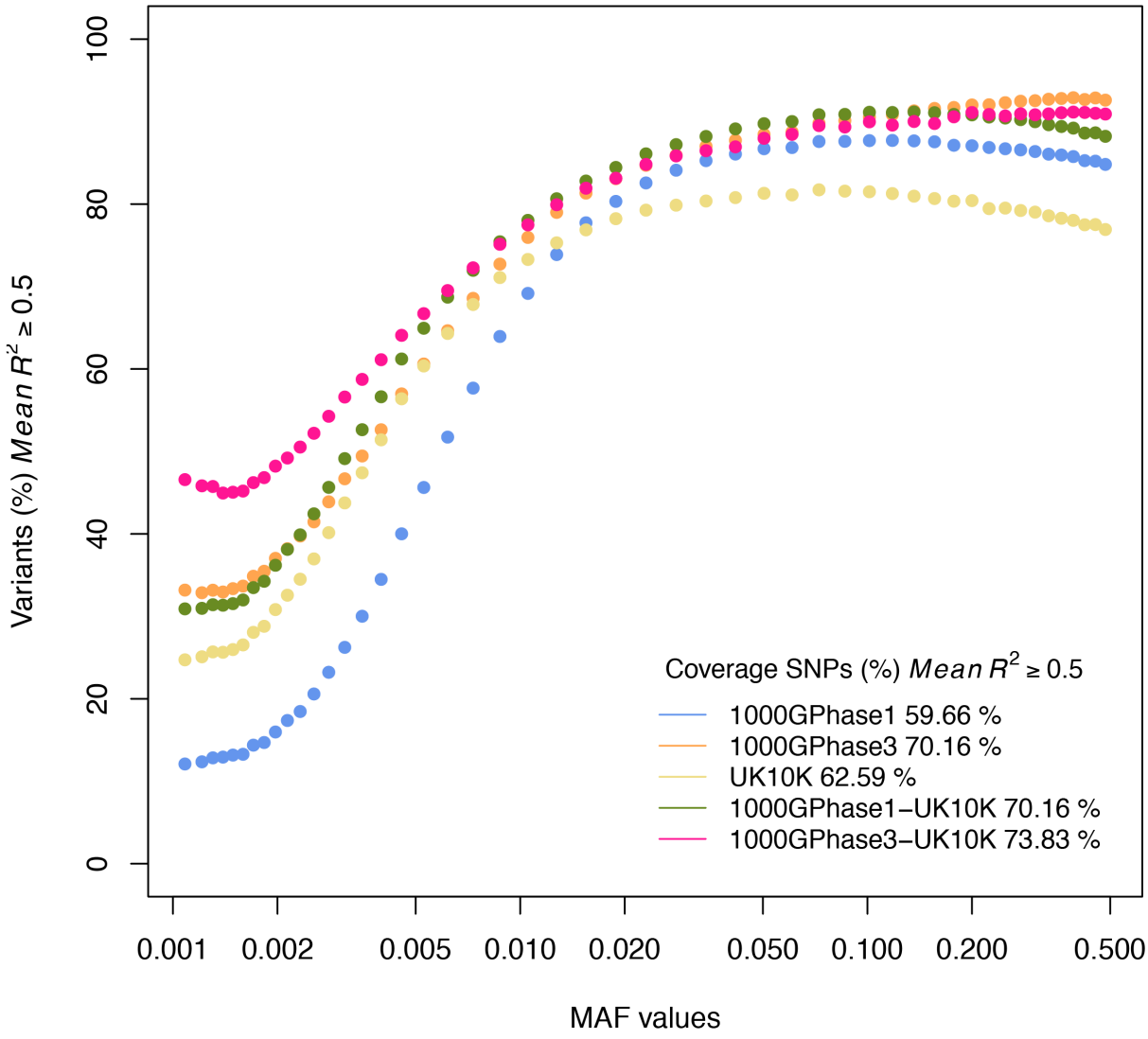


Figure 21. Genome-wide representation of power across different MAF ranges. The y-axis, represents the percentage of SNPs with a mean imputed $R^2 \geq 0.5$. The x-axis represents different MAF values. Of note, we did not include the chromosome 21 because the UK10K release lacked a substantial fraction of this chromosome at least in the initial release that we had at our disposal. The reference panels evaluated were 1000G-Phase1 (blue), UK10K (yellow), 1000G-Phase3 (red), 1000G-Phase1-UK10K (green) and 1000G-Phase3-UK10K (fuchsia).

1.2.2 Preventing the occurrence of spurious association from errors in genotyping

We hypothesized that wrongly genotyped (FP) variants could highly contribute to the accuracy of imputation. In order to find a method to correct these spurious associations, we tested whether the *-pgs* option from IMPUTE2, which imputes the already genotyped variants based on the surrounding SNPs, could be of use to identify these wrongly genotyped variants. The strategy is based on comparing the imputed results with the real genotyped results. Several metrics were evaluated in order to select the one with the greatest discriminator power. The best metric was *diffBeta*, defined as the absolute difference between 1 and the slope comparing the real and imputed genotypes (see methods, Figure 11 (A)). Note that we expect a perfect correlation and a slope of 1 when there is complete agreement between the imputed and the real results. Using the NBS genotyped data, we set a *diffBeta* cut-off of 0.10, which is able to filter out 87.1% of all false positive associations, while still retaining 78.3% of good quality SNPs (SNPs that do not show GW significant when comparing Affy vs IL).

We used the 58C cohort to test how this method was able to eliminate 46 genomic regions imputed on chromosome 1 that showed at least one genome-wide significant SNP (i.e. a false positive due to a genotype imputation or a genotyping artefact). For most of these regions, the genome-wide significant SNP was an SNP imputed from both platforms. We eliminated all the SNPs with *diffBeta* higher than 0.10, as this was able to eliminate 87.1% of biased SNPs, as trained in the NBS cohort. After eliminating all the SNPs with *diffBeta* higher than 0.10, as trained in the NBS cohort, we phased and imputed the genotypes again (Figure 11 (B)). We then evaluated the coverage and percentage of false positives using different IMPUTE2-*info* thresholds. We observed that in order to decrease in at least a 90% the ratio of FP observed when not applying this *–pgs* pre-filtering and IMPUTE2-*info* of 0.7, an IMPUTE2-*info* filter of 0.828 should be applied in conjunction with the *–pgs* via. This alternative pipeline resulted in a -10.28 fold-change for the ratio of FP. With respect to the number of variants retained, the approach based on the *diffBeta* with the IMPUTE2-*info* score 0.828 resulted in a loss of 19.91% in the number of variants captured in the original pipeline (no *–pgs* filtering and IMPUTE2-*info* score = 0.7) see Figure 22, (A).

In order to test this strategy in another dataset with a different genotyping array, we compared a fraction of the 58C samples that were genotyped by both the Affymetrix 500K arrays and Illumina 1.2 arrays. After imputing both datasets independently, there were 18 genomic regions with FP associations in chromosome 1. We showed that applying blindly the previously mentioned filters (*diffBeta* < 0.10 and IMPUTE2-*info* cut-off 0.828) in comparison with a standard pipeline (without –pgs pre-processing and IMPUTE2-*info* cut-off 0.7) we were able to reduce the

number of false positive to 2 out of 149, resulting in a 97.84% reduction in the ratio of FP of the standard pipeline. However, the number of retained SNPs observed for the original pipeline was decreased in a 37.79% (Figure 22 (B)).

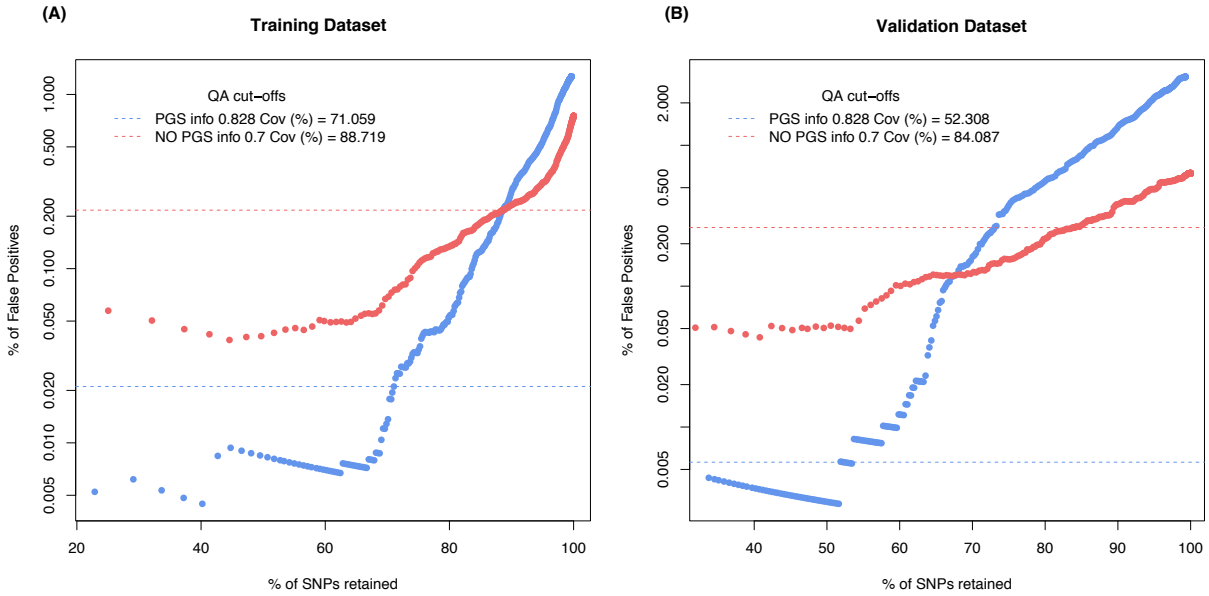


Figure 22 Evaluation of accuracy and statistical power of standard and the pre-filtering (*diffBeta*) based pipeline for genotype imputation. The x-axis corresponds to the percentage of SNPs retained with respect to the whole fraction of variants obtained by genotype imputation without filtering on the genotyped data and no post-imputation filter has been applied. On the y-axis, the percentage of FP associations is represented in the log scale. Each dot corresponds to an IMPUTE2-*info* threshold used to filter the imputed results from Affy and IL when using the standard pipeline (red) and the *diffBeta* pipeline, which previously eliminates the wrongly genotyped SNPs (blue). Dashed lines represent the percentage of FP when only filtering the imputed results with and IMPUTE2-*info* cut-off of 0.7 (red) and when filtering genotyped data with the *diffBeta* parameter previous to genotype imputation and with a post-imputation IMPUTE2-*info* of 0.828 (blue). (A) Analysis based on 46 regions on chromosome 1 with at least one FP association when comparing the imputed results of the 58C cohort from Affy and IL. (B) The analysis was based on the 18 genomics regions on chromosome 1 that showed at least one FP association when comparing the imputed results of the 58C cohort from Affy500K and IL.

1.2.3 Exploring the impact of genotype imputation in meta-analysis approaches

We compared the NBS (labeled as controls) against 58C (labeled as cases) samples to obtain two independent GWAS results, as these datasets were genotyped by both Affy and IL genotyping platforms (Figure 12). We did not expect major differences between NBS and 58C, as they are both sets of Caucasian controls. However, we expected some real differences due to slight population genetic differences. We meta-analysed the results of NBSvs58C genotyped and imputed from Affy with the results of NBSvs58C genotyped by IL. We used the I^2 score as a measure of heterogeneity, as suggested in the literature (Higgins and Thompson 2002; Evangelou and Ioannidis 2013).

We found that when imputed variants with low IMPUTE2-*info* score are included in a meta-analysis, the mean heterogeneity is much higher than when using IMPUTE2-*info* score 0.7 to filter out inaccurate imputed SNPs (mean I^2 when IMPUTE2-*info* score < 0.7 = 6.142; mean I^2 for IMPUTE2-*info* score ≥ 0.7 = 0.155; Wilcoxon-test p-value < $2.2x10^{-16}$). In addition, the percentage of SNPs with high heterogeneity (I^2 higher than 50) increased as the IMPUTE2-*info* cut-off decreased. For example, a relaxed filtering (IMPUTE2-*info* ≥ 0.3) led to 193,108 variants (1.5%) with I^2 higher than 50, while only 7,128 variants (0.08%) had I^2 higher than 50 considering an IMPUTE2-*info* cut-off of 0.7 (Figure 23). These results suggest that, setting as missing a given variant only in those cohorts where that variant has an IMPUTE2-*info* score < 0.7, is a more powerful approach than eliminating all the variants that show high heterogeneity without a previous filtering based on IMPUTE2-*info* score. This was consistent across different ranges of allele frequencies as reported in Figure 24. All the results showed correspond to a sample-size meta-analysis but we realized the same evaluation with the inverse variance fixed effects meta-analysis, providing equivalent performances (Supplementary Material 1, Supplementary Material 2).

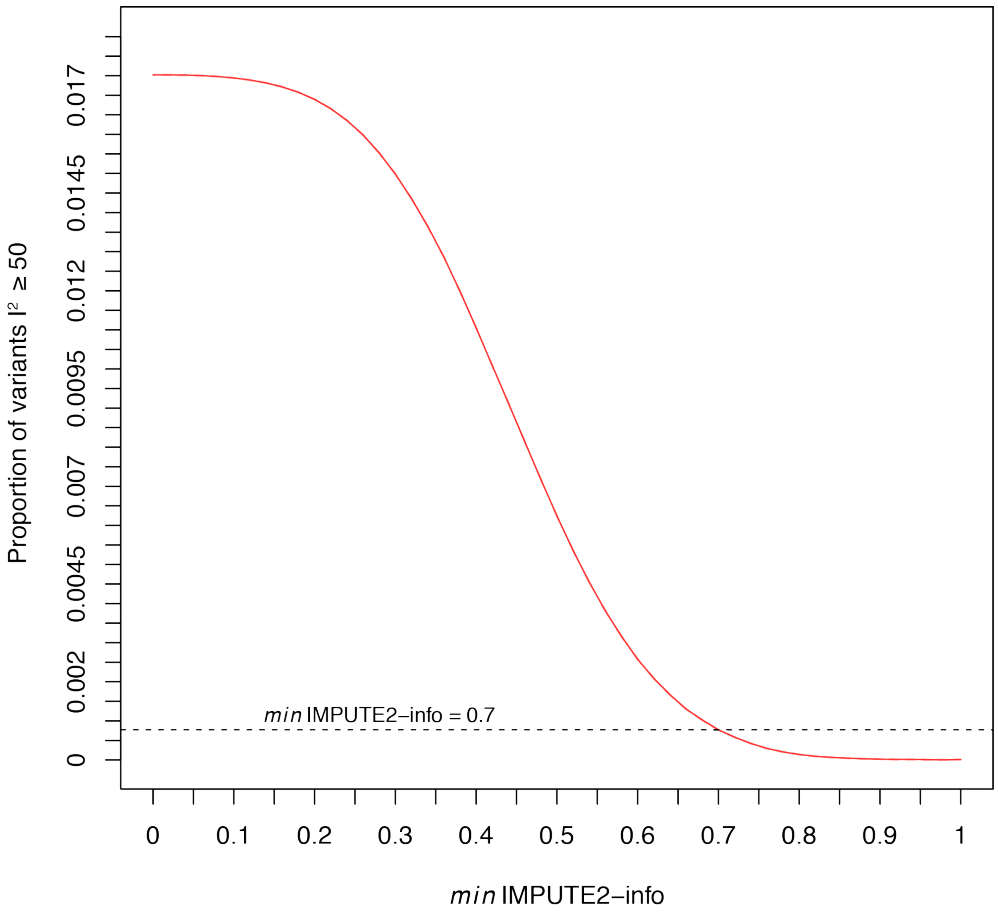


Figure 23 Proportion of high heterogeneity SNPs (I^2 ≥ 50, y-axis), across different IMPUTE2-*info* score cut-offs (x-axis)

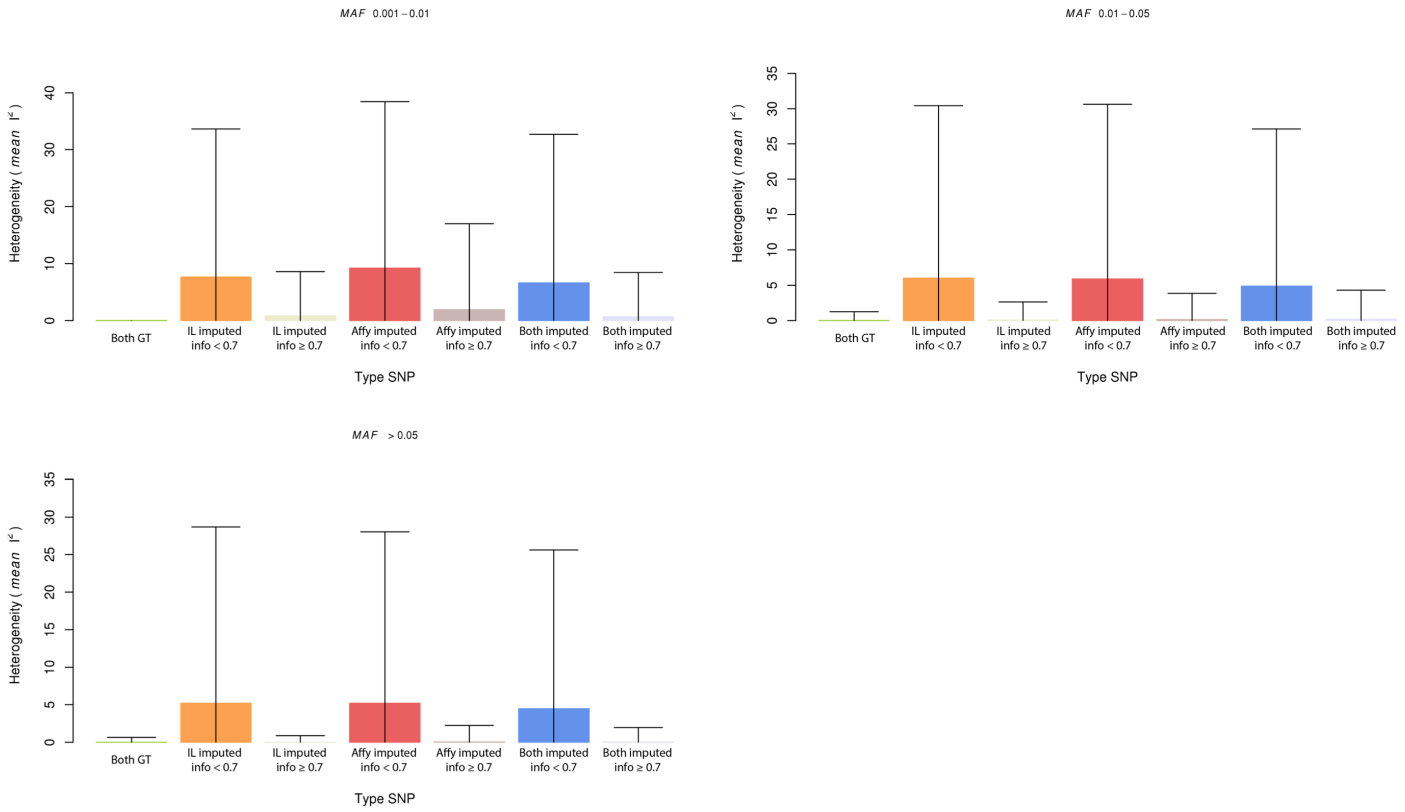


Figure 23 Comparison of the levels of heterogeneity using different post-imputation QCs across different ranges of allele frequency. Bars represent mean heterogeneity (I^2) across different allele frequency ranges after filtering for IMPUTE2-*info* ≥ 0.7. Different groups of variants are analysed: Both GT (variants genotyped by both platforms), IL imputed (variants genotyped by Affymetrix and not in Illumina), Affy imputed (variants genotyped by Illumina and imputed by Affymetrix) and Both imputed (variants imputed for both sets of data)

2. Novel insights of the genetic architecture of T2D: crossing the boundaries of common variants

Bonàs-Guarch S, Guindo-Martínez M, Miguel-Escalada I, Garup N, Sebastian D, Rodriguez-Fos E, Sánchez F, Planas-Fèlix M, Cortés-Sánchez P, Morgan CC, Moran I, González JR, Andersson E, Díaz C, Badia RM, Udler M, Flannick J, Jorgensen T, Linneberg A, Jorgensen ME, Witte DR, Christensen C, Brandslund I, Appel EV, Scott R, Luan J, Sigma Type 2 Diabetes Consortium, The InterAct Consortium, Pedersen O, Zorzano A, Flórez JC, Hansen T, Ferrer J, Mercader JM, Torrents D. Sequencing-based imputation and reanalysis of 70,000 individuals from publicly available datasets reveals novel *loci* associated with type 2 diabetes. *In preparation*.

Contribution of the PhD candidate

- Quality control of the genotyped data.
- Genotype imputation with two sequence-based reference panels following the new quality-control guidelines for imputed data generated in the previous section.
- Fine-mapping with the 99% credible set of variants approach the associated *loci*.
- Functional annotation of coding variants.
- Costumed statistical analysis of the novel rare X-chromosome variant.
- Involvement in the manuscript preparation.

2.1 Overall analysis strategy

As shown in Figure 25, we first obtained all T2D case-control GWAS individual-level data that was available through EGA and dbGaP databases. To harmonize the data of all these cohorts, we developed a pipeline to standardize the quality control and filtering of low-quality variants and samples (see Methods). After this process, we were able to gather a total of 70,127 subjects (70KforT2D; 12,931 cases and 57,196 controls; Supplementary Material 3). Each of the cohorts was then imputed to 1000G-Phase1 and UK10K reference panels. To perform the imputation and association analysis we developed an integrated tool that performs phasing, genotype imputation and association testing by exploiting the internal parallelism of the multiple tasks involved (under review) (Sanchez et al. 2016).

Following this strategy (see Figure 25), a total of 15,115,281 variants with good imputation quality were tested for association in a total of 12,931 T2D cases and 57,196 controls (IMPUTE2-*info* score≥0.7, MAF≥0.001 and I^2 heterogeneity score<0.75). Of these, 6,845,408 variants were common (MAF≥0.05), 3,100,848 variants were low-frequency (0.01≤MAF<0.05) and 5,169,025 variants were rare (0.001≤MAF<0.01). Interestingly, merging the association results from both the UK10K and the 1000G-Phase1 reference panels substantially improved the number of high-confidence imputed SNVs and INDELs, compared to the association results obtained with each of the reference panels alone, especially in the low-frequency and rare variant spectrum. For example, a total of 5,169,025 high-confidence imputed rare variants resulted from combining 1000G-Phase1 and UK10K results, while only 2,878,263 and 4,066,210 rare variants were imputed with 1000G-Phase1 and UK10K respectively (Figure 26A). The combination of the results from both reference panels also allowed the imputation of 1,357,753 high-confidence INDELs (Figure 26B).

We used three main meta-analytic strategies to take the greatest advantage of all publicly available data, including summary statistics, individual-level genotype datasets, and individual queries possible through the T2D Portal (http://www.type2diabetesgenetics.org/). First, we meta-analyzed all summary statistics results from the DIAGRAM trans-ethnic meta-analysis (DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium et al. 2014) (26,488 cases and 83,964 controls) consisting mostly of HapMap variants (1,918,233 variants), with the fraction of the publicly available cohorts, obtained through dbGaP and EGA, that had no overlap with the DIAGRAM dataset (i.e. GERA cohort and the NuGENE, 7,522 cases and 50,446 controls) (Figure 25, Supplementary Material 3). Second, the rest of the variants, which were not common (0.001≤MAF<0.05) or not tested in DIAGRAM, were further meta-analyzed using all the cohorts for whom individual-level data were available in dbGaP and EGA (12,931 cases and 57,196 controls, 13,197,048 variants; 70KforT2D resource). Finally, low-frequency coding variants with p-value≤$1x10^{-4}$ were meta-analyzed using the non-overlapping fraction of samples with the data from the T2D Portal through the interrogation of exome array data

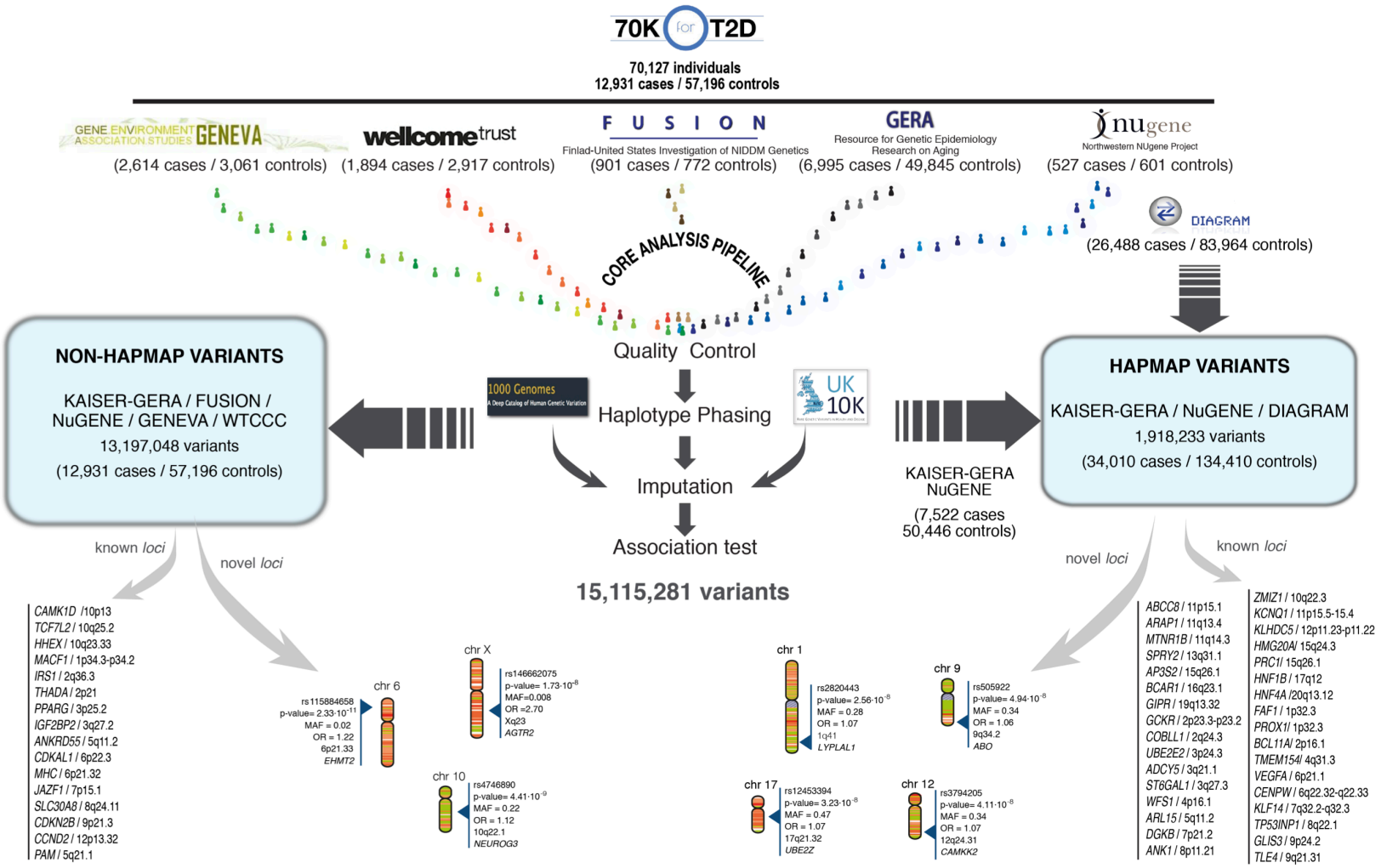


Figure 25. Discovery and replication strategy. Publicly available GWAS datasets representing a total of 12,931 cases and 57,196 controls (70KforT2D) were first quality controlled, phased and imputed using 1000G-Phase1 and UK10K separately. For those variants that were present in the DIAGRAM trans-ethnic meta-analysis we used the summary statistics to meta-analyze our results with the cohorts that had no overlap with any of the cohorts included in the trans-ethnic meta-analysis. With this first meta-analysis we discovered five novel *loci*. For the rest of the variants we meta-analyzed all the 70KforT2D datasets, which resulted in two novel *loci*. All the variants that were coding and that showed a p-value ≤ $1x10^{-4}$ were tested for replication by interrogating the summary statistics in the Type 2 Diabetes Genetics portal (http://www.type2diabetesgenetics.org/). This resulted in a novel low-frequency variant in the *EHMT2* gene.

from ~80,000 individuals and ~17,000 individuals that were whole-exome sequenced (Flannick et al. 2014; Mahajan et al. 2015; Fuchsberger et al. 2016).

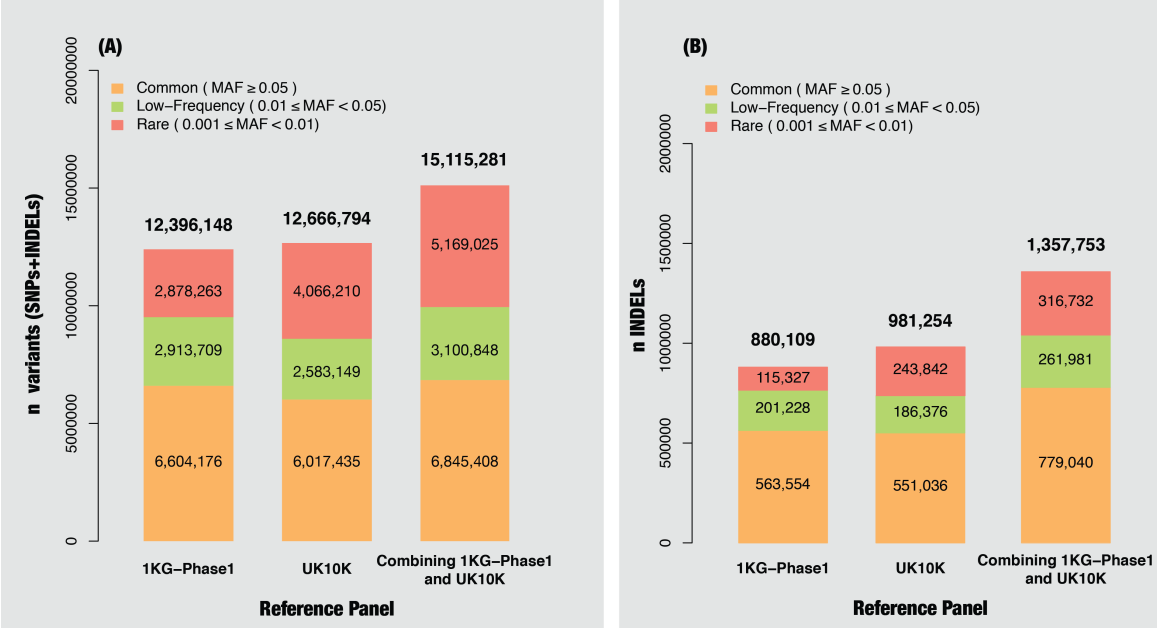


Figure 26. Description of the genomic coverage from genotype imputation attained after combining multiple reference panels compared to only using a single reference panel. A) All variants. B) INDELs and large deletions. Each bar represents the genomic coverage from the final meta-analysis for the 70KforT2D cohort according to the reference panel used: from left to right, (1) 1000G-Phase1 release, (2) UK10K and (3) when combining the best-guessed variants from 1000G-Phase1 and UK10K reference panels. Each bar was stratified according to the range of allele frequency: rare variants (0.001 ≤ MAF < 0.01) in blue, low frequency variants (0.01 ≤ MAF < 0.05) and common variants (MAF ≥ 0.05) in red. Y-axis shows the absolute number of variants that passed all post-imputation quality filters, including IMPUTE2-*info* score ≥ 0.7.

2.2 Pathway analysis

As a first exploration of how our association results recapitulate the pathophysiology of T2D we performed gene-set enrichment analysis with DEPICT, using all the variants with p-value≤1x10^{-5} as input. DEPICT has been successfully used to identify whether genes in associated GWAS *loci* are enriched for tissue-specific expression and reconstituted gene sets (modified versions of canonical gene sets) (Pers et al. 2015) (see Methods). This analysis showed enrichment of genes expressed in pancreas (ranked 1st, p-value=7.8x10^{-4}, FDR<0.05, Supplementary Material 4) and cellular response to insulin stimulus (ranked 2nd, p-value=3.9x10^{-8}, FDR=0.05, Figure 27, Supplementary Material 5), in concordance with the current knowledge of the pathophysiology of T2D.

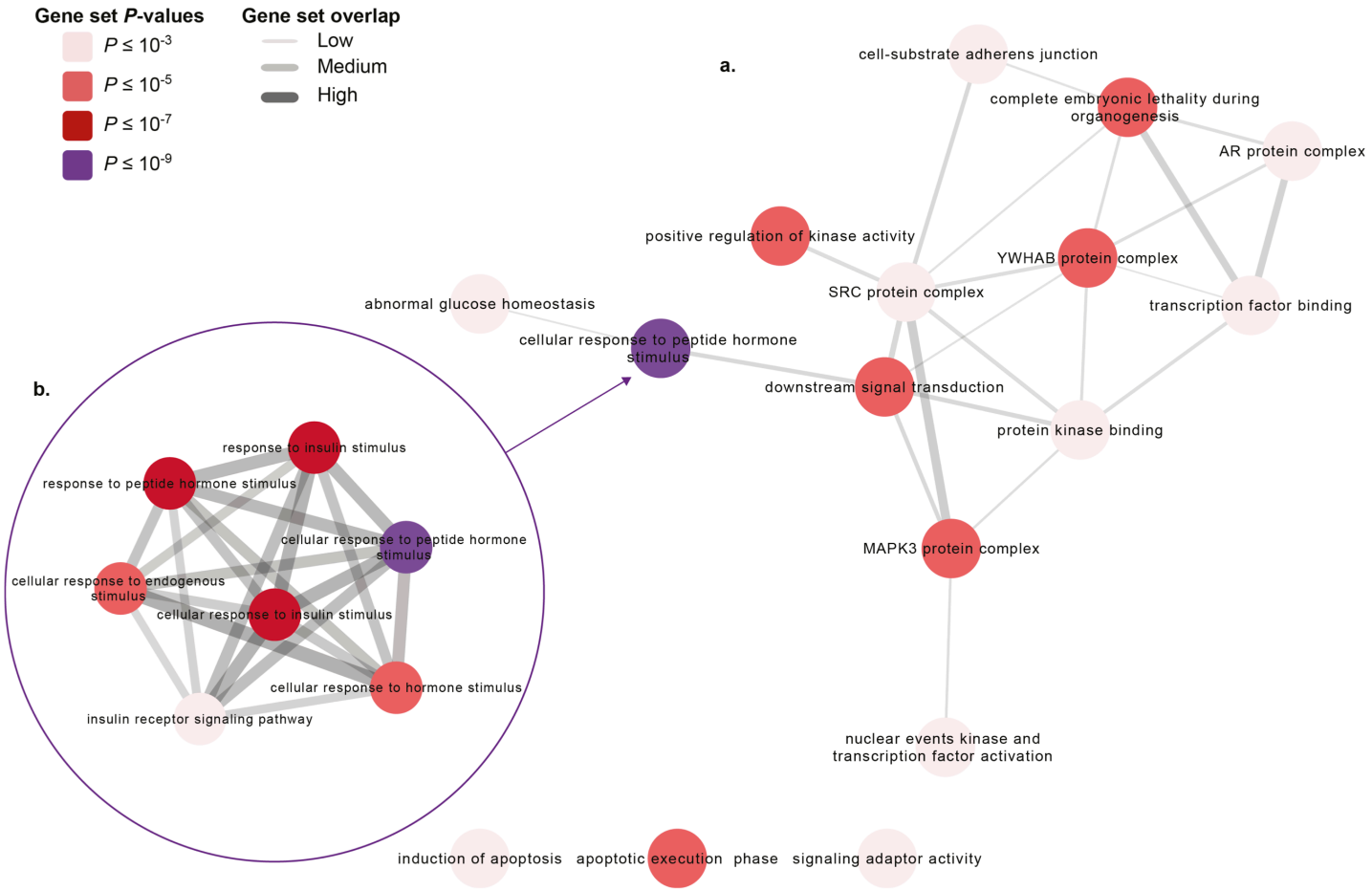


Figure 27. Network plot representing all pathway clusters that were significantly enriched cluster pathways (FDR<5%) using DEPICT. As input, we used all summary statistics with p-value≤1x10^{-5} in the 70KforT2D meta-analysis. Significantly enriched pathways were clustered by merging all pathways showing correlation higher than 0.3 into a single cluster using the affinity Propagation tool. The dependency between clusters is represented by the width of the edges. An expanded version of the "cellular response to peptide hormone stimulus" cluster is represented, showing all the pathways that showed enrichment within this cluster and their dependencies.

2.3 Identification, fine-mapping and functional characterization of novel and previously known *loci*

Following the homogenization, the imputation with 1000G-Phase1 and UK10K reference panels, and the meta-analysis of all the publicly available datasets, we identified 56 genome-wide significant associated *loci* (p-value≤$5x10^{-8}$). Among them, 7 *loci* were not previously reported as associated with T2D (Table 4). The remaining 49 *loci* were already known, and included the two recently identified low-frequency variants in Europeans, a *CCND2* intronic variant and a missense variant in *PAM* (Steinthorsdottir et al. 2014).

As a quality control of our dataset, we confirmed that the magnitude and direction of effect of all the associated variants (with p-value≤0.001) were highly consistent with those reported previously (Rho=0.92, p-value=$1x10^{-248}$, Figure 28).

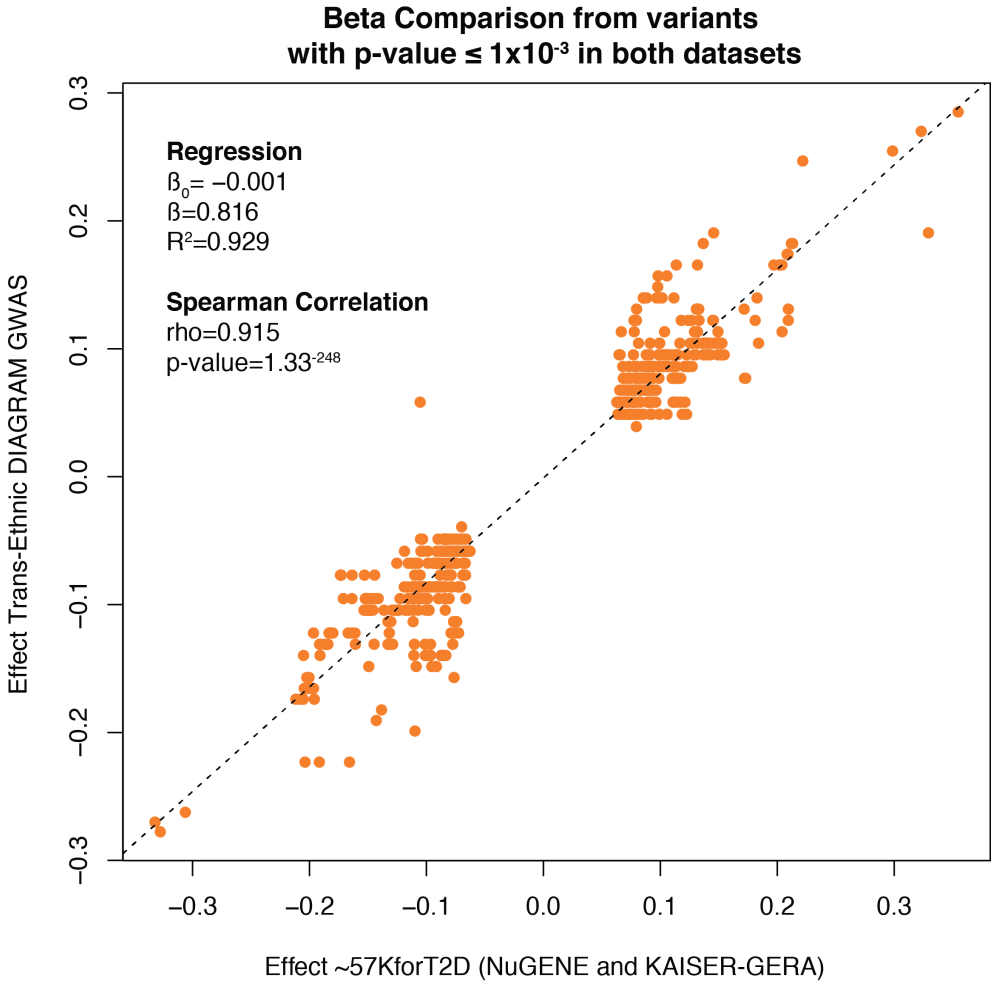


Figure 28. Comparison of effect sizes in the non-overlapping cohorts from the 70KforT2D meta-analysis and previously published results from DIAGRAM trans-ethnic meta-analysis. Each dot corresponds to a previously reported risk variant and its corresponding log-odds ratio. The analysis comprised all variants with p-value ≤ $1x10^{-3}$ in both datasets.

Table 4. Novel T2D associated *loci*.

Novel Locus	Chr	rsID – Risk Allele	OR (95% CI) P-value			MAF
			Stage1 Discovery meta-analysis	Stage2 Replication meta-analysis	Stage1 + Stage2 Combined meta-analysis	
LYPLAL1/ZC3H11B (1q41)	1	rs2820443-T	1.08 (1.04-1.13) $2.94x10^{-4}$ †	1.06 (1.03-1.09) $2.10x10^{-5}$ §	1.07 (1.04-1.09) $2.56x10^{-8*}$	0.28
EHMT2 (6p21.33)	6	rs115884658-A	1.37 (1.21-1.54) $3.57x10^{-7}$ ¥	1.17 (1.09-1.26) $2.90x10^{-6}$ ‡*	1.22 (1.15-1.30) $2.33x10^{-11*}$	0.02
ABO (9q34.2)	9	rs505922-C	1.07 (1.03-1.11) $6.93x10^{-4}$ †	1.06 (1.03-1.09) $1.90x10^{-5}$ §	1.06 (1.04-1.09) $4.94x10^{-8*}$	0.34
NEUROG3 (10q22.1)	10	rs4746890-T	1.12 (1.08-1.16) $4.41x10^{-9}$ ¢	-	-	0.22
CAMKK2 (12q24.31)	12	rs3794205-G	1.09 (1.05-1.14) $4.18x10^{-5}$ †	1.06 (1.03-1.09) $1.60x10^{-4}$ §	1.07 (1.04-1.10) $4.11x10^{-8*}$	0.34
CALCOCO2/ATP5G1/ UBE2Z/SNF8/GIP (17q21.32)	17	rs12453394-A	1.08 (1.04-1.12) $7.86x10^{-5}$ †	1.07 (1.03-1.11) $9.60x10^{-5}$ §	1.07 (1.05-1.10) $3.23x10^{-8*}$	0.47
AGTR2 (Xq23)	X	rs146662075-T	3.09 (2.06-4.60) $3.24x10^{-8}$ Δ	1.82 (0.92-3.61) 0.09 Ø	2.70 (1.91-3.81) $1.73x10^{-8}$	0.008

Abbreviations are as follows: Chr, Chromosome; OR, odds ratio; MAF, minor allele frequency.

† Imputed based public GWAS discovery meta-analysis (NuGENE + GERA cohort, 7,522 cases and 50,446 controls)

¥ Imputed based public GWAS discovery meta-analysis (NuGENE + GERA cohort + GENEVA, 8,136 cases and 53,507 controls)

¢ Full imputed based public GWAS meta-analysis (NuGENE + GERA cohort + GENEVA+FUSION+WTCCC, 12,931 cases and 57,196 controls)

Δ 70KforT2D Men Cohort (GERA cohort + GENEVA + FUSION, 5,277 cases and 15,702 controls older than 55 years old)

‡ T2D Diabetes Genetic Portal (Exome-Chip+Exome Sequencing, 35,789 cases and 56,738 controls)

§ Transancestry DIAGRAM Consortium (26,488 cases and 83,964 controls)

Ø Replication Men Cohort SIGMA UK10K imputation + InterAct + Danish Cohort (case-control and follow-up) (9,529 cases and 4,664 controls older than 55 years old and OGTT>7.8 mmol/l, when available).

*Meta P-value Estimated using a weighted Z-score method due to unavailable SE information from the Stage 2 replication cohorts.

This study also allowed us to fine-map known and novel *loci* and therefore provides a more accurate functional annotation of the associated *loci* to propose candidate causal variants. To do this, we first constructed the 99% credible sets (i.e. the subset of variants that have, in aggregate, 99% probability of containing the true causal variant) (Wellcome Trust Case Control et al. 2012) of all the *loci* that contained at least one GWAS significant variant (p-value=$5x10^{-8}$).

The fine-mapping approach increased the resolution of previously and newly identified *loci*, with special improvement in the identification of structural variants. We observed that our credible sets for all the *loci* contained a total of 8,305 variants, of which 922 were INDELs. Of all these INDELs, 105 were genome-wide significant. 672 out of the 922 INDELs in the credible sets were confirmed to be present in 1000G-Phase3 release. Of these 672, 188 and 90 INDELs were only present in the UK10K and 1000G-Phase1 reference panels, respectively, while only 394 INDELs were identified by both UK10K and 1000G-Phase1. This analysis emphasizes the advantage of combining the results from several reference panels. Overall, the INDELs represent 11.1% of the variants within our 99% credible sets. In fact, for 15 of all the 71 known *loci* that we were able to replicate (p-value≤$5.3x10^{-4}$, to correct for multiple testing) we found that the top variant was a previously undescribed INDEL, suggesting that other types of variation rather than SNVs may have a substantial role in the susceptibility for T2D. A valuable example of the possible role of INDELs is the fine-mapping of the well-known associated region within the *IGF2BP2* intron. Although this is a well-established and functionally validated *locus* (Diabetes Genetics Initiative of Broad Institute of et al. 2007; Dai et al. 2015), the causal variant for this gene has not been identified yet. In this *locus*, 12 of the 57 variants within the 99% credible set were INDELs, showed a collective posterior probability of 18.4%, and all of them were genome-wide significant ($5.6x10^{-16}$<p-value<$2.4x10^{-15}$) and described here for the first time. Notably, a common 10 base-pair deletion that was not found in 1000G-Phase1 (rs755826890, OR=1.14, p-value=$1.13x10^{-15}$) falls within a curated regulatory element (OREG1275562) (Portales-Casamar et al. 2009), suggesting a potential regulatory mechanism that contributes to increase the susceptibility for T2D in the *IGF2BP2 locus*. These results therefore highlight the potential of our strategy to discover associated INDELs, which represent additional candidate causal variants for known and novel *loci.*

To identify or prioritize the causal variants, effector transcripts, or tissues underlying novel and previously known *loci*, we performed a detailed annotation analysis. To this end, we analyzed all the credible sets using the variant effector predictor (VEP), a tool that provides functional annotation of coding variants (McLaren et al. 2010), and extended the functional annotation to non-coding variants with Combined Annotation Dependent Depletion (CADD) scores (Kircher et al. 2014). Additionally, we tested the effect of all variants on expression across diverse tissues by interrogating GTEx (GTEx Consortium 2013; Carithers and Moore 2015; Mele et al. 2015) and RNA-sequencing gene expression data from pancreatic islets (Fadista et al. 2014). For example, in the *MACF1* region, a

detailed characterization of the region resulted in several causal variants that should be further tested. First, there were three missense variants within the 99% credible set, which could collectively influence protein function. Furthermore, we found a robust association between a large number of variants in the credible set with expression of several nearby genes, including *PABPC4, OXCT2P1,* but also *MACF1*.

The analysis of pancreatic islet expression datasets (Fadista et al. 2014) showed that in multiple known T2D associated variants were eQTLs for nearby genes. For example, we showed that variants in the 15q26.1 region are associated with expression of the *AP3S2-C15orf38* read-through transcript (rs7111, R^2 with top variant=0.93, p-value=$1.1x10^{-14}$).

2.4 Identification, fine-mapping and functional characterization of novel and previously known *loci*

Besides providing a comprehensive characterization of known T2D associated regions, we also identified 5 novel *loci* driven by common variants, which have modest effect sizes (Table 4, Figure 29, Supplementary Material 6 and 7) that could be relevant for a better understanding of the biology of T2D. A comprehensive genetic and functional characterization of each of these novel *loci* based on the methodology explained above and extensive literature search is described below.

Within the novel T2D-associated *locus* in chromosome 1q41 (*LYPLAL1-ZC3H11B*, rs2820443, OR=1.07 [1.04-1.09], p-value=$2.56x10^{-8}$), several variants in this region have been previously associated with waist-to-hip ratio in women, height, visceral adipose fat in women, adiponectin levels, and fasting insulin (Dastani et al. 2012; Fox et al. 2012; Manning et al. 2012; Berndt et al. 2013; Randall et al. 2013). Among the genes captured within the credible set, *LYPLAL1*, which encodes for lysophospholypase-like 1, is downregulated in mouse models of diet induced obesity and upregulated during adipogenesis, which implicates *LYPLAL1* as a plausible effector gene (Lei et al. 2015). Among the potential causal variants in the same *locus*, rs10779358, which is in strong LD with the top variant (R^2=0.75), is associated with expression of the pseudogene *RIMKLBP2* in adipose tissue according to GTEx (rs10779358; beta=0.235; p-value=$1.77x10^{-5}$).

At the chromosome 10q22.1 (*NEUROG3/COL13A1/RPL5P26*, rs4746890, OR=1.12 [1.08-1.16], p-value=$4.41x10^{-9}$), *NEUROG3* (Neurogenin3) is an essential regulator of pancreatic endocrine cell differentiation (Gradwohl et al. 2000; Wang et al. 2006). Mutations in this gene have thus been reported to cause permanent neonatal diabetes (Rubio-Cabezas et al. 2011), but a role of this gene in T2D has not been previously reported (del Bosque-Plata et al. 2001).

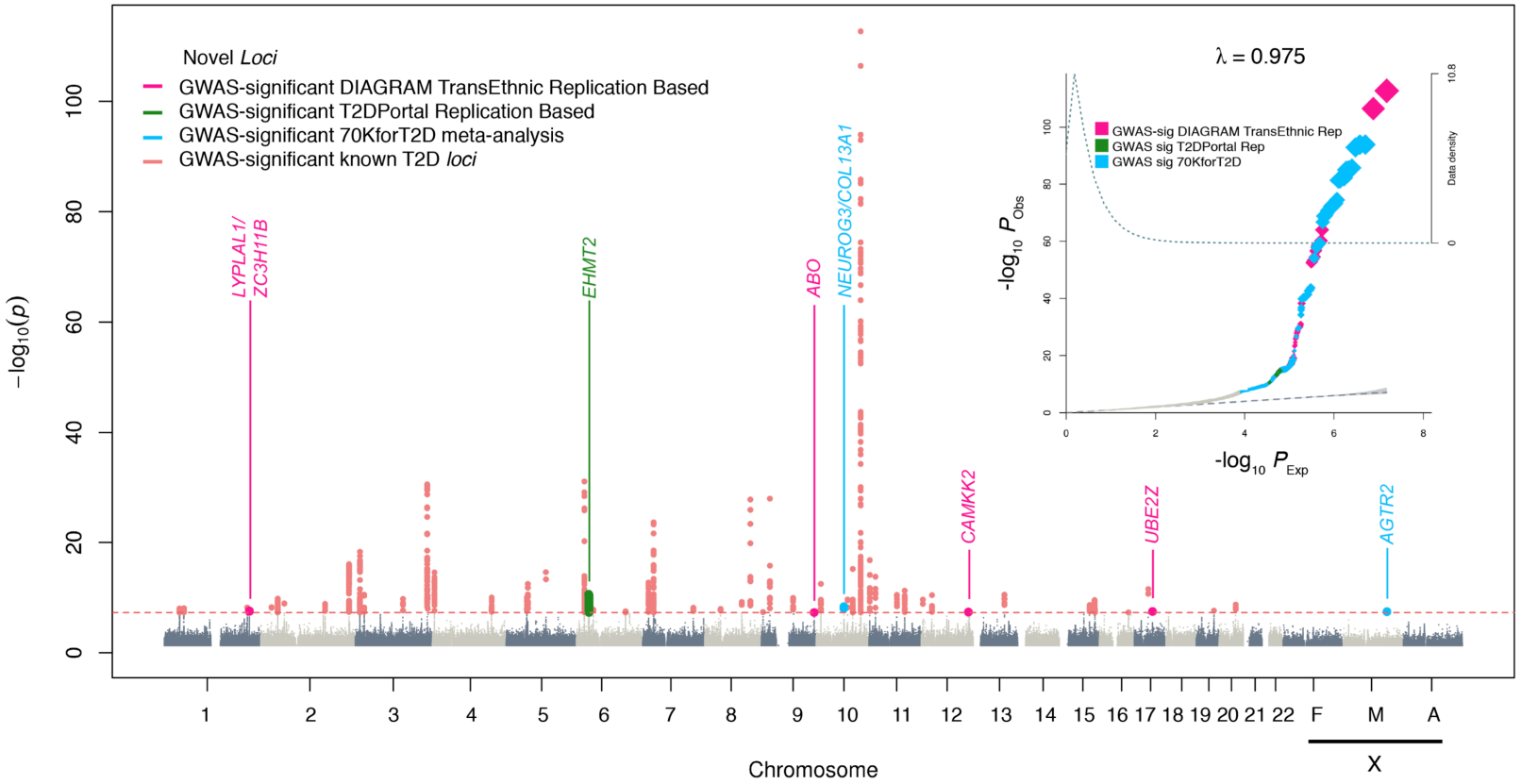


Figure 29. Manhattan and Quantile-Quantile plot (Q-Q-plot) of the discovery and replication genome-wide meta-analysis. The upper corner represents the Quantile-quantile plot. Expected $-\log_{10}$ p-values under the null hypothesis are represented in the x-axis while observed $-\log_{10}$ p-values are represented in the y-axis. Observed p-values were obtained according to the suitable replication dataset used (as shown in Figure 25) and were depicted using different colors. HapMap variants meta-analyzed using the Trans-Ethnic summary statistics from the DIAGRAM study and our meta-analysis based on the GERA cohort and the NuGENE and that resulted in novel associations are depicted in magenta. The rest of non-HapMap variants meta-analyzed using the full 70KforT2D cohort are represented in grey, highlighting in light blue the fraction of novel GWAS-significant variants. Coding low-frequency variants meta-analyzed using the 70KforT2D and the T2D Portal data that resulted in novel GWAS-significant associations are depicted in green. The shaded area of the Q-Q-plot indicates the 95% confidence interval under the null and a density function of the distribution of the p-values was plotted using a dashed line. The λ is a measure of the genomic inflation and corresponds to the observed median χ2 test statistic divided by the median expected χ^2 test statistic under the null hypothesis.

The lead variant at the chromosome 12q24.31 *locus* (rs3794205, OR=1.07 [1.04-1.10], p-value=4.11x10^{-8}) lies in the intron of *CAMKK2*, a gene suggested to be involved in cytokine induced beta cell-death (Beck et al. 2011). Interestingly, other variants within the credible set at this *locus* could be responsible for the molecular link; a missense variant within the *P2RX7* gene, that has been previously associated with glucose homeostasis humans and mice (Todd et al. 2015), as well as other variants (rs11065504, R^2 with lead variant=0.81) that are associated with the regulation of the *P2RX4* gene in tibial artery and in whole blood according to GTEx. Further fine-mapping efforts and functional studies will be needed to disentangle which is the most likely effector transcript.

The chromosome 9q34.2 *locus* (*ABO*, rs505922, OR=1.06 [1.04-1.09], p-value=4.94x10^{-8}) comprises several variants that have been previously linked to other metabolic disorders. For example, a variant in LD with rs505922 (rs651007, R^2=0.507) has been recently associated with fasting glucose (Wessel et al. 2015), whereas other SNPs, such as rs514659 (R^2 with top=1), have been associated with an increased risk for cardio-metabolic disorders (The CARDIoGRAMplusC4D Consortium 2015). One of the variants within the credible set is the 1 base-pair frame-shift deletion underlying the blood group O (Yamamoto et al. 1990). In addition, several variants within the credible set of this *locus* are associated with expression of the *ABO* gene in esophagus and blood.

Within the chromosome 17q21.32 *locus* (rs12453394, OR=1.07 [1.05-1.10], p-value=3.23x10^{-8}) three missense variants are located in *CALCOCO2*, *SNF8* and *GIP*. Variants in the glucose-dependent insulinotropic polypeptide regulatory protein (GIPR) have been previously associated with insulin response to oral glucose challenge (Saxena et al. 2010) and beta-cell function, which makes *GIP* a plausible candidate gene for this *locus* (Lyssenko et al. 2011).

2.5 Identification of a new signal driven by a low-frequency variant

To identify low-frequency coding variants associated with T2D, we meta-analysed all the coding variants with p-value≤1x10^{-4} in our meta-analysis, by meta-analysing the results from the non-overlapping 70KforT2D samples (NuGENE, GENEVA and GERA) with exome array data from ~80,000 individuals and ~17,000 individuals that were whole-exome sequenced (Flannick et al. 2014; Mahajan et al. 2015; Fuchsberger et al. 2016).

This resulted in a novel genome-wide association driven by a low-frequency missense variant (Figure 29, Supplementary Material 6 and 7) within the *EHMT2* gene at chromosome 6p21.33 (rs115884658, OR=1.22 [1.15-1.30], p-value=2.33x10^{-11}). *EHMT2* is involved in the mediation of FOXO1 translocation induced by insulin (Arai et al. 2015). Since this variant was less than 1 Mb from *HLA-DQA1*, which was recently reported to be associated with T2D (Cook and Morris 2016), we performed a series of conditional analyses to exclude that our finding was capturing previously reported T2D (Ng et al. 2014; Cook and Morris 2016) or T1D (Hakonarson et al. 2007; Wellcome Trust Case Control 2007; Barrett et al. 2009a) signals. The results show that this signal is

independent from the other previously reported variants (Supplementary Material 8). However, despite that the association at the *EHMT2 locus* was identified by replication with whole-exome sequencing datasets, other low frequency variants within the credible set of this region may also have the potential of being causal. Among them, rs115333512 (R^2 with lead variant=0.28) is associated with the expression of *CLIC1* in several tissues according to GTEx (multi-tissue Meta Analysis p-value=$8.95x10^{-16}$). In addition, this same variant is also associated with expression of the first and second exon of *CLIC1* mRNA in pancreatic islet donors (Fadista et al. 2014) (p-value(exon 1)=$1.4x10^{-19}$; p-value(exon 2)=$1.93x10^{-13}$). Interestingly, *CLIC1* has been reported as a direct target of metformin by mediating the anti-proliferative effect of this drug in human glioblastoma (Gritti et al. 2014). The results suggest *CLIC1* as a possible effector transcript, although larger sample sizes with high quality imputation or direct targeted sequencing will be needed in order to narrow the association interval at this *locus*.

2.6 Identification of a novel rare variant in the X chromosome associated with 2.7-fold increased risk for T2D

As for many other complex diseases, published large-scale T2D GWAS studies have generally excluded the analysis of the X chromosome, with the notable exception of the identification of a region near *DUSP9* in 2010 (Voight et al. 2010). To fill this gap, we aimed to thoroughly test the X chromosome genotyped and imputed variants for association with T2D. To account for heterogeneity of effects and for the differences in imputation performance between males and females, the association was stratified by sex and tested separately, as well as together. We were able to replicate the rs5945326 variant (OR=1.15, p-value=0.049). Interestingly, we identified an INDEL in high LD with the previously reported variant (R^2=0.62), which suggested a novel candidate variant for this *locus*.

We identified a genome-wide significant signal in males at the Xq23 *locus* driven by a rare variant (rs146662075, MAF=0.008, OR=2.94 [2.00-4.31], p-value=$3.52x10^{-8}$; Figure 30A). We tested the accuracy of the imputation of this variant by comparing the imputed results from the same individuals genotyped by two different platforms (see Methods) and noted that the imputation for this variant was highly accurate in men, and when using UK10K, but not in women or when using 1000G-Phase1 ($R^2_{[UK10K,males]}$=0.94; $R^2_{[UK10K,females]}$=0.66, $R^2_{[1000G,males]}$=0.62, $R^2_{[1000G,females]}$=0.46, Supplementary Material 9). Therefore, further studies will be needed in order to clarify whether this association is specific to men or if the risk is also increased in female carriers but not observed due to poorer imputation in women.

In order to confirm this association of the rs146662075 variant, we analyzed two independent cohorts by performing imputation with the UK10K reference panel (SIGMA, INTERACT) and a third cohort by *de-novo* genotyping of the rs146662075 variant in several Danish sample sets to further exclude the possibility of any genotype imputation artifact. The initial meta-analysis, once including the replication

datasets, did not result in genome-wide significance (OR=1.72, p-value=$1.6x10^{-04}$) (Figure 30B), and the results showed strong heterogeneity (heterogeneity p-value=0.002), presumably driven by the replication. We also genotyped the rs146662075 variant in a prospective study of 1,652 non-diabetic male subjects older than 45 from the Inter99 cohort who were followed-up for a median time of 11 years, of whom 158 developed T2D, and found replication of our initial findings. In agreement with the initial discovery results, the carriers of the rare T risk allele were more likely to develop diabetes during this period, compared to the C carriers (Cox-proportional Hazards Ratio (HR)=3.17 [1.3-7.7], p-value=0.011, Figure 30C).

In order to explore the discrepancy between the replication results in the three case-control studies as compared to the Inter99 prospective study and in an attempt to explain the observed heterogeneity in the meta-analysis, we compared basic characteristics of the study subjects. In fact, we found that INTERACT, SIGMA and the Danish cohort replication datasets contained controls who tended to be younger than the average age at onset of T2D recorded in cases; this was particularly true for the Danish cohort (age controls [95%CI]=46.9 [46.6-47.2]) and INTERACT (age controls [95%CI]=51.7 [51.4-52.1]) (Supplementary Material 10). For this reason, we repeated the meta-analysis using a stricter definition of controls for both the discovery and replication datasets, including only controls that were older than 55 years old. While this analysis did not result in genome-wide significant results, we performed a more strict analysis including only controls who were older than 55 years old, and with measured 2 hours plasma glucose during an oral glucose tolerance test (OGTT) below 7.8 mmol/l in the Danish cohort to further ensure the absence of pre-diabetes cases in our set of controls. OGTT is employed to diagnose T2D as well as impaired glucose tolerance, which is a strong risk factor of developing T2D (Bartoli et al. 2011). This strict definition of controls was only possible in the Danish study, as OGTT was not available for other datasets. In order to meta-analyse all the case-control studies, including the Cox-proportional hazards results, we used a meta-analysis method that accounts for overlapping subjects (MAOS) (Lin and Sullivan 2009), since there were samples who were included in both the longitudinal (follow-up) and the case-control study. The overall meta-analysis resulted in genome-wide significant results and no significant heterogeneity (OR=2.7 (1.91, 3.81), p-value=$1.73x10^{-08}$, p-value het=0.51, Figure 30D). These results therefore indicate the existence of a genetic association with T2D in elder male subjects that is driven by a rare variant.

The lead SNP rs146662075 is located near *AGTR2*, encoding for the angiotensin II receptor type 2. Several lines of evidence have implicated *AGTR2* in insulin action (Kim et al. 2006; Shum et al. 2013; Underwood and Adler 2013). Although the exact role that *AGTR2* exerts in the insulin function is still unclear, a previous study reported that the deletion of *Agtr2* protects from diet-induced insulin resistance in mice (Yvan-Charvet et al. 2005).

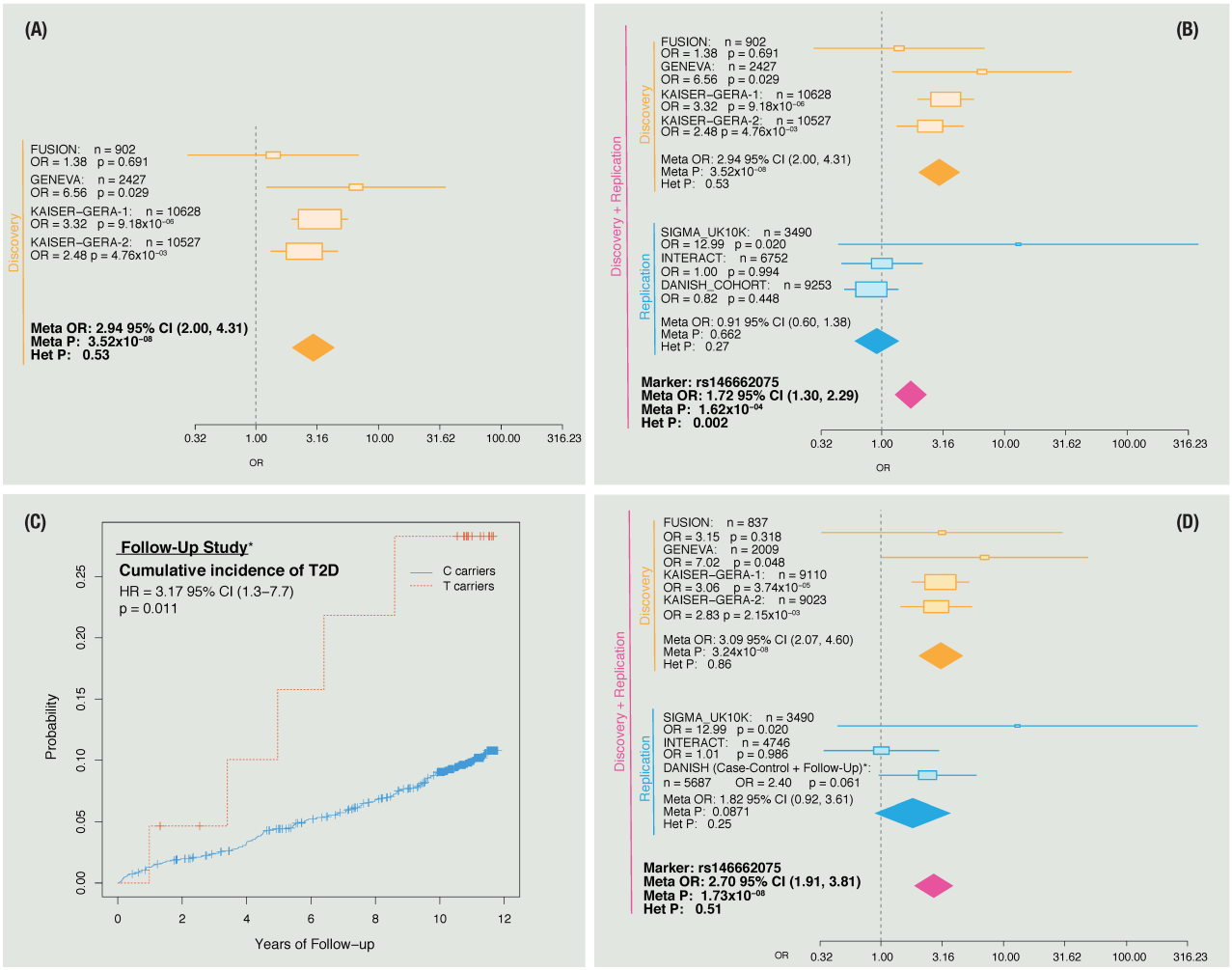


Figure 30. Discovery and replication of rs14666075 association signal. Forest plots for rs146662075 using data from the discovery and replication datasets. Cohort-specific odds ratios (95% CIs) are denoted by blue boxes (blue lines). The combined OR estimate for all the datasets is represented by a green diamond, where the diamond width corresponds to 95% CI bounds. The p-value for the meta-analysis (Meta P) and for the heterogeneity (Het P) of odds ratio is shown. A) Discovery meta-analysis. B) Discovery and replication. C) Plot showing the cumulative incidence of type 2 diabetes for tertiles of the genetic risk score (median follow-up 11 years). The red line represents the T carriers and light blue represents C carriers (n=1,652, cases=158.). D) Discovery and replication after excluding controls younger than 55 years old and OGTT>7.8 mmol/l in both the discovery and replication cohorts when possible.

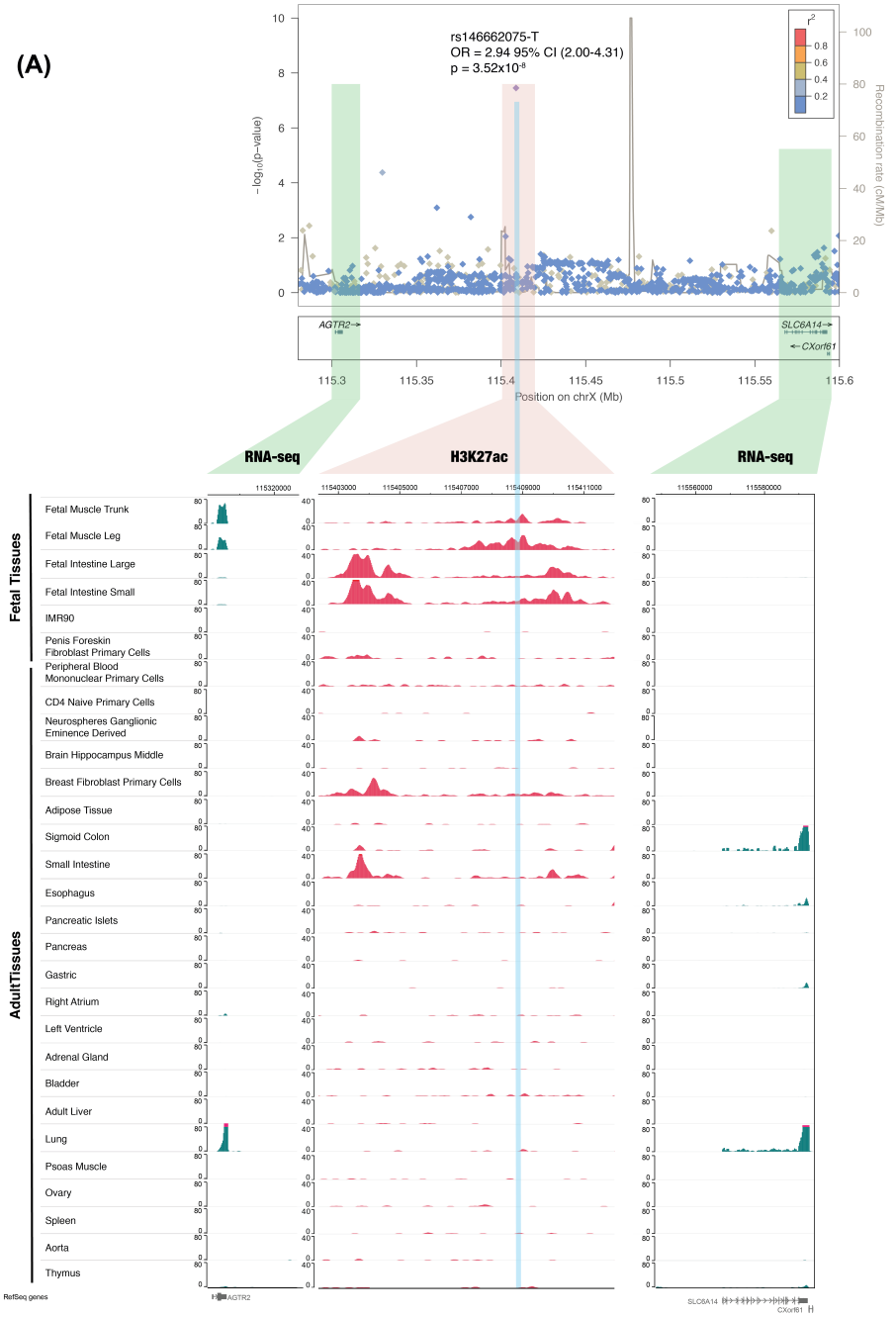
To find additional support for *AGTR2* as the causal gene for this association, we analysed exome sequences of 25,982 (26K) individuals from 5 ancestry groups (European, South Asian, African American, East Asian, and Hispanic) and performed both single-variant and gene-level analyses of T2D risk. We identified 38 low-frequency or population-specific variants (MAF<0.03) variants that were predicted to modify the protein sequence. Among them, we were able to perform association testing for eight variants that were present in more than one cohort and that contained more than 10 allele counts. Only one variant was nominally significant (rs121917810, MAF=0.003, OR=1.7, p-value=0.013), but was not significant after correcting for multiple testing. Burden tests did not yield significant results either. Despite the results were not significant, this dataset was still underpowered

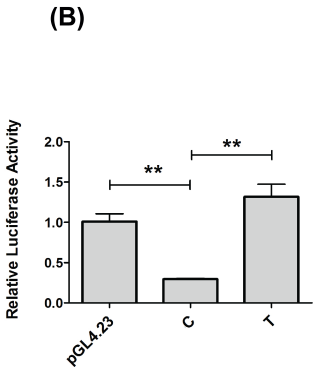
to identify associations in this low allele frequency range. These results thus failed to provide supportive evidence of the protein-coding variants associated with T2D in this gene.

2.7 The rs146662075 T risk allele is associated with 5-fold greater enhancer activity and disruption of allele specific nuclear protein binding

We next investigated the potential functional impact of rs146662075. The variant is located in a highly conserved intergenic region that contains DNase I hypersensitivity sites (Fetal Muscle Leg, Fetal Muscle Trunk, Fetal Kidney and Fetal Lung) and H3K27ac marks for active enhancers (Foreskin Keratinocyte Primary Cells, Fetal Muscle Leg, Fetal Muscle Trunk, Fetal Intestine Small, Fetal Intestine Large and Rectal Mucosa) according to HaploRegv4.1 (Ward and Kellis 2012a; Ward and Kellis 2016). These evidences suggested a potential role of this region in lineage-specific regulatory programs. This variant is located 103 kb downstream to *AGTR2*, the closest gene.

The analysis of the epigenomic data from the Roadmap project (Roadmap Epigenomics et al. 2015) showed that the rs146662075 variant lies in a genomic region that contains conspicuously strong active enhancer chromatin marks (H3K27ac) in human fetal muscle. Furthermore, the analysis of epigenome datasets across multiple tissues indicated that H3K27ac enrichment correlates with expression of *AGTR2*, suggesting that this enhancer may regulate the expression of *AGTR2* (Figure 31A). We thus evaluated whether the region encompassing the rs146662075 variant could act as a transcriptional enhancer and whether allelic variants could affect its activity. We linked DNA segments containing either T or C alleles to a minimal promoter and performed luciferase assays in a mouse myoblast cell line. Luciferase assays showed that the disease-associated T allele consistently exhibited 5-fold greater activity than the C allele, suggesting an activating role of the T allele, or a repressive role of a protein complex that specifically binds the C allele (Figure 31B). Consistent with these findings, electrophoretic mobility shift assays using nuclear protein extracts from mouse myoblast cell lines, differentiated myotubes and human muscle fetal primary tissue revealed that the DNA segment containing rs146662075 exhibited sequence-specific binding activity. This activity was disrupted by the rare T allele, thus pointing to a potential repressive function of the common C allele (Figure 31C). Overall, these results indicate that the rs146662075-T variant maps to a tissue-specific enhancer that correlates with *AGTR2* gene activity, and further indicate that the variant modifies the function of this enhancer. These results suggest that the rs146662075 variant is a regulatory allele for *AGTR2*, a gene that is known to control insulin action.


(A)
rs146662075-T
OR = 2.94 95% CI (2.00-4.31)
p = 3.52x10-8
r2
0.8
0.6
0.4
0.2
−log10(p-value)
Recombination rate (cM/Mb)
AGTR2
SLC6A14
CXorf61
115.3
115.35
115.4
115.45
115.5
115.55
115.6
Position on chrX (Mb)
RNA-seq
H3K27ac
RNA-seq
Fetal Tissues
AdultTissues
Fetal Muscle Trunk
Fetal Muscle Leg
Fetal Intestine Large
Fetal Intestine Small
IMR90
Penis Foreskin Fibroblast Primary Cells
Peripheral Blood Mononuclear Primary Cells
CD4 Naive Primary Cells
Neurospheres Ganglionic Eminence Derived
Brain Hippocampus Middle
Breast Fibroblast Primary Cells
Adipose Tissue
Sigmoid Colon
Small Intestine
Esophagus
Pancreatic Islets
Pancreas
Gastric
Right Atrium
Left Ventricle
Adrenal Gland
Bladder
Adult Liver
Lung
Psoas Muscle
Ovary
Spleen
Aorta
Thymus
RefSeq genes
AGTR2
SLC6A14
CXorf61


(B)
Relative Luciferase Activity
2.0
1.5
1.0
0.5
0.0
**
**
pGL4.23
C
T

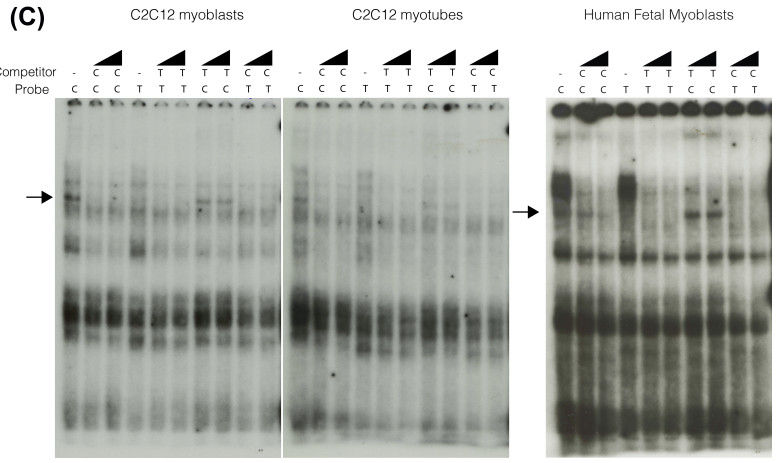

(C)
C2C12 myoblasts
C2C12 myotubes
Human Fetal Myoblasts
Competitor
Probe

Figure 31. Functional characterization of rs146662075 association signal. A) Signal plot for X chromosome region surrounding rs146662075. Each point represents a variant, with its p-value (on a $-\log_{10}$ scale, Y axis) derived from the meta-analysis results from association testing in males. The x-axis represents the genomic position (Hg19). Representation of H3K27ac and RNA-seq in a subset of cell-types for which RNA-seq and H3K27ac was available is also shown. The association between RNA-seq signals and H3K27ac marks suggest that *AGTR2* is the most likely regulated gene by the enhancer that harbors rs146662075. B) The presence of the common allelic variant rs146662075-C reduces enhancer activity in luciferase assays performed in a mouse myoblast cell line. C) Electrophoretic mobility shift assay in C2C12 myoblast cell lines, C2C12 differentiated myotubes and human fetal myoblasts showed allele-specific binding of a ubiquitous nuclear complex. The arrows indicate the allele-specific binding event. Competition was carried out using 50- and 100- fold excess of the corresponding unlabeled probe.

Discussion

This thesis focused on pushing forward the understanding of the genetic basis of complex diseases by putting a great deal of effort in accurate and efficient implementations of genotype-imputation based analytical workflows. This rationale was applied to T2D due to several reasons. T2D has become a modern threat for global health, favoured by the upsurge of obesity tied to unhealthy lifestyles (International Diabetes Federation 2015). Moreover, the high heritable component for T2D liability (Willemsen et al. 2015) underscored the value of strengthening genetic research initiatives, which led to a myriad of large-scale genetic studies previous to this thesis. Thus, the substantial amount of genetic data at our disposal through public repositories gave us the opportunity to test our methodological-driven approach.

1. Challenging the genetic architecture of complex diseases

The study of the genetics of T2D was intensified during the last decade with the emergence of GWAS, which became the most resourceful approach. **The study sample size** increased from a few thousands to more than ~100,000 individuals, **diverse ethnicities** were studied and **the genomic coverage** was extended through genotype imputation with sequence-based reference panels and by directly sequencing the participants. However, the collective effect of ~100 T2D known associated *loci* only explains 10-15% of T2D heritability (Flannick et al. 2016). Are these approaches underpowered to capture the genetic basis of T2D?

By reviewing past genetic studies, this thesis was able to ascertain the flaws and the opportunities on the path to better understand the genetic architecture of T2D. The small effect-sizes (OR ~ 1.1-1.2) found for the majority of the risk variants suggested that a **(i) much larger number of common susceptibility variants showing weaker effect sizes** should be identified. Larger sample sizes and the study of diverse ethnicities may still lead to fruitful results. Additionally, the recurrent findings of **(ii) overlapping genes between monogenic forms of DM and T2D** (Sidransky 2006; Sandhu et al. 2007; Voight et al. 2010; Rees et al. 2011; Cho et al. 2012; Tallapragada et al. 2015) stressed that future genetic research for common forms of DM (T2D) may also benefit from the methodology and the knowledge obtained for rare forms of DM. Finally, the large proportion of rare variants discovered in individual sequenced genomes advocated for **(iii) a more relevant role of rare variants** into disease aetiology (Lupski et al. 2011; Agarwala et al. 2013). This is one of the most controversial matters but what did we observe in these several waves of large-scale genetic studies? First, a few of all the T2D discovered *loci* are driven by rare and low-frequency variants (Steinthorsdottir et al. 2014). Second, many T2D GWAS *loci* found in one ancestry were also replicated in other ethnicities, which would be unlikely if these *loci* were driven by a rare allele. Third, the last compelling WGS study suggested a minor role of low-frequency and rare variants in the T2D liability (Fuchsberger et al. 2016). Looking at these and many more evidences, do we have to mainly steer all our efforts towards the identification of many more common variants with ever-weaker effect sizes? From my point of view, the answer is no. We are just beginning to thoroughly explore the rare and low-

frequency variant spectrum. **WGS studies are still confined to small sample sizes**, and **genotype imputation** is the only strategy statistically powered to study the contribution of the whole landscape of allele frequency to the T2D susceptibility. Additionally, even though low-frequency and rare variants are not governing the heritability of T2D, variants with large effect sizes are crucial to elucidate the molecular mechanisms underlying T2D and to provide novel therapeutic targets. For instance, the protective effect of loss-of-function variants in *SLC30A8* suggests that the inhibition of this islet zinc transporter may lead to a therapeutic prevention strategy for T2D (Flannick et al. 2014).

Although the genetic research efforts for the study of T2D have been intensified since 2005-2006, which has led to a notable pace of discovery, many challenges are still ahead. Thereafter, what sort of research approach did we take and why?

2. Mining existing GWAS data

The take-home message of this thesis is that by applying all the **methodological and computational developments now at our disposal**, we can reveal novel hints of the T2D aetiology still **hidden in a huge amount of publicly available large-scale genetic data.** Based on this rationale, we were committed to contributing to this on-going debate about which spectrum of variants is underlying the genetic architecture of T2D and other common diseases.

Why do we argue that this work can be extremely resourceful at this point?

First, **data-sharing initiatives** such as dbGaP or EGA have been completely **embraced by the epidemiological and the genetic research community**. The dbGaP repository has already collected 733 studies (accessed in 2016-09-06), and the data supplier institutions transcend the academic research field (Paltoo et al. 2014). An example proving this success is the availability of the **Genetic Epidemiology Research on Adult Health and Aging (GERA) project**. The GERA cohort linked genetic data to medical information for more than 78K aged individuals from an ethnically diverse population and represents an immense opportunity to study the genetic risk for a broad range of health conditions and to thoroughly explore pleiotropic effects within a single large dataset (Kvale et al. 2015). Free access to public datasets with huge sample sizes is fundamental **to foster high-profile research projects** capable to detect disease associations with **small effect sizes** or **at the bottom range of the allele frequency spectrum**. The generation and gathering of this amount of clinical and biological samples, until the disposal of the genotyped data, is laborious and expensive. Thus, only big consortia involving multiple institutions can afford it. The downside of this common scenario is that only the members of these consortia may have access to the raw genetic and phenotypic data. Even between them, the exchange of data is in most of the cases minimal because of data sharing polices. To foster scientific advances and innovation, data democratization is essential. We should work on a **framework able to reconcile ethical and privacy policies with a**

wider access to phenotypic and genetic datasets. This scenario will stimulate the development of novel and more adequate methodological approaches.

This need to push forward the boundaries of data-sharing is imperative and is gradually being articulated in several ways. I only focused on decentralising genomic analysis but **large-scale genetic findings should also be interpreted through functional studies** in order to extract a fuller value of this data. Large GWAS consortia partially highlight the value of this genetic data by performing thorough genetic analyses. But to gain the biological insights that these genetic associations can really provide, rigorous custom analyses or functional experiments should be followed. However, if the access to the genomic data sets is confined to a single consortium across disparate locations, the biological community is not able to make a more intelligent use of all the genetic data produced. In order to facilitate this translation of genetic association results to biological insights, the **Type 2 Diabetes Knowledge Portal (T2D Portal)** web-server was launched to enrich the connections between GWAS consortia and experimentalists and other public and private sectors (Flannick and Florez 2016). Therefore, 'consumers' of genomic data, which are usually non-expert users of genomic datasets, are able to interpret through a meaningful interface a variety of genetic analyses across hundreds of thousands of individuals from several GWAS consortia. By removing the barriers around GWAS consortia, genetic associations can be translated into a mechanistic insight.

This approach is thereafter only focused on extending the value from the genetic association data. However, dbGaP or EGA acts on a deeper level, because it empowers researchers outside of big consortia to perform novel and versatile genetic analyses, only attainable by handling the raw genetic-individual level data. This novel re-analyses have the potential to significantly increase the number of GWAS discoveries, which also could be meaningfully evaluated through the integrative T2D Portal. Thus, secondary research uses from dbGaP and the T2D Portal are two non-exclusive initiatives that serve to promote data-sharing. Both cases stress that in order to advance in our comprehension of the T2D biology and to contribute in clinical decision-making, collaboration and common public resources are mandatory. This thesis has focused on demonstrating how a **more democratized use of large-scale genetic data** through public repositories offers the opportunity to take advantage of **novel and powerful analytical approaches**.

Second, when I started this thesis, the phase1 of the 1000G (which was updated multiple times) was the only sequence-based reference panel available for genotype imputation, besides the HapMap reference panel (based on genotyping data). Since then, **more diverse and larger sequence-based reference panels** have emerged such as the final Phase3 release of the 1000G (The 1000 Genomes Project Consortium et al. 2015), the UK10K project (UK10K Consortium et al. 2015) or the recent Haplotype-Reference Consortium (HRC) panel, which comprises 64,976 haplotypes (McCarthy et al. 2016). These novel reference data is expected to truly lead to a significant shift in

the statistical power and accuracy of genotype imputation approaches. Obviously, imputation still falls short of the genomic coverage compared to the one obtained from directly sequencing the participants. However, **conducting WGS on large sample sizes such as in GWAS is still prohibitive** (Huang et al. 2015). As a practical example, what would be the cost of sequencing 70K individuals, which is the sample size analysed for T2D during this thesis? According to the NIH Institute Pricing (www.cidr.jhmi.edu/services/pricing.pdf), low pass (4x) WGS data (approximately corresponds to the specifications from the GoT2D project, ~5x mean coverage, 101-bp paired-end reads (Fuchsberger et al. 2016)) cost 850$. Therefore, the sequencing budget for 70K individuals will roughly be 59.5M of dollars. This example clearly shows the limitation of sequencing technologies in the context of GWAS, as the sample size required cannot be economically afforded. Thus, to study lower and rarer variants through GWAS approaches, genotype imputation based on WGS reference panels is fundamental. However, clear **guidelines** to select the most suitable panel or to precisely discard **badly imputed variants** are necessary if we are resolved to thoroughly explore the broad allele frequency spectrum. This thesis addressed this issue as we noted that the **inclusion of inaccurate imputed genotypes** leads **to spurious associations** and **can wipe out true association signals**.

Third, **parallel computing techniques and HPC environments** are unavoidable in order to handle the **computational burden** derived from: **(1)** analysing gigantic GWAS cohorts such as GERA and **(2)** the application of computationally intense calculations such genotype imputation with ever-larger reference panels. To keep these analyses doable into thousands of samples, HPC infrastructures are necessary (Das et al. 2016), but most research groups **cannot afford these computational resources** or **lack the expertise needed to implement them**. Within large consortia, which own exclusive GWAS case-control data, data analyses are usually performed separately by each of the institutions and then, the results are shared and meta-analysed. This can lead to systematic heterogeneity in the GWAS results generated within each of the groups, which may ultimately result into a loss of statistical power.

This thesis has been developed using the ***MareNostrum* supercomputer**, close to a large community of **experts in computational sciences**, which indeed is one of the main reasons that made this whole study possible. We were outsiders of the majority of T2D consortia, without access to most of the large cohorts generated by them. However, we have been able of gathering public GWAS data that reached a reasonably statistically powerful sample size. Therefore, our trump card was maximizing the statistical power of this data by applying computational intense calculations from genotype imputation. These analyses were thereafter only attainable by exploiting all the computational techniques and resources at our disposal. This centralised large-scale genetic study for T2D involving thousands of subjects would have been almost impossible without our computationally advantageous position in the Barcelona Supercomputing Center (BSC).

After delineating the cornerstones of the hypothesis that drove this thesis, in the next section, I will review the outcomes from the implementation of methodological and computational improvements of GWAS analyses. Afterwards, in section 2.2 I will focus on the novel insights that we gained for T2D.

2.1 Implementation of efficient computational and analytical frameworks for imputation based GWAS

The first main section of this thesis comprises all the computational and methodological implementations required to perform genotype imputation and GWAS-based pipelines in the most accurate and efficient manner. This work has been absolutely relevant to maximize the statistical power of all the publicly available GWAS data, which has been translated into novel insights for T2D. This work also allowed me to participate in other large-scale genetic studies for T2D and other complex diseases, which led to several publications (Bonnelykke et al. 2014; Horikoshi et al. 2016).

2.1.1 Automatization and packaging of GWAS analytical workflows: computationally optimizing a workflow for Quality Control (QC) for genotyped data

Quality control of genotyped data is crucial to avoid **systematic biases in GWAS,** which are amplified by genotype imputation and may lead to spurious associations and decreased statistical power (de Bakker et al. 2008; Anderson et al. 2010). Moreover, ever-increasing large-scale datasets such as the GERA cohort challenged the efficiency of this basic step and urged the **integration of computational solutions.** Inspired by these two premises, this thesis worked on computationally optimizing **widely accepted QC practices** for an effective identification and removal of problematic markers and samples. This packaged protocol facilitated performing, as a single execution, hundreds of QC analysis required for different on-going projects.

A QC protocol is an illustrative example of a pipeline encompassing multiple steps that can require **manual intervention** at several points of the execution and can become a real computational headache with increased sample sizes. The packaged QC only requires the input files and the appropriate parameters that are adjusted using a basic flag scheme. In addition, this automatized workflow is able to optimize and select the **precise computational resources** required at each step, as well as **adjusting the adequate level of parallel tasks** at each point of the execution. We showed that a case-control dataset of 5,828 samples and 741,192 genetic markers followed the full protocol in 2h 51' 14'' without any need of user intervention in a cluster environment. Despite the great advantage of having at our disposal a quick computationally packaged QC, I want to point out some limitations and future improvements that are mandatory. To facilitate the analyses of huge sample sizes, the whole execution was split in three parts to accommodate a separate branch to perform ancestry clustering analyses based on the flashpca software (Abraham and Inouye 2014). This specific analysis should be integrated into the main code. Moreover, this packaged QC is only able to run in cluster environments similar to the *MareNostrum* in terms of the execution queuing

system. This limitation has to be generalized in order to make this tool accessible to all the community.

Finally, working in this computationally optimization made me realize that without the **collaboration with computer science experts**, sophisticated solutions for more complex analytical workflows such as imputation and association testing could not have been attained. Therefore, I have been fortunate to also participate in a join effort between our genomics group and the computer science department from the BSC to create an integrated framework to efficiently deal with large-scale autosomal and X-wide GWAS, genotype imputation and cross-phenotype analysis through parallel computing infrastructures. The final outcome is **GUIDANCE**, which has been conceived to be an easy-to-use efficient solution for large imputed-based GWAS analysis that can exploit cluster infrastructures or cloud computing environments (Sanchez et al. 2016).

2.1.2 Fostering guidelines for accurate genotype imputation of common and low-frequency variants for GWAS and sequence-based reference panels

This thesis addressed the **accuracy of imputed variants** according to **novel sequence-based reference panels.** We established **practical guidelines** for the broad implementation of this strategy to better exploit any GWAS scenario and across a wider allele frequency spectrum. The focus of association studies has been extended beyond common variants to the rare and low allele frequency range, which indeed are more difficult to impute because the lower number of haplotype carriers. Therefore, we aimed to address the performance of a large population specific reference panel such as the UK10K in comparison with a global reference panel such as the 1000G-Phase1 and ultimately, the Phase3 release, for a better analysis of variants at lower allele frequency ranges. This work has been challenging because it was performed while imputation tools and reference panels were suffering a viral evolving progress and all the study had to be redone at different time points. Moreover, while we were able to individually assess the accuracy of the imputation of the novel rare variant associated with T2D in the X chromosome, future work will focus on performing the same assessment for the X-chromosome in a systematic manner. Performing the same assessment in the X-chromosome has additional layers of complexity due to the fact that the **X-chromosome** analysis forced us to **stratify** this evaluation **between males and females**, which compromised our power to evaluate the quality of imputed low-frequency and rare variants in this chromosome. Additionally, **our results are confined to a single methodology**, and thus, we recognize that this work should be extended to other imputation-software tools and resources, as there is no consensus in which is the best genotype imputation algorithmic approach. We decided to work with the Oxford statistical tools because the uncertainty of imputed genotypes is retained by delivering them as probabilities (i.e. an imputed genotype with a probability of 80% of being homozygous for the rare allele and 20% of being heterozygous will be provided as 0 0.20 0.80), they facilitate performing association tests under different inheritance modes and handle X-chromosome data through SNPTEST. All the experiences

learnt from this work were of great importance to foster the discovery of two novel T2D *loci* driven by a low-frequency and a rare variant.

2.1.2.1 Fixing appropriate quality thresholds across genotyping platforms

- *Quality Control filtering for imputed variants*

Our first step was identifying a cut-off able to filter wrongly imputed variants without compromising the genomic coverage that can ultimately wipe out true disease associations. We fixed a **minimal threshold to discard badly imputed variants** at ***info* score ≥ 0.7** (average R^2 of ~0.5 between the true and the imputed allele dosages as determined in the original MACH paper to discern well imputed from badly imputed variants (Li et al. 2010; Pistis et al. 2015)). Interestingly, this cut-off was also suggested for low-frequency variants (MAF < 1%) imputed using reference panels genetically close to the study population (Pistis et al. 2015) or in a recent study focused on comparing the rate of false positives associations between a global and a population specific reference panel (Surakka et al. 2016). This thesis exposes that this threshold **transcends reference panels** and is valid for UK10K, the 1000G-Phase1 and the 1000G-Phase3.

Of note, for this evaluation of the degree of accuracy of genotype imputation, we simulated a GWAS based on a dataset of controls genotyped using different arrays (see Methods). The improved performance obtained for this particular setup in this thesis is consistent with other studies that successfully demonstrated the advantage of using past powered sets of controls (Ho and Lange 2010; Bonnelykke et al. 2014). The full deployment of data-sharing initiatives will continually **challenge common GWAS designs** and will inevitably demand for novel methods capable **to handle with increasing sources of heterogeneity**.

Therefore, we consistently adjusted the quality filters for badly imputed variants and we proved that this **protocol is valid for any reference panel**. Our approach also exhibited greater performances in comparison with previous studies for challenging scenarios such as a study design based on "convenient controls". Our results exposed the **accuracy** and the **versatility** of **genotype imputation** techniques, and encouraged us to struggle to **gain the maximum coverage** by characterizing sequence-based reference panels.

- *Seeking for an increased genomic coverage across the allele frequency spectrum*

Once we were able to determine the fraction of well-imputed variants, we evaluated **genotype imputation accuracy across the whole allele spectrum** through 1000G-Phase1 (and lastly, the 1000G-Phase3) and UK10K reference panels. UK10K outperformed the Phase1 and even the Phase3 release of the 1000G at every allele frequency range in terms of the proportion of false positive associations. However, we saw that the coverage (variants with IMPUTE2-*info* score ≥ 0.7) of the 1000G-Phase1 (and in addition the Phase3) was the most extensive for common variants in

comparison with UK10K, highlighting that these reference panels were complementary. Thus, we sought for **integrating** the contribution of a **cosmopolitan reference panel** with a **population-specific reference panel** and we initially demonstrated that **the combination of 1000G-Phase1 and UK10K better captured the whole spectrum of allele frequency** in autosomes. Huang and colleagues (Huang et al. 2015) also realized about the opportunity of integrating the non-exclusive contributions of distinctive reference panels in order to better reflect the broad allele frequency spectrum. They built a **unified reference panel** from the independent contributions of 1000G-Phase1 and UK10K (Huang et al. 2015), which increased the number of high-confidence variants. In concordance with our results, UK10K was the major contributor for rare variants but 1000G-Phase1 also provided a fraction of rare variants not well captured by UK10K. These variants might have reached very low frequencies in the UK population, but they prevailed elsewhere in Europe or in other ethnicities. As future work, **both strategies for combining reference panels should be compared**, but clearly the advent of novel, larger and diverse sequencing resources is spurring the development of additional methodologies.

Of note, in section 2.2 section, we relied on the integration of 1000G-Phase1 and UK10K reference panels, which was the most powerful approach at that point. But once it was released, the potential of the 1000G-Phase3 could not be overlooked. Indeed, our comparison of reference panels ultimately included the 1000G-Phase3 and for rare variants it increased by a ~68% the number of well-imputed variants respect to the 1000G-Phase1. This increased coverage of rare genetic variation is likely to arise from more **diverse populations** and the **larger sample size** of the panel (Browning and Browning 2016). Hence, we evaluated the integration of the UK10K and the 1000G-Phase3 reference panels that exhibited the most powerful performance at any range of allele frequency. In addition, the integration of the 1000G-Phase3 instead of the 1000G-Phase1 with the UK10K panel also minimized the number of FP associations, denoting the increased accuracy of this new release of the 1000G. The coverage of high confidence rare variants was extremely improved (15.49% increase with respect to the 1000G-Phase1 and UK10K integrated results), suggesting a clear distinctive contribution of rare genetic variants from 1000G-Phase3 and UK10K. Following the rationale of Huang and colleagues (Huang et al. 2015), an integrated reference panel comprising the 1000G-Phase3 and UK10K was built and is available through the EGA repository (https://ega-archive.org/studies/EGAS00001000713). Worthy of note is that highly deleterious rare alleles have been of the outmost interest as they are crucial to elucidate key biological processes from disease aetiology and to provide novel therapeutic breakthroughs. Therefore, these and many other sophisticated solutions are directed towards a better characterization of this fraction of genetic variation. Novel designs such as performing genotype imputation using a combination of SNP and exome chip data for the target dataset and a reference panel based on WES, SNP and exome chip data, offered an increased quality of imputation for rare variants (Kim et al. 2015). Hence, genotype imputation based on **integrating population-specific WGS** reference panels (UK10K) and novel

large-scale WES data such as the ExaC catalogue (Lek et al. 2016) or the recent Genome Aggregation Database (http://gnomad.broadinstitute.org) is a promising approach still unexplored. Moreover, the first release of the HRC comprises the largest collection of human genetic variation, which will be continuously expanded (McCarthy et al. 2016). This reference panel is expected to provide an unprecedented quality of imputation at any range of allele frequency and future work should focus on testing its performance through the limited subset available through the EGA repository (https://www.ebi.ac.uk/ega/studies/EGAS00001001710). Besides the emergence of novel reference panels, the sample size of genotyped datasets has increased from a few thousands until hundreds of thousands of individuals that will facilitate exploring lower disease associated alleles. The new SHAPEIT3 (O'Connell et al. 2016) or the new EAGLE (Loh et al. 2016) algorithm will mitigate the technical bottleneck of large-scale haplotype phasing derived from the analysis of these huge biobanks. In the near future, we will witness a development of imputation techniques parallel to novel software solutions that will require a closer collaboration with the computer science community.

2.1.2.2 Preventing the occurrence of spurious associations from errors in genotyping

We realized that despite applying multiple protocols for genotyped and imputed data, false positive associations still escape and propagate through genotype imputation. This is especially relevant in study designs based on different genotyping arrays, with an even worse rate of spurious associations. We developed a novel **pre-filtering protocol** for the identification of genotyping errors based on comparing the true genotypes from the array with the corresponding ones imputed from the LD structure with the surrounding markers (i.e. re-imputing SNPs that were already typed). We showed that our novel filter was able to achieve a **10.28-fold reduction in the rate of false positives,** which was accompanied by only a ~20% loss of coverage. This strategy may lead to a substantial improvement for strategies based on using convenient controls, especially if a previous study proposed using the fraction of intersecting SNPs between multiple platforms to generate larger sets of controls (Johnson et al. 2013). Limiting the genotyping coverage to the fraction of intersecting SNPs represents that a limited number of typed markers are available, which will compromise the degree of accuracy from genotype imputation. Other approaches for ameliorating this rate of spurious associations consisted in restricting the analysis to imputed SNPs showing high accuracy (MACH R^2 = 0.99) or genotyping a subset of controls on the array of the patient data (Sinnott and Kraft 2012). We believe that our strategy can be easily implemented for checking those doubtful associated *loci*, driven by a single or a few imputed variants, which are likely to be an artefact, before exploring any possible functional effect.

2.1.2.3 Exploring the impact of genotype imputation in meta-analysis

We also addressed which **appropriate filters** should be applied for genotype imputation across case-control datasets genotyped with **the same array technology** and the **effect on meta-analysis**.

Discarding inaccurate imputed variants at each cohort can benefit the **final performance of the meta-analysis** by decreasing the degree of heterogeneity. In line with this, Li, J. and colleagues (Li et al. 2012) discussed that those markers genotyped in a subset of the cohorts from a meta-analysis but imputed in the rest of studies led to a 25% reduction of power, and in some cases the statistical power was even smaller than that of individual studies. This thesis extended this initial observation, and we demonstrated that by relying in **our definition** of **well-imputed variants** (i.e. IMPUTE2-*info* score higher than 0.7), there is no loss of statistical power due to the meta-analysis of imputed and genotyped variants. Therefore, we demonstrated that there is no need of applying additional filters such as looking for variants imputed or genotyped in all the cohorts that minimizes the statistical power of meta-analysis.

To summarize this main section, we addressed several technological and methodological issues as well as many challenges still ahead. Overall, our results underscore many scenarios that can benefit from a broad application of genotype imputation by following our guidelines. All this knowledge has provided me the criteria to widely apply genotype imputation with novel sequence-based reference panels according to multiple scenarios in order to better exploit all the publicly available GWAS datasets. Specifically, our efforts are focused on enlarging the spectrum of allele frequency attainable with deeper large-scale genetic studies. Therefore, once we had achieved a robust methodological basis, as a proof of concept, we proceeded to the analysis of all publicly available GWAS data for T2D.

2.2 Novel insights into the genetic architecture of T2D: crossing the boundaries of common variants

This thesis has been able to unlock novel insights into the T2D aetiology by **re-analysing with genotype imputation and two sequence-based reference panels all publicly T2D GWAS data** available in the dbGaP and EGA repositories. Our approach has benefited from the disposal of the GERA cohort, which corresponds to almost ~70% of the 70KforT2D sample size. Importantly, this cohort gave us the opportunity to **simultaneously phase a large set of individuals,** which is indispensable to achieve higher accurate haplotypes, not attainable with only a few thousands of samples (Williams et al. 2012). Better-phased haplotypes minimizes spurious associations and **increases the chances of detecting associations driven by low-frequency variants** (O'Connell et al. 2016). Moreover, our approach looked for a **unified re-analysis** of independent **published GWAS cohorts** with systematic and homogenous analytical frameworks. Hence, this thesis sought to demonstrate the advantage of applying homogenous analytical pipelines with the aim to minimize between-study heterogeneity, which is a major concern in meta-analysis (Thompson and Sharp 1999). This source of statistical heterogeneity arises from population structure or the variability caused by distinctive analytical approaches undertaken by different research institutions in, for instance, the context of large consortia. Actually, analysing homogenously and systematically

multiple GWAS case-control datasets was effective in the identification of 108 *loci* for schizophrenia including three of them in the X-chromosome (Schizophrenia Working Group of the Psychiatric Genomics 2014). Thus, we had an ideal basis to capture any contribution from the rare and low allele frequency range, and to bring new insights into the vivid debate about the genetic architecture of T2D.

The 70KforT2D meta-analysis enclosed 70,127 individuals, which were imputed with a global (1000G-Phase1) and a population-specific reference panel (UK10K). This strategy substantially improved the number of good quality SNVs and INDELs available to test for association, generating a final resource of ~15 M variants. GWAS has mainly focused on common variants of small effect sizes (Hindorff et al. 2009), and the effect of low-frequency and rare non-coding variants has been started to be explored in isolated populations (Gudbjartsson et al. 2015; Sulem et al. 2015). This work underscored how UK10K, a large-scale population-specific sequence-based reference panel, allowed us addressing this latter question across heterogeneous European-ancestry populations. Moreover, a recent meta-analysis based on genotype imputation with 1000G-Phase1 and UK10K (based on the approach of (Huang et al. 2015)) has also demonstrated that larger reference panels enriched with the most relevant ethnicities fostered the discovery of low-frequency variants of large effects associated with bone mineral density and fracture (Zheng et al. 2015).

2.2.1 Pathway analysis

In order to illustrate the high potential of the 70KforT2D meta-analysis to capture the pathophysiology of T2D, we performed pathway and gene-set enrichment analysis. The overall results showed enrichment for the **insulin-signalling function**, exemplified by capturing the ***response to insulin stimulus*** category within the top ranked cluster of pathways. This result was also supported by the identification of **MAPK signalling related pathways** (Frojdo et al. 2009), which are involved in the insulin-signalling cascade. For instance, the p38-mitogen activated protein kinase (MAPK) belongs to the MAPK super-family and it has been reported to be a critical regulator of hepatic glucose production (Cao et al. 2005; Wu et al. 2006) and lipid metabolism (Xiong et al. 2007). Additionally, SRC and BCAR1 protein complexes were highlighted, and the latter one has been suggested to mediate a **crosstalk between the insulin-signalling cascade and the mitochondrial biology** in one of the papers in which I have participated during this thesis (Mercader et al. 2012). Interestingly, the GERA cohort is the main contributor to the 70KforT2D resource and is based on aged population (median age of 63). We believe that this cohort has extensively enriched our picture of the insulin resistance molecular biology. Decline in insulin function has been associated with a variety of age related changes (Facchini et al. 2001) such as increased adiposity, decreased lean muscle mass, mitochondrial dysfunction and changes in the diet and the physical activity (Petersen et al. 2003; Morley 2010; Michalakis et al. 2013; Atkins et al. 2014; Leon-Latre et al. 2014). Tissue specific expression from enriched genes prioritized **pancreas** as the most relevant tissue influenced by T2D

susceptibility variants. Some of the contributor genes such as *VEGFA*, *ADAMTS9-AS1* and *SLC30A8* were related with the pancreatic islet vascularization and the insulin deliver to the bloodstream, the recruitment and release of insulin from insulin granules, and the maturation, storage and secretion of insulin, respectively (Brissova et al. 2006; Wijesekara et al. 2009; Simonis-Bik et al. 2010; Pound et al. 2011). These biological processes specifically underscore ß-cell related functions. Despite our limitation in the interpretation of this data, we believe that these results taken all together bring a quite consistent glimpse **of the T2D pathophysiology that agrees with widely accepted molecular knowledge** and invited us to dig deeper into the novel discoveries.

2.2.2 Fine-mapping and functional characterization of T2D loci

The discovery power from our 70KforT2D meta-analysis was complemented by the contribution of **publicly available summary statistics** of the DIAGRAM trans-ethnic study (Replication et al. 2014) and the T2D Portal (Flannick and Florez 2016). This strategy led to the identification of **56 GWAS** *loci* significantly associated with T2D of which **7 were novel**.

We better characterized known and novel associated *loci* for T2D with the **99% credible sets of variants,** that benefited from the increased genomic coverage offered by the two-sequence based reference panels used in genotype imputation. The 99% credible set approach allowed us narrowing wide associated *loci* to a small set of variants that have in aggregation 99% probability of including the true causal variant. Afterwards, for each 99% credible set of variants, we provided the corresponding **functional annotation** that addressed the impact on the protein and gene expression. Actually, we thoroughly analysed how these variants were influencing gene expression in a large collection of tissues (Mele et al. 2015) and specifically in pancreatic islets (Fadista et al. 2014). Moreover, the denser coverage from our imputed data was translated into a notable involvement of INDELs in the 99% credible sets of variants. The importance of this type of variation is supported by independent studies that observed how hundreds of small and large structural variants are in high LD with known trait-associated SNPs (Sudmant et al. 2015; Hehir-Kwa et al. 2016). Moreover, other studies have also benefited from genotype imputation with sequence-based reference panels, such as a large meta-analysis for fibrinogen, in which a known and four novel associated *loci* were led by INDELs (de Vries et al. 2016). Therefore, INDELs can be a **novel source of functional candidate variants**.

To make the data more easily available, all these results will be accessible to the whole community by an initial deposit of the summary statistics to the T2D Portal (www.type2diabetesgenetics.org/).

2.2.3 Identification of novel signals driven by common variants

Five novel *loci* were driven by common variants with modest effect sizes but they may provide highly valuable insights of the pathophysiology of T2D. Of note, the *NEUROG3* gene in the 10q22.1 *locus* is a **key regulator of the development of pancreatic islets and enteroendocrine cells.** *NEUROG3*

was reported to be involved in the genetics of a Permanent Neonatal Diabetes Mellitus (PNDM) (Tallapragada et al. 2015) and recessive mutations were attributed to a PNDM subtype with a severe congenital mal-absorptive diarrhea (Rubio-Cabezas et al. 2011). Our data joints other evidences showing the **overlap between the molecular and genetic background of rare and common forms of DM** (Flannick and Florez 2016). These observations are increasingly delineating DM as a disease continuum, in which common and rare forms are two different end-points of this landscape of phenotypes.

2.2.4 Identification of a novel locus driven by a low-frequency variant

Our unified re-analysis of public GWAS data with genotype imputation and global and population-specific panels led to an excellent performance at lower allele frequency ranges. The majority of known T2D *loci* driven by low-frequency variants identified in European populations have been replicated by our approach, including the low-frequency variants in the *PAM* and the *CCND2* gene (Steinthorsdottir et al. 2014). The **rs76895963 *CCND2* intronic variant** only reached GWAS significance in our study when using the **UK10K reference panel**. The last state of the art paper for T2D identified this *locus* by means of a WGS-based association study. However, in that particular study, the *CCND2* signal was lost in the final meta-analysis that also included imputed data, despite the study had nearly 100% statistical power to identify this *locus* (Fuchsberger et al. 2016). The loss of the *CCND2* signal from the WGS study after the meta-analysis with the imputed data suggested that a **sub-optimal quality filter of genotype imputed data and an inadequate reference panel** might constrain the identification of low-frequency variants, even with sufficiently powered datasets.

These evidences suggested that our 70KforT2D resource had a good coverage at the low allele frequency range, which in fact resulted into the identification of a novel low-frequency missense variant in the *EHMT2* gene showing modest effect size (rs115884658, OR=1.22, p-value=2.33×10^{-11}). The 99% credible set of variants did not clarify which is the most plausible effector transcript and further larger scale fine-mapping efforts will be needed to elucidate this. However, the *EHMT2* gene was reported to be crucial in the FOXO1 translocation induced by insulin, which by decreasing the expression of *PCK1*, **represses gluconeogenesis** (Arai et al. 2015). Alternatively, another potential effector gene can be the ***CLIC1*** gene, for which we identified an active eQTL across multiple tissues and in pancreatic islets. Moreover, through the dSysMap web-service (Mosca et al. 2015) we addressed if protein-coding variants in *CLIC1* can **interfere with binding interfaces** and in **the protein structure**. We noticed that CLIC1 was interacting with the carnitine palmitoyltransferase 1 (CPT1A), which is crucial for the mitochondrion fatty acid oxidation (FAO) (Eaton 2002; Schooneman et al. 2013). The role of FAO in relation to insulin resistance has not been elucidated, and it is unclear whether improves insulin resistance by decreasing lipid accumulation (Krssak et al. 1999; Dobbins et al. 2001; Bruce et al. 2006; Holland et al. 2007) or the excess of FAO intermediaries worsens insulin resistance (Koves et al. 2008; Mihalik et al. 2010; Muoio and Neufer 2012).

Interestingly, recent studies based on disease networks linked protein coding variants with protein-protein interactions (Wang et al. 2012b) or novel insights from the aetiology of T2D were gained by the integration of GWAS and systems biology approaches (Mercader et al. 2012). These approaches demonstrated how the understanding of interaction networks might be an effective approach to disentangle the genetics underlying complex diseases. Therefore, we advocate for integrating large catalogues of protein coding variants such as ExAC (Lek et al. 2016) with resources such as dSysMap (Mosca et al. 2015), an extensive resource for mapping human disease-related mutations on the structural interactome to facilitate rationalizing the underlying mechanism of action.

2.2.5 Identification of a novel rare variant in the X-chromosome

In contrast with the majority of previous T2D meta-analyses, **we incorporated the X-chromosome in the analysis.** We nominally replicated the unique T2D known *locus* found in the X-chromosome (rs5945326, OR=1.15, p-value=4.97×10^{-2}; new lead chrX:152889460:I, OR=1.25, p-value=3.50×10^{-4}, R^2 with rs5945326=0.62) (Voight et al. 2010) and we identified a **novel rare variant association** close to the *AGTR2* gene. The risk for T2D was increased in males by nearly three fold (rs146662075, OR=2.72, p-value=1.73×10^{-8}). The well ascertainment of this rare variant, with the largest OR ever identified in European ancestry for T2D, was tied to the **UK10K** reference panel**,** which underlined the importance of using a population-specific reference panel**.** Additionally, since the phasing in males is unnecessary, the **X-chromosome analysis** may offer a bright opportunity to **dig into the role of rare variants** in the **genetics beneath complex diseases**. Indeed, we align with other initiatives that suggested the potential contribution of the X-chromosome, beyond T2D risk, which resulted in specific methodologies for enhancing X-wide association studies (Chang et al. 2014).

Multiple previous studies have implicated angiotensin II in insulin sensitivity (Kim et al. 2006; Shum et al. 2013; Underwood and Adler 2013), and for instance the deletion of *AGTR2* was reported to be protective for diet-induced insulin resistance by exerting a negative control on lipid utilization in muscles (Yvan-Charvet et al. 2005). However, the stimulation of the angiotensin II receptor type 2 by the C21 agonist accompanied with the PPARγ activation was shown to reduce insulin resistance in T2D mice by enhanced adipocyte differentiation and possibly, by a protective effect on pancreatic islets ß-cells (Ohshima et al. 2012). Therefore, the role of *AGTR2* in insulin resistance is still unclear. Our study proved that this rare variant perturbs the activity of a distal enhancer of *AGTR2* in cell lines and in primary human tissue. Our results pointed out that a gain of function of *AGTR2* is contributing to the pathophysiology of T2D in the T risk allele carriers. Therefore, by blocking this protein in the most appropriate tissue, a novel therapeutic strategy for T2D for this highly-risk group of individuals may be efficient. Nonetheless, we should previously focus our efforts on the identification of the regulatory protein that binds in an allele-specific manner to the enhancer to elucidate the molecular mechanism underlying this association.

2.2.6 Future goals: beyond additive genetic variance

By only relying on data-sharing and improved analytical approaches, this study has been able to undercover new susceptibility genes as well as to propose novel molecular mechanisms which may led to new preventing and therapeutic breakthroughs. Moreover, our results indicated that **rare and low-frequency variants** are **true contributors** to the susceptibility of T2D and are crucial for the comprehension of the underlying key biological processes. We hypothesise that more low-frequency and rare variants may be uncovered by applying the same strategy exposed in this thesis but in even larger sample sizes, such as the used in previous T2D GWAS meta-analyses.

Besides the allele spectrum, the genetic architecture of T2D has many other open fronts. For instance, genetic variation is assumed to contribute to the risk of complex diseases under an additive model (Balding 2006). Less unexplored still is how missing heritability can also be explained under **non-additive models**. Actually, common variants influencing disease risk in a recessive manner (Vukcevic et al. 2011) can be captured under the additive model as the high number of available homozygotes leads to a strong signal (Vukcevic et al. 2011). At lower allele frequencies, the power of the additive model to capture recessive associations is notably reduced. Therefore, the analysis of the **role of recessive effects** in the genetic basis of complex diseases may lead to novel insights. Specifically, non-additive rare disease alleles can provide fruitful discoveries considering the high proportion of rare Mendelian alleles conferring risk under a dominant or a recessive model. In addition, detecting which genetic model best fits a disease association may be important to better optimize the use for predictive purposes (Salanti et al. 2009). To illustrate the importance of non-additive modes, the novel signal in *TBC1D4* found in Greenlandic population exposed the recessive inheritance for T2D (Moltke et al. 2014).

Non-additive effects also include **epistatic interactions**, which are arduous to identify and even to study and they are one potential argument of the small fraction of heritability explained for complex diseases. Additionally, cis-regulatory common variants have been estimated to affect 20% of protein-coding variants in a tissue-specific manner for a single subject (Dimas et al. 2008). Actually, a previous study demonstrated that interactions such as **common cis-regulatory variation** that **modify the penetrance of rare putatively deleterious coding variants** are likely to contribute to the genetic architecture of complex diseases (Lappalainen et al. 2011). Considering the intense genetic research towards loss-of-function coding variants, the integration with RNA-sequencing data from relevant tissues may elucidate how the predicted functional effect is translated in the downstream pathways and in the phenotype.

Our approach has indeed extensively benefited from the disposal of novel functional annotation resources for regulatory functional variants. Therefore, among other approaches as the

aforementioned to push forward the study of rare and low allele frequency disease alleles, **burden tests** that have been confined to protein coding units can be now **extended to non-coding regions**.

These and many more evidences and emerging lines of research are underscoring the lack of perspective when applying a reductionist view to T2D predisposition. Clearly, the genetic contribution to T2D transcends the common allele frequency range but to fully understand this complex picture, diverse exposures involved during lifetime should be taken into account or for instance the role of transgenerational epigenetic inheritance of diabetes risk (Jimenez-Chillaron et al. 2016). Therefore, before elucidating which elements are governing the individual risk for T2D, many factors should be thoroughly characterized. We have still a long path ahead to fully understand the complete mutational load of an individual genome, with much rarer variants and even the interaction with *de novo* and somatic variants.

3. CONCLUDING REMARKS

The dissertation of this thesis aimed to **maximize the discovery power of GWAS data** by a **thorough assessment of genotype imputation analytical methodologies** and the **opportunities from computational techniques and HPC infrastructures**. We demonstrated that the methodological experience gained during these years empowered this thesis to extract novel insights into the pathophysiology of T2D that were hidden in publicly available GWAS data but also this knowledge has been applied in other large-scale analyses for several complex diseases. Of note, two of our novel T2D *loci* were driven by low-frequency and rare variants, respectively. This suggests that the failure in linking lower allele frequencies to the T2D aetiology in previous studies, with larger sample sizes than the one analysed here, may come from a technical limitation rather than limited statistical power. Overall, this work represents a proof of concept of how **data sharing initiatives** in conjunction with **appropriate methodologies** can provide **novel biological hypothesis about the molecular mechanisms governing complex diseases**. We are encouraged to translate this strategy into the analysis of all publicly available GWAS data for many complex diseases. We believe that by building a map for all the spectrum of allele-frequency variants on the disease risk for the majority of complex diseases, we can capture complex interactions and pleiotropic effects. This next layer will be crucial for the translation of GWAS results into predictive, preventive, and personalised medicine.

Conclusions

1. Implementation of efficient computational and analytical frameworks for imputation based GWAS

1.1) Quality controlled imputed data with the IMPUTE2-*info* score ≥ 0.7 filter minimizes the occurrence of false positive associations across global and population-specific sequence-based reference panels at any range of allele frequencies.

1.2) This filter was optimal to maximize the statistical power of meta-analytic approaches by reducing the levels of between-study heterogeneity.

1.3) The integration of the imputed results from a population-specific (UK10K) and a global reference panel (1000G-Phase3) exhibited the highest performance across all the spectrum of allele frequencies.

1.4) Our filtering protocol based on the *diffBeta* parameter identifies genotyping errors that escape standard QC practices and propagate through genotype imputation. This facilitates discarding artefacts from association analyses and the adoption of strategies such as reusing convenient controls.

2. Novel insights into the genetic architecture of T2D: crossing the boundaries of common variants

2.1) By re-analysing individual-level genetic data from 70,127 individuals with genotype imputation and sequence-based reference panels, seven novel *loci* associated with T2D were identified.

2.2) Genotype imputation with multiple sequence-based reference panels provides a better characterization of all the T2D associated regions and facilitates highlighting the more plausible candidate causal variants.

2.3) Our results showed that INDELs and large deletions are potential candidate causal variants for known and novel *loci*.

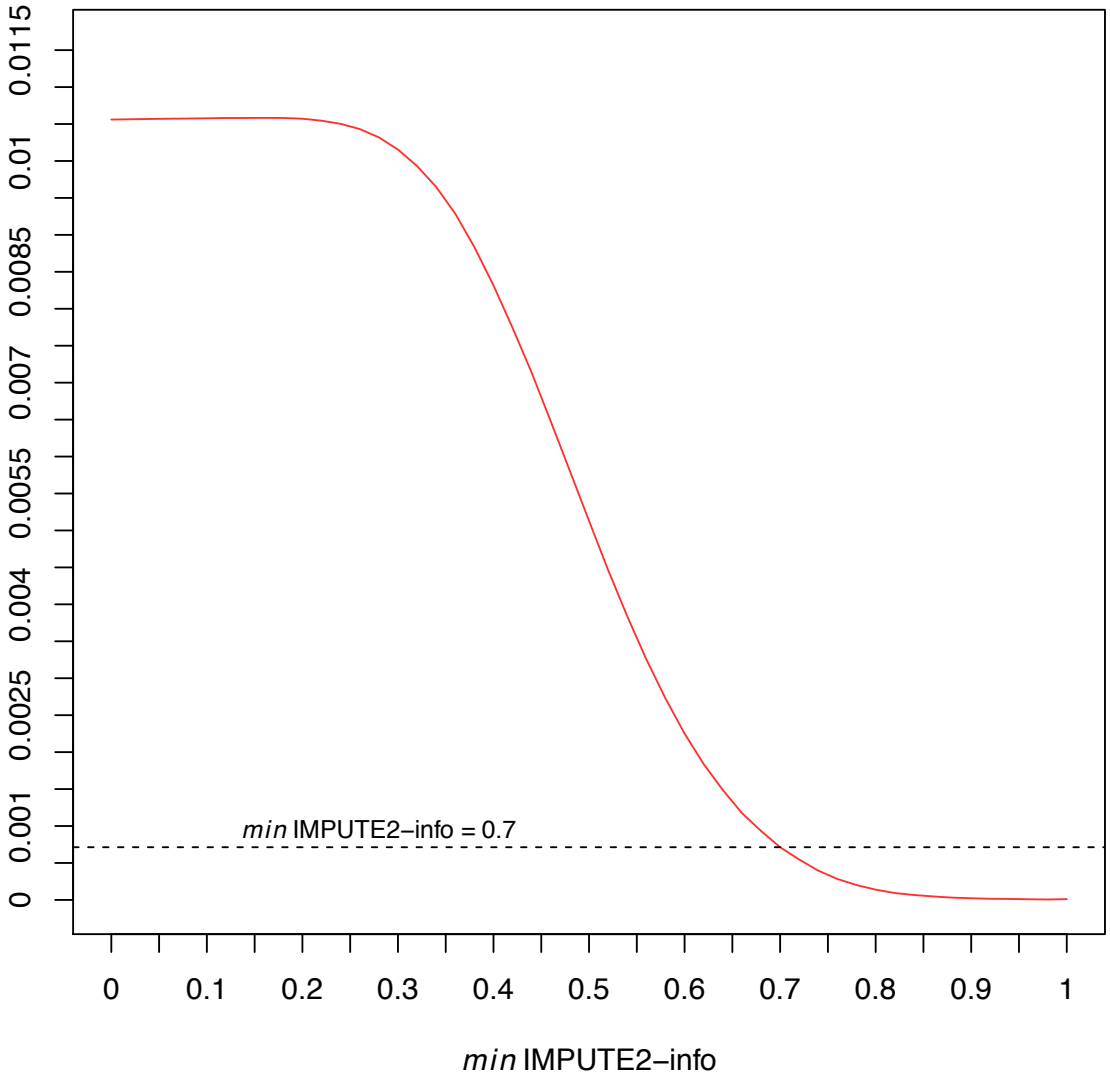
2.4) We identified two novel *loci* driven by a low-frequency and a rare variant, which suggests that the genetic architecture of T2D extends well beyond common genetic variants.

2.5) A novel *locus* driven by a rare variant and showing the largest OR found in European populations for T2D was identified in the X-chromosome and underscored the importance of genome-wide analyses of the sexual chromosomes.

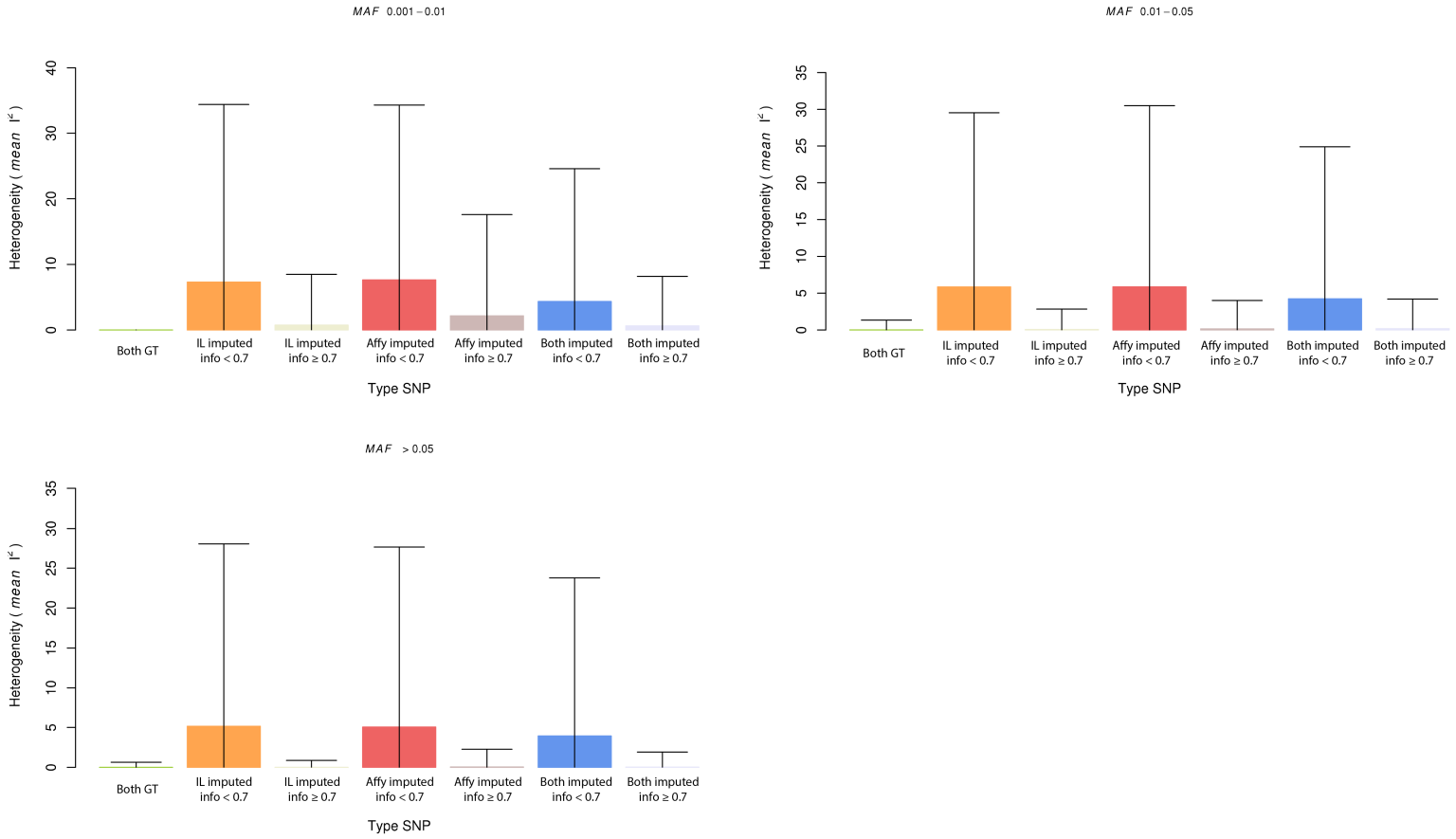
2.6) This rare variant alters the function of a distal enhancer, resulting in a gain of function of the *AGTR2* gene, which may lead to a novel therapeutic strategy based on blocking the angiotensin II receptor type 2.

2.7) The use of publicly available GWAS data through the implementation of accurate genotype imputation with sequence-based reference panels is a cost-effective approach to obtain novel insights into the genetic basis of T2D.

Supplementary Material



Supplementary Material 1. Proportion of high heterogeneity SNPs ($I^2 \geq 50$, y-axis), across different IMPUTE2-*info* score cut-offs (x-axis) using an inverse variance fixed effects meta-analysis.



Supplementary Material 2. Comparison of the levels of heterogeneity using different post-imputation QCs across different ranges of allele frequency (inverse-variance fixed effects meta-analysis). Bars represent mean heterogeneity (I^2) across different allele frequency ranges after filtering for IMPUTE2-*info* ≥ 0.7. Different groups of variants are analysed: Both GT (variants genotyped by both platforms), IL imputed (variants genotyped by Affymetrix and not in Illumina), Affy imputed (variants genotyped by Illumina and imputed by Affymetrix) and Both imputed (variants imputed for both sets of data)

Supplementary Material 3. Description of T2D cohorts included in the meta-analysis.

Study ID	Study	Ethnic group (country of origin)	Case-control status	Sample size	QCed Sample characteristics		Genotyping array
				(males/females)	Sample size (males/females)	Co-variates	
dbGaP phs000237.v1.p1	NuGENE NORTHWESTERN	European (USA)	Cases	556 (306/250)	527 (292/235)	BMI[1], sex, PC[2]1:7	Illumina Human1M-Duov3_B
			Controls	614 (270/344)	601 (264/337)		
dbGaP phs000100.v4.p1	FUSION	European (Finland)	Cases	919(532/387)	901 (526/375)	BMI, age, sex, PC1:7	Illumina HumanHap300v1.1
			Controls	787 (412/375)	772 (402/370)		
dbGaP phs000091.v2.p1	GENEVA NHS/HPFS	European (USA)	Cases	2680 (1164/1516)	2614 (1130/1484)	BMI, age, sex, PC1:7	Affymetrix AFFY_6.0
			Controls	3148 (1338/1810)	3061 (1299/1762)		
EGAS00000000005 (EGAD00000000001 + EGAD00000000002 + EGAD00000000009)	WTCCC	European (UK)	Cases	1999 (1162/837)	1894 (1097/797)	sex, PC1:7	Affymetrix 500K
			Controls	2999 (1470/1529)	2917 (1432/1485)		
dbGaP phs000674.v1.p1	GERA	European (USA)	Cases	7703 (4124/3579)	6995 (3741/3254)	BMI_CAT, BIRTH YEAR (categorical), sex, PC1:7	Affymetrix Axiom_KP_UCSF_EUR
			Controls	54578 (20818/33760)	49845 (19073/30772)		
			TOTAL	**75983 (31596/44387)**	**70127 (29256/40871)**		

BMI: Body Mass Index
PC1-7: Principal Components used to adjust for population stratification.
BMI_CAT: In the GERA the BMI was presented as a categorical variable defined by 5 different BMI intervals.
BIRTH_YEAR_CAT: In the GERA the age was presented as a categorical variable defined by 14 different birth year intervals.

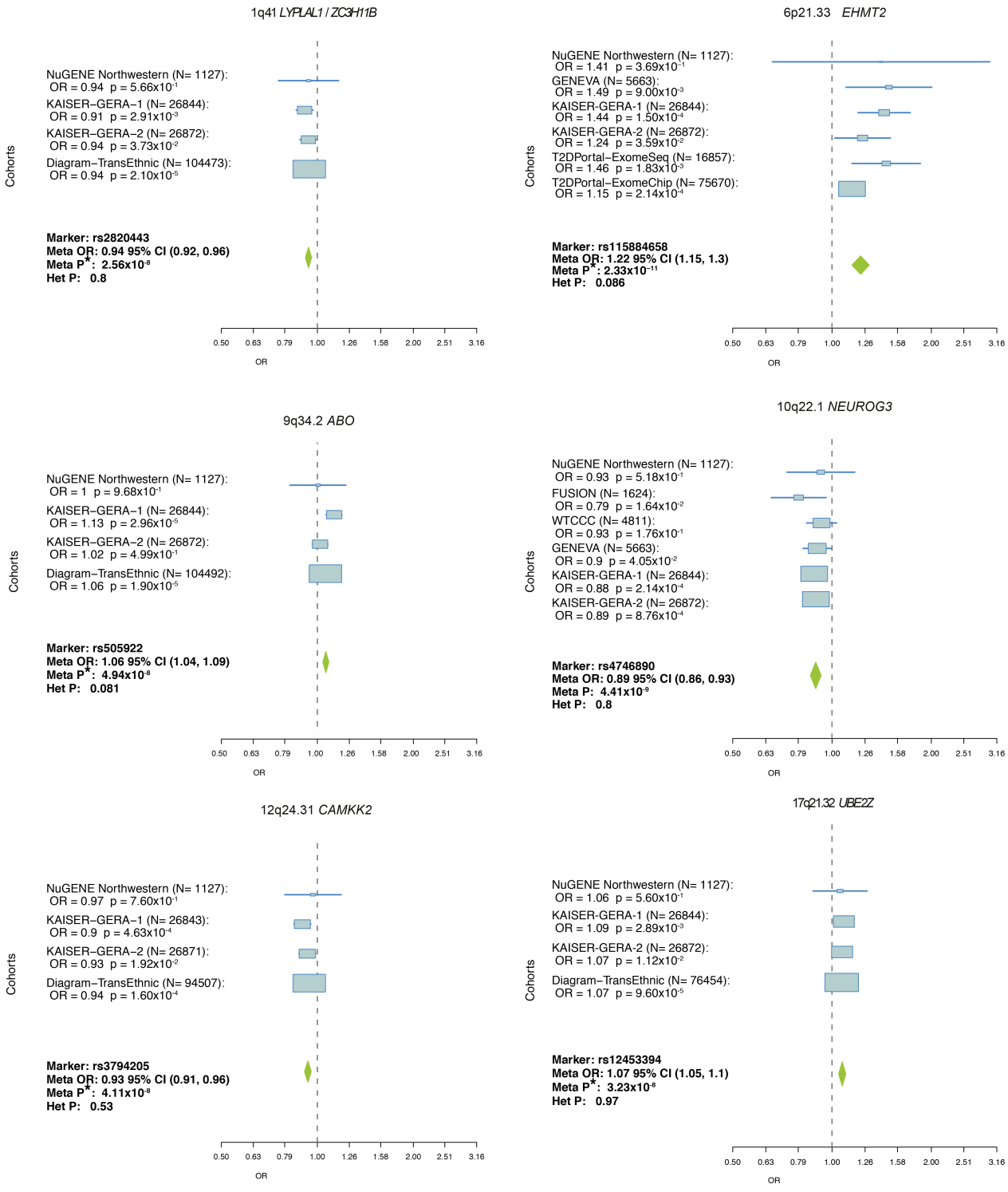
Supplementary Material 4. Tissue enrichment of genes at T2D associated *loci* (p-value<$1e^{-05}$, FDR≤0.20)

MeSH term	Name	MeSH first level term	MeSH second level term	Nominal p-value	FDR
A03.734	Pancreas	Digestive System	Pancreas	7,78E-04	<= 0.05
A05.360.319.679.690	Myometrium	Urogenital System	Genitalia	3,83E-03	<= 0.20
A03.556.875.500	Esophagus	Digestive System	Gastrointestinal Tract	5,06E-03	<= 0.20
A07.541	Heart	Cardiovascular System	Heart	9,65E-03	<= 0.20
A07.541.560	Heart Ventricles	Cardiovascular System	Heart	9,85E-03	<= 0.20
A03.556.249.249.209	Cecum	Digestive System	Gastrointestinal Tract	0,01	<= 0.20
A10.165.114	Adipose Tissue	Tissues	Connective Tissue	0,02	<= 0.20
A03.556.124.526.767	Rectum	Digestive System	Gastrointestinal Tract	0,02	<= 0.20
A07.541.358	Heart Atria	Cardiovascular System	Heart	0,02	<= 0.20
A10.165.114.830.750	Subcutaneous Fat	Tissues	Connective Tissue	0,02	<= 0.20
A10.165.114.830	Adipose Tissue White	Tissues	Connective Tissue	0,02	<= 0.20
A03.556.249.249.356.668	Colon Sigmoid	Digestive System	Gastrointestinal Tract	0,02	<= 0.20
A06.407.071.140	Adrenal Cortex	Endocrine System	Endocrine Glands	0,02	<= 0.20
A06.407.071	Adrenal Glands	Endocrine System	Endocrine Glands	0,02	<= 0.20
A07.231.114	Arteries	Cardiovascular System	Blood Vessels	0,02	<= 0.20
A03.556.249	Lower Gastrointestinal Tract	Digestive System	Gastrointestinal Tract	0,03	<= 0.20
A03.556.249.249	Intestine Large	Digestive System	Gastrointestinal Tract	0,03	<= 0.20
A07.541.358.100	Atrial Appendage	Cardiovascular System	Heart	0,03	<= 0.20
A05.360.444.492	Penis	Urogenital System	Genitalia	0,03	<= 0.20
A03.556.875	Upper Gastrointestinal Tract	Digestive System	Gastrointestinal Tract	0,03	<= 0.20
A03.556.249.249.356	Colon	Digestive System	Gastrointestinal Tract	0,03	<= 0.20
A10.165.114.830.500	Abdominal Fat	Tissues	Connective Tissue	0,03	<= 0.20
A10.165.114.830.500.750	Subcutaneous Fat Abdominal	Tissues	Connective Tissue	0,03	<= 0.20

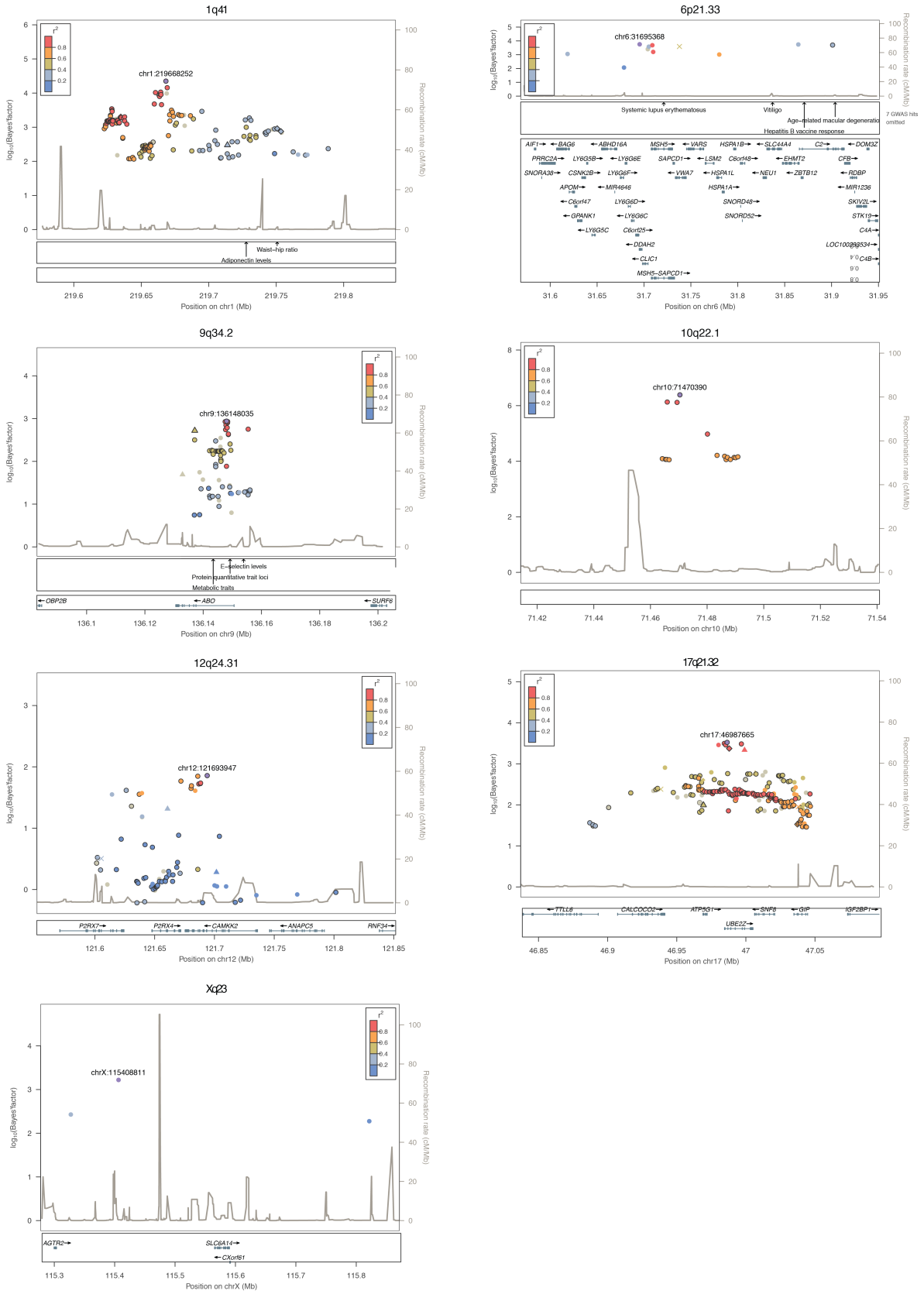
Supplementary Material 5. Tissue enrichment of genes at Type II Diabetes associated *loci* (P-value<$1e^{-05}$, FDR≤0.20)

Original gene set ID	Original gene set description	Nominal p-value	FDR
GO:0071375	cellular response to peptide hormone stimulus	1.47e-10	<= 0.05
GO:0032868	response to insulin stimulus	3.97e-08	<= 0.05
GO:0032869	cellular response to insulin stimulus	4.95e-08	<= 0.05
GO:0043434	response to peptide hormone stimulus	8.04e-08	<= 0.05
GO:0032870	cellular response to hormone stimulus	3.57e-07	<= 0.05
GO:0071495	cellular response to endogenous stimulus	1.38e-06	<= 0.05
GO:0008286	insulin receptor signaling pathway	6.25e-05	<= 0.05
GO:0045860	positive regulation of protein kinase activity	2.03e-06	<= 0.05
GO:0051347	positive regulation of transferase activity	2.95e-06	<= 0.05
GO:0033674	positive regulation of kinase activity	6.11e-06	<= 0.05
GO:0000165	MAPK cascade	7.09e-05	<= 0.05
ENSG00000134308	YWHAQ PPI subnetwork	4.21e-06	<= 0.05
ENSG00000165699	TSC1 PPI subnetwork	3.87e-05	<= 0.05
ENSG00000108953	YWHAE PPI subnetwork	4.24e-05	<= 0.05
ENSG00000103197	TSC2 PPI subnetwork	7.94e-05	<= 0.05
ENSG00000166913	YWHAB PPI subnetwork	0.000116	<= 0.05
ENSG00000082701	GSK3B PPI subnetwork	0.000191	<= 0.05
MP:0011100	complete preweaning lethality	7.36e-06	<= 0.05
MP:0001722	pale yolk sac	2.11e-05	<= 0.05
MP:0000295	trabecula carnea hypoplasia	2.18e-05	<= 0.05
MP:0011098	complete embryonic lethality during organogenesi	2.85e-05	<= 0.05
MP:0005312	pericardial effusion	3.01e-05	<= 0.05
GO:0001701	in utero embryonic development	5.69e-05	<= 0.05
MP:0011101	partial prenatal lethality	8.35e-05	<= 0.05
MP:0001651	necrosis	0.000124	<= 0.05
MP:0004076	abnormal vitelline vascular remodeling	0.00013	<= 0.05
MP:0004255	abnormal spongiotrophoblast layer morphology	0.00013	<= 0.05
MP:0004787	abnormal dorsal aorta morphology	0.000176	<= 0.05
MP:0001698	decreased embryo size	0.000211	<= 0.05
MP:0003984	embryonic growth retardation	0.000241	<= 0.05
MP:0008803	normal placental labyrinth vasculature morpholo	0.000318	<= 0.05
MP:0002652	thin myocardium	0.000352	<= 0.05
ENSG00000138685	FGF2 PPI subnetwork	0.000417	<= 0.05
ENSG00000166908	PIP4K2C PPI subnetwork	0.000489	<= 0.05
MP:0003229	abnormal vitelline vasculature morphology	0.000514	<= 0.05
ENSG00000100030	MAPK1 PPI subnetwork	9.4e-06	<= 0.05
ENSG00000102882	MAPK3 PPI subnetwork	6.32e-05	<= 0.05
ENSG00000166501	PRKCB PPI subnetwork	0.000149	<= 0.05
ENSG00000169032	MAP2K1 PPI subnetwork	0.000566	<= 0.05
OME_SIGNALING_BY_	REACTOME_SIGNALING_BY_SCF:KIT	9.48e-06	<= 0.05
GG_PROSTATE_CANC	KEGG_PROSTATE_CANCER	9.52e-06	<= 0.05
TREAM_SIGNALING_C	ME_DOWNSTREAM_SIGNALING_OF_ACTIVAT	0.000103	<= 0.05
TOME_SIGNALLING_B	REACTOME_SIGNALLING_BY_NGF	0.000125	<= 0.05
TOME_SIGNALING_BY	REACTOME_SIGNALING_BY_FGFR	0.000135	<= 0.05
HRONIC_MYELOID_LE	KEGG_CHRONIC_MYELOID_LEUKEMIA	0.000237	<= 0.05
GO:0019902	phosphatase binding	0.000258	<= 0.05
NSULIN_SIGNALING_P	KEGG_INSULIN_SIGNALING_PATHWAY	0.00033	<= 0.05
NG_VIA_TRKA_FROMF	_SIGNALLING_VIA_TRKA_FROM_THE_PLAS	0.000395	<= 0.05
WNSTREAM_SIGNAL_C	TOME_DOWNSTREAM_SIGNAL_TRANSDUC	0.000506	<= 0.05
REACTOME_AMYLOIDS	REACTOME_AMYLOIDS	9.86e-06	<= 0.05
MP:0001219	thick epidermis	0.000169	<= 0.05
APOPTOTIC_EXECUTI	EACTOME_APOPTOTIC_EXECUTION__PHAS	0.000219	<= 0.05
GO:0006917	induction of apoptosis	1.67e-05	<= 0.05
GO:0012502	induction of programmed cell death	2.07e-05	<= 0.05

ENSG00000197122	SRC PPI subnetwork	3.79e-05	<= 0.05
ENSG00000186716	BCR PPI subnetwork	4.44e-05	<= 0.05
ENSG00000171105	INSR PPI subnetwork	6.24e-05	<= 0.05
GO:0004713	protein tyrosine kinase activity	9.97e-05	<= 0.05
ENSG00000050820	BCAR1 PPI subnetwork	0.00031	<= 0.05
ENSG00000090020	SLC9A1 PPI subnetwork	0.000316	<= 0.05
ENSG00000176105	YES1 PPI subnetwork	0.000464	<= 0.05
GO:0019897	extrinsic to plasma membrane	0.000548	<= 0.05
MP:0002078	abnormal glucose homeostasis	3.88e-05	<= 0.05
MP:0000187	abnormal triglyceride level	6.06e-05	<= 0.05
MP:0001783	decreased white adipose tissue amount	9.38e-05	<= 0.05
MP:0005331	insulin resistance	0.000132	<= 0.05
GO:0042803	protein homodimerization activity	0.000148	<= 0.05
MP:0005668	decreased circulating leptin level	0.000186	<= 0.05
MP:0005459	decreased percent body fat	0.000345	<= 0.05
MP:0003566	abnormal cell adhesion	4.66e-05	<= 0.05
ENSG00000066032	CTNNA2 PPI subnetwork	0.000315	<= 0.05
GO:0005924	cell-substrate adherens junction	0.000432	<= 0.05
ENSG00000150867	PIP4K2A PPI subnetwork	0.000486	<= 0.05
GO:0005925	focal adhesion	0.000511	<= 0.05
GO:0019901	protein kinase binding	6.4e-05	<= 0.05
GO:0019900	kinase binding	0.000128	<= 0.05
GO:0008134	transcription factor binding	8.04e-05	<= 0.05
GO:0003705	tal enhancer sequence-specific DNA binding tran	0.000338	<= 0.05
GO:0040008	regulation of growth	0.000451	<= 0.05
GO:0051427	hormone receptor binding	0.000468	<= 0.05
GO:0043566	structure-specific DNA binding	0.000473	<= 0.05
GO:0035591	signaling adaptor activity	8.08e-05	<= 0.05
GO:0005070	SH3/SH2 adaptor activity	0.000175	<= 0.05
TOME_ERKMAPK_TAR	REACTOME_ERKMAPK_TARGETS	9.94e-05	<= 0.05
_KINASE_AND_TRANSR	_EVENTS_KINASE_AND_TRANSCRIPTION_I	0.000176	<= 0.05
ENSG00000169083	AR PPI subnetwork	0.000116	<= 0.05
ENSG00000135679	MDM2 PPI subnetwork	0.000149	<= 0.05
ENSG00000170315	UBB PPI subnetwork	0.000408	<= 0.05
ENSG00000110092	CCND1 PPI subnetwork	0.000503	<= 0.05



Supplementary Material 6. ForestPlots of the novel T2D associated *loci.*



Supplementary Material 7. 99% credible sets of variants representation of the novel T2D associated *loci.*

Supplementary Material 8. Conditional Analysis of the rs115884658 variant leading the novel *EHMT2* association and T2D and T1D leading variants from the MHC region.

Test Variant	Conditioning Variant	R2 LD between tested and conditioning SNP	MAF	OR	95% CI	P-value	Direction	HetISq
rs115884658	-	-	0,02	1,31	(1.17-1.47)	3,62E-06	?+-+++	42,8
rs115884658	rs9268835	0,06	0,02	1,24	(1.10-1.39)	4,20E-04	+-+++	34,8
rs115884658	rs9273401	0,11	0,02	1,18	(1.05-1.34)	7,65E-03	+-+++	56,0
rs115884658	rs2647044	3,22E-05	0,02	1,30	(1.16-1.46)	6,59E-06	+-+++	41,5
rs115884658	rs9268645	0,05	0,02	1,28	(1.14-1.44)	3,57E-05	+-+++	36,8
rs115884658	rs9272346	0,026	0,02	1,28	(1.14-1.44)	2,46E-05	+-?+++	38,4
rs9268835	-	-	0,29	1,14	(1.09-1.18)	1,368E-11	?+++++	0,0
rs9273401	-	-	0,10	0,85	(0.80-0.89)	1,619E-09	?-----	0,0
rs2647044	-	-	0,12	1,04	(0.97-1.11)	2,497E-01	?---++	30,1
rs9268645	-	-	0,39	0,93	(0.90-0.97)	1,470E-04	?++---	28,2
rs9272346	-	-	0,41	1,06	(1.03-1.10)	2,222E-04	++--++	3,4
rs9268835	rs115884658	0,06	0,29	1,12	(1.08-1.17)	1,234E-09	+++++	0,0
rs9273401	rs115884658	0,11	0,10	0,87	(0.82-0.92)	1,738E-06	-----	0,0
rs2647044	rs115884658	3,22E-05	0,12	1,04	(0.97-1.11)	2,85E-01	---++	16,9
rs9268645	rs115884658	0,05	0,39	0,94	(0.91-0.98)	1,41E-03	+++--	24,0
rs9272346	rs115884658	0,026	0,42	1,05	(1.02-1.09)	3,95E-03	+--++	17,5

rs115884658, *EHMT2* lead variant from 70KforT2D

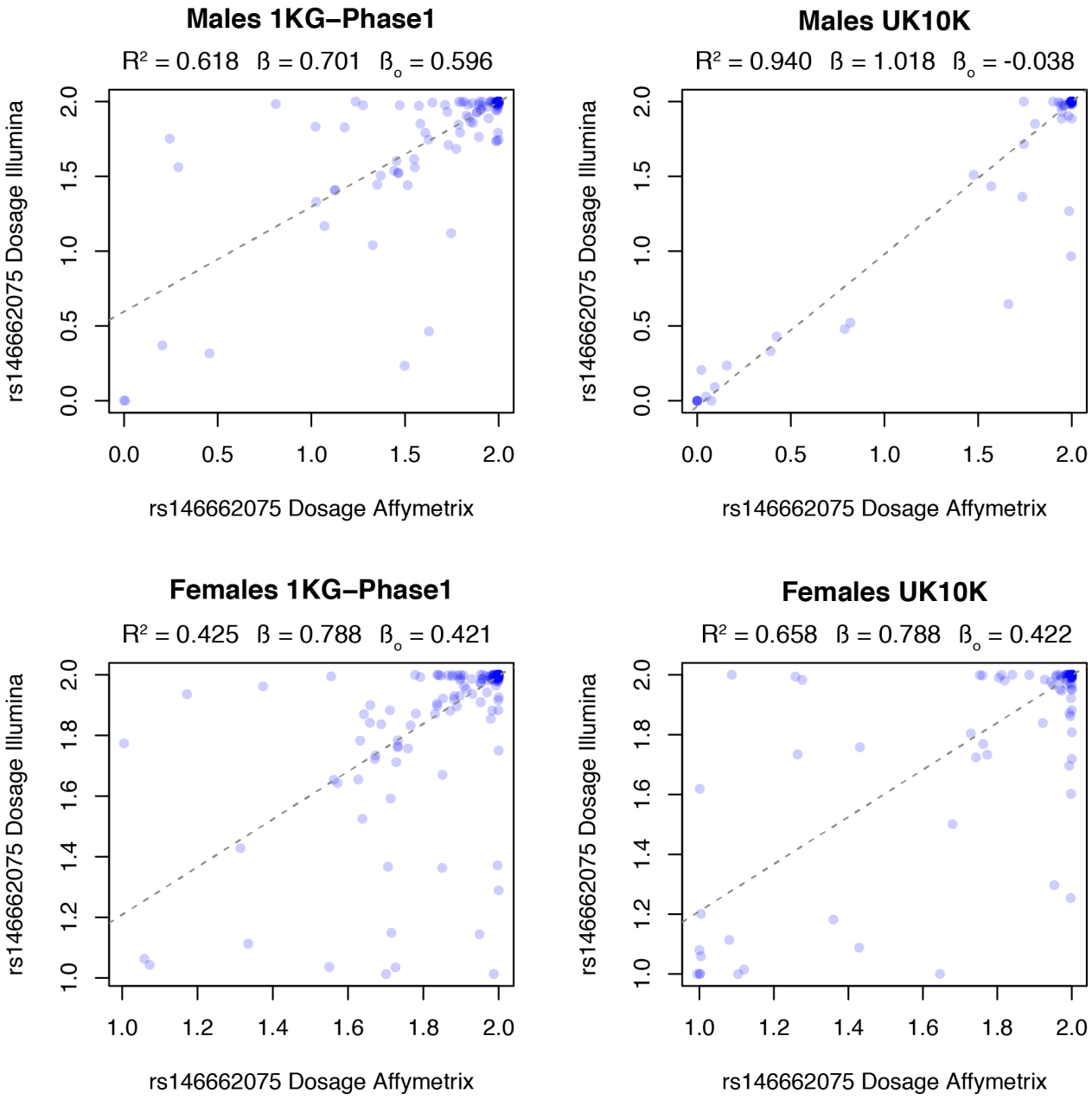
rs9268835, *MHC* T2D *l*ead variant from 70KforT2D

rs9273401, *MHC* T2D lead variant from Cook JP and Morris AP. *EJHG* (2016)

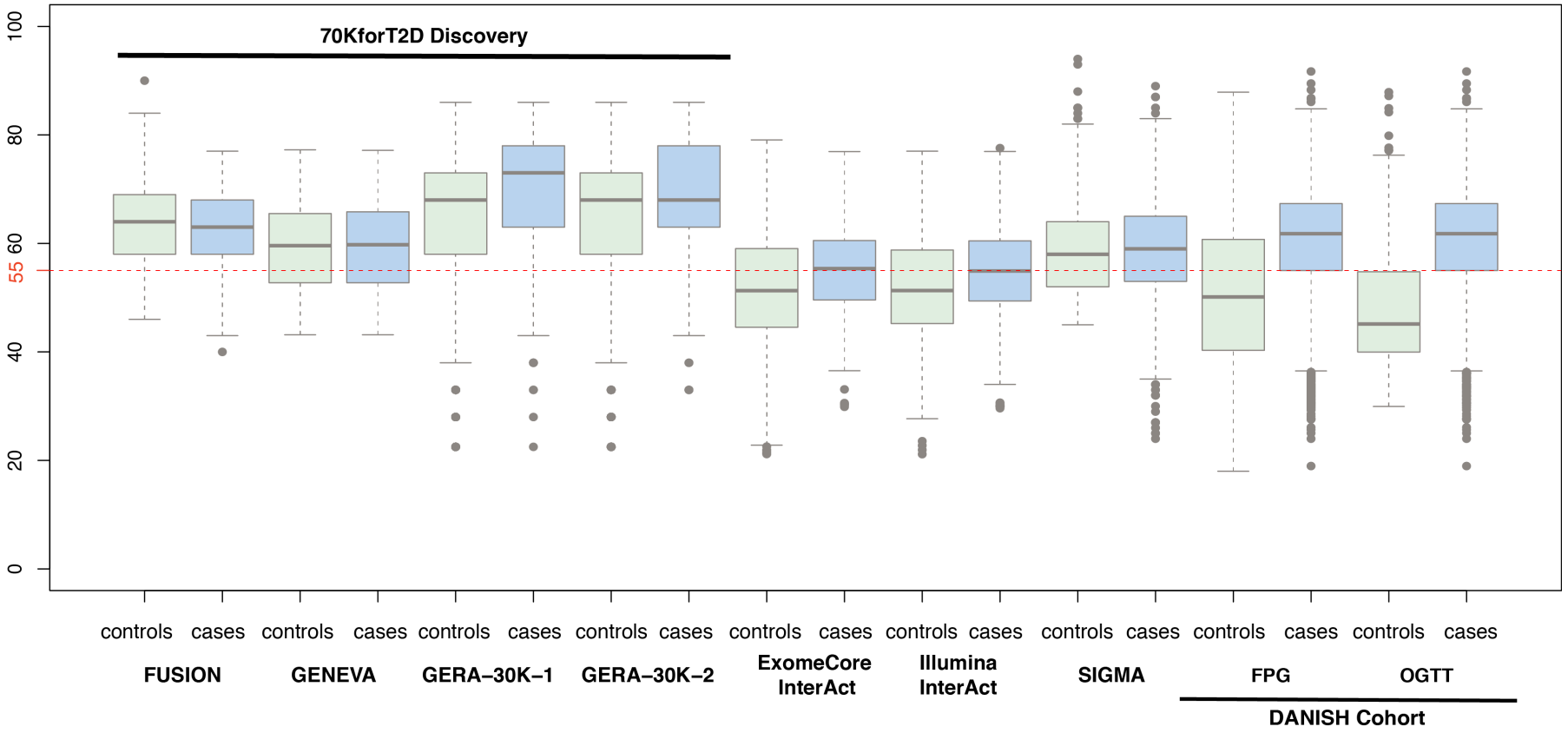
rs2647044, *MHC* T1D lead variant from Hakonarson H. *et al. Nature* (2007)

rs9268645, *MHC* T1D lead variant from Barrett JC, *et al. Nat Genet.* (2009)

rs9272346, *MHC* T1D *l*ead variant from WTCCC *Nature* (2007)



Supplementary Material 9. Comparison of imputation quality across males and females and 1000G-Phase1 and UK10K reference panels. 58C cohort from the WTCCC (~3,000 individuals) that was genotyped by both Affymetrix 6.0 (Affy) and Illumina 1.2M (IL) platforms have been imputed independently. We computed the allelic dosage R^2 coefficient between the dosages from imputing using Affy and Illumina as the backbone for each scenario.



Supplementary Material 10. Boxplot representing the distribution of ages in cases and controls across cohorts. The red line represents 55 years old, which is the average age at onset of T2D in the Danish cohorts.

References

2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**: 661-678.

Abraham G, Inouye M. 2014. Fast principal component analysis of large-scale genome-wide data. *PLoS One* **9**: e93766.

Adams D, Altucci L, Antonarakis SE, Ballesteros J, Beck S, Bird A, Bock C, Boehm B, Campo E, Caricasole A et al. 2012. BLUEPRINT to decode the epigenetic signature written in blood. *Nat Biotechnol* **30**: 224-226.

Agarwala V, Flannick J, Sunyaev S, Go TDC, Altshuler D. 2013. Evaluating empirical bounds on complex disease genetic architecture. *Nat Genet* **45**: 1418-1427.

Almgren P, Lehtovirta M, Isomaa B, Sarelin L, Taskinen MR, Lyssenko V, Tuomi T, Groop L, Botnia Study G. 2011. Heritability and familiality of type 2 diabetes and related quantitative traits in the Botnia Study. *Diabetologia* **54**: 2811-2819.

Altshuler D, Daly MJ, Lander ES. 2008. Genetic mapping in human disease. *Science* **322**: 881-888.

Altshuler D, Hirschhorn JN, Klannemark M, Lindgren CM, Vohl MC, Nemesh J, Lane CR, Schaffner SF, Bolk S, Brewer C et al. 2000. The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nat Genet* **26**: 76-80.

Amberger J, Bocchini CA, Scott AF, Hamosh A. 2009. McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res* **37**: D793-796.

Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. 2015. OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. *Nucleic Acids Res* **43**: D789-798.

Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. 2010. Data quality control in genetic case-control association studies. *Nat Protoc* **5**: 1564-1573.

Arai T, Kano F, Murata M. 2015. Translocation of forkhead box O1 to the nuclear periphery induces histone modifications that regulate transcriptional repression of PCK1 in HepG2 cells. *Genes Cells* **20**: 340-357.

Atkins JL, Whincup PH, Morris RW, Wannamethee SG. 2014. Low muscle mass in older men: the role of lifestyle, diet and cardiovascular risk factors. *J Nutr Health Aging* **18**: 26-33.

Atkinson MA, Eisenbarth GS, Michels AW. 2014. Type 1 diabetes. *Lancet* **383**: 69-82.

Baker M. 2012. Structural variation: the genome's hidden architecture. *Nat Methods* **9**: 133-137.

Balding DJ. 2006. A tutorial on statistical methods for population association studies. *Nat Rev Genet* **7**: 781-791.

Barrett JC, Clayton DG, Concannon P, Akolkar B, Cooper JD, Erlich HA, Julier C, Morahan G, Nerup J, Nierras C et al. 2009a. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat Genet* **41**: 703-707.

Barrett JC, Lee JC, Lees CW, Prescott NJ, Anderson CA, Phillips A, Wesley E, Parnell K, Zhang H, Drummond H et al. 2009c. Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the HNF4A region. *Nat Genet* **41**: 1330-1334.

Bartoli E, Fra GP, Carnevale Schianca GP. 2011. The oral glucose tolerance test (OGTT) revisited. *Eur J Intern Med* **22**: 8-12.

Bateson W, Saunders ER, Punnett MA. 1905. Experimental studies in the physiology of heredity. *Reports to the Evolution Committee of the Royal Society* **2**: 1-55, 80-99.

Bays H, Mandarino L, DeFronzo RA. 2004. Role of the adipocyte, free fatty acids, and ectopic fat in pathogenesis of type 2 diabetes mellitus: peroxisomal proliferator-activated receptor agonists provide a rational therapeutic approach. *J Clin Endocrinol Metab* **89**: 463-478.

Beck A, Isaac R, Lavelin I, Hart Y, Volberg T, Shatz-Azoulay H, Geiger B, Zick Y. 2011. An siRNA screen identifies transmembrane 7 superfamily member 3 (TM7SF3), a seven transmembrane orphan receptor, as an inhibitor of cytokine-induced death of pancreatic beta cells. *Diabetologia* **54**: 2845-2855.

Begum F, Ghosh D, Tseng GC, Feingold E. 2012. Comprehensive literature review and statistical considerations for GWAS meta-analysis. *Nucleic Acids Res* **40**: 3777-3784.

Berndt SI Gustafsson S Magi R Ganna A Wheeler E Feitosa MF Justice AE Monda KL Croteau-Chonka DC Day FR et al. 2013. Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. *Nat Genet* **45**: 501-512.

Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, Kellis M, Marra MA, Beaudet AL, Ecker JR et al. 2010. The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol* **28**: 1045-1048.

Billings LK, Florez JC. 2010. The genetics of type 2 diabetes: what have we learned from GWAS? *Ann N Y Acad Sci* **1212**: 59-77.

Birney E, Soranzo N. 2015. Human genomics: The end of the start for population sequencing. *Nature* **526**: 52-53.

Bonnefond A, Clement N, Fawcett K, Yengo L, Vaillant E, Guillaume JL, Dechaume A, Payne F, Roussel R, Czernichow S et al. 2012. Rare MTNR1B variants impairing melatonin receptor 1B function contribute to type 2 diabetes. *Nat Genet* **44**: 297-301.

Bonnefond A, Froguel P. 2015. Rare and common genetic events in type 2 diabetes: what should biologists know? *Cell Metab* **21**: 357-368.

Bonnelykke K, Sleiman P, Nielsen K, Kreiner-Moller E, Mercader JM, Belgrave D, den Dekker HT, Husby A, Sevelsted A, Faura-Tellez G et al. 2014. A genome-wide association study identifies CDHR3 as a susceptibility locus for early childhood asthma with severe exacerbations. *Nat Genet* **46**: 51-55.
Botstein D, White RL, Skolnick M, Davis RW. 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet* **32**: 314-331.
Bouatia-Naji N, Bonnefond A, Cavalcanti-Proenca C, Sparso T, Holmkvist J, Marchand M, Delplanque J, Lobbens S, Rocheleau G, Durand E et al. 2009. A variant near MTNR1B is associated with increased fasting plasma glucose levels and type 2 diabetes risk. *Nat Genet* **41**: 89-94.
Bowler PJ. 2003. *Evolution: the history of an idea*. Berkeley: University of California Press.
Brissova M, Shostak A, Shiota M, Wiebe PO, Poffenberger G, Kantz J, Chen Z, Carr C, Jerome WG, Chen J et al. 2006. Pancreatic Islet Production of Vascular Endothelial Growth Factor-A Is Essential for Islet Vascularization, Revascularization, and Function. *Diabetes* **55**: 2974.
Browning BL, Browning SR. 2016. Genotype Imputation with Millions of Reference Samples. *Am J Hum Genet* **98**: 116-126.
Bruce CR, Thrush AB, Mertz VA, Bezaire V, Chabowski A, Heigenhauser GJ, Dyck DJ. 2006. Endurance training in obese humans improves glucose tolerance and mitochondrial fatty acid oxidation and alters muscle lipid content. *Am J Physiol Endocrinol Metab* **291**: E99-E107.
Buchanan AV, Weiss KM, Fullerton SM. 2006. Dissecting complex disease: the quest for the Philosopher's Stone? *Int J Epidemiol* **35**: 562-571.
Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Schizophrenia Working Group of the Psychiatric Genomics C, Patterson N, Daly MJ, Price AL, Neale BM. 2015. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**: 291-295.
Cann RL, Stoneking M, Wilson AC. 1987. Mitochondrial DNA and human evolution. *Nature* **325**: 31-36.
Cao W, Collins QF, Becker TC, Robidoux J, Lupo EG, Jr., Xiong Y, Daniel KW, Floering L, Collins S. 2005. p38 Mitogen-activated protein kinase plays a stimulatory role in hepatic gluconeogenesis. *J Biol Chem* **280**: 42731-42737.
Carithers LJ, Moore HM. 2015. The Genotype-Tissue Expression (GTEx) Project. *Biopreserv Biobank* **13**: 307-308.
Carpentier A, Mittelman SD, Bergman RN, Giacca A, Lewis GF. 2000. Prolonged elevation of plasma free fatty acids impairs pancreatic beta-cell function in obese nondiabetic humans but not in individuals with type 2 diabetes. *Diabetes* **49**: 399-408.
Castano L, Eisenbarth GS. 1990. Type-I diabetes: a chronic autoimmune disease of human, mouse, and rat. *Annu Rev Immunol* **8**: 647-679.
Centers for Disease Control and Prevention (CDC). 2014. Centers for Disease Control and Prevention National diabetes statistics report: estimates of diabetes and its burden in the United States. U.S. Department of Health and Human Services, Atlanta, GA.
Chang AM, Halter JB. 2003. Aging and insulin secretion. *Am J Physiol Endocrinol Metab* **284**: E7-12.
Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**: 7.
Chang D, Gao F, Slavney A, Ma L, Waldman YY, Sams AJ, Billing-Ross P, Madar A, Spritz R, Keinan A. 2014. Accounting for eXentricities: analysis of the X chromosome in GWAS reveals X-linked genes implicated in autoimmune diseases. *PLoS One* **9**: e113684.
Chargaff E, Zamenhof S, Green C. 1950. Composition of human desoxypentose nucleic acid. *Nature* **165**: 756-757.
Charlesworth B, Charlesworth D. 2009. Darwin and genetics. *Genetics* **183**: 757-766.
Charlesworth B, Charlesworth D. 2016. Population genetics from 1966 to 2016. *Heredity (Edinb)* doi:10.1038/hdy.2016.55.
Cho YS, Chen CH, Hu C, Long J, Ong RT, Sim X, Takeuchi F, Wu Y, Go MJ, Yamauchi T et al. 2012. Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians. *Nat Genet* **44**: 67-72.
Chong JX, Buckingham KJ, Jhangiani SN, Boehm C, Sobreira N, Smith JD, Harrell TM, McMillin MJ, Wiszniewski W, Gambin T et al. 2015. The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *Am J Hum Genet* **97**: 199-215.
Cobb M. 2006. Heredity before genetics: a history. *Nat Rev Genet* **7**: 953-958.
Collins F. 2010. Has the revolution arrived? *Nature* **464**: 674-675.
Cook JP, Morris AP. 2016. Multi-ethnic genome-wide association study identifies novel locus for type 2 diabetes susceptibility. *Eur J Hum Genet* **24**: 1175-1180.
Cooper DN, Krawczak M, Polychronakos C, Tyler-Smith C, Kehrer-Sawatzki H. 2013. Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. *Hum Genet* **132**: 1077-1130.
Cornell S. 2015. Continual evolution of type 2 diabetes: an update on pathophysiology and emerging treatment options. *Ther Clin Risk Manag* **11**: 621-632.
Crawford DC, Nickerson DA. 2005. Definition and clinical importance of haplotypes. *Annu Rev Med* **56**: 303-320.
Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA et al. 2010. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A* **107**: 21931-21936.
Crick F. 1970. Central dogma of molecular biology. *Nature* **227**: 561-563.

Crow JF. 2010. Wright and Fisher on inbreeding and random drift. *Genetics* **184**: 609-611.
Cusi K, Maezono K, Osman A, Pendergrass M, Patti ME, Pratipanawatr T, DeFronzo RA, Kahn CR, Mandarino LJ. 2000. Insulin resistance differentially affects the PI 3-kinase- and MAP kinase-mediated signaling in human muscle. *J Clin Invest* **105**: 311-320.
Dahm R. 2010. From discovering to understanding. Friedrich Miescher's attempts to uncover the function of DNA. *EMBO Rep* **11**: 153-160.
Dai N, Zhao L, Wrighting D, Kramer D, Majithia A, Wang Y, Cracan V, Borges-Rivera D, Mootha VK, Nahrendorf M et al. 2015. IGF2BP2/IMP2-Deficient mice resist obesity through enhanced translation of Ucp1 mRNA and Other mRNAs encoding mitochondrial proteins. *Cell Metab* **21**: 609-621.
Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES. 2001. High-resolution haplotype structure in the human genome. *Nat Genet* **29**: 229-232.
Darwin CR. 1859. *The Origin of Species*, London.
Darwin CR. 1868. *Variation of Animals and Plants Under Domestication*, London.
Das S, Forer L, Schonherr S, Sidore C, Locke AE, Kwong A, Vrieze SI, Chew EY, Levy S, McGue M et al. 2016. Next-generation genotype imputation service and methods. *Nat Genet* doi:10.1038/ng.3656.
Dastani Z Hivert MF Timpson N Perry JR Yuan X Scott RA Henneman P Heid IM Kizer JR Lyytikainen LP et al. 2012. Novel loci for adiponectin levels and their influence on type 2 diabetes and metabolic traits: a multi-ethnic meta-analysis of 45,891 individuals. *PLoS Genet* **8**: e1002607.
de Bakker PI, Ferreira MA, Jia X, Neale BM, Raychaudhuri S, Voight BF. 2008. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum Mol Genet* **17**: R122-128.
De Vries H. 1889. *Intracellulare Pangenesis*. Open Court Publishing Co., Chicago.
De Vries H. 1901-1903. *Die Mutationstheorie*. Viet à Co., Leipzig.
de Vries PS Chasman DI Sabater-Lleal M Chen MH Huffman JE Steri M Tang W Teumer A Marioni RE Grossmann V et al. 2016. A meta-analysis of 120 246 individuals identifies 18 new loci for fibrinogen concentration. *Hum Mol Genet* **25**: 358-370.
Defronzo RA. 2009. Banting Lecture. From the triumvirate to the ominous octet: a new paradigm for the treatment of type 2 diabetes mellitus. *Diabetes* **58**: 773-795.
DeFronzo RA, Hompesch M, Kasichayanula S, Liu X, Hong Y, Pfister M, Morrow LA, Leslie BR, Boulton DW, Ching A et al. 2013. Characterization of renal glucose reabsorption in response to dapagliflozin in healthy subjects and subjects with type 2 diabetes. *Diabetes Care* **36**: 3169-3176.
del Bosque-Plata L, Lin J, Horikawa Y, Schwarz PE, Cox NJ, Iwasaki N, Ogata M, Iwamoto Y, German MS, Bell GI. 2001. Mutations in the coding region of the neurogenin 3 gene (NEUROG3) are not a common cause of maturity-onset diabetes of the young in Japanese subjects. *Diabetes* **50**: 694-696.
Delaneau O, Marchini J, Zagury JF. 2012. A linear complexity phasing method for thousands of genomes. *Nat Methods* **9**: 179-181.
Delaneau O, Zagury JF, Marchini J. 2013. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods* **10**: 5-6.
Diabetes Genetics Initiative of Broad Institute of H, Mit LU, Novartis Institutes of BioMedical R, Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PI, Chen H, Roix JJ et al. 2007. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* **316**: 1331-1336.
DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium Asian Genetic Epidemiology Network Type 2 Diabetes C South Asian Type 2 Diabetes C Mexican American Type 2 Diabetes C Type 2 Diabetes Genetic Exploration by Nex-generation sequencing in muylti-Ethnic Samples C Mahajan A Go MJ Zhang W Below JE Gaulton KJ et al. 2014. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat Genet* **46**: 234-244.
Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB. 2010. Rare variants create synthetic genome-wide associations. *PLoS Biol* **8**: e1000294.
Dimas AS, Stranger BE, Beazley C, Finn RD, Ingle CE, Forrest MS, Ritchie ME, Deloukas P, Tavaré S, Dermitzakis ET. 2008. Modifier Effects between Regulatory and Protein-Coding Variation. *PLoS Genetics* **4**: e1000244.
Dobbins RL, Szczepaniak LS, Bentley B, Esser V, Myhill J, McGarry JD. 2001. Prolonged inhibition of muscle carnitine palmitoyltransferase-1 promotes intramyocellular lipid accumulation and insulin resistance in rats. *Diabetes* **50**: 123-130.
Dupuis J Langenberg C Prokopenko I Saxena R Soranzo N Jackson AU Wheeler E Glazer NL Bouatia-Naji N Gloyn AL et al. 2010. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat Genet* **42**: 105-116.
Eaton S. 2002. Control of mitochondrial beta-oxidation flux. *Prog Lipid Res* **41**: 197-239.
Encode Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57-74.
Encode Project Consortium Birney E Stamatoyannopoulos JA Dutta A Guigo R Gingeras TR Margulies EH Weng Z Snyder M Dermitzakis ET et al. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**: 799-816.
European Environment Agency (EEA). 2015. Changing disease burdens and risks of pandemics (GMT 3). In *The European environment — state and outlook 2015 (SOER 2015)*. European Environment Agency.
Evangelou E, Ioannidis JP. 2013. Meta-analysis methods for genome-wide association studies and beyond. *Nat Rev Genet* **14**: 379-389.

Facchini FS, Hua N, Abbasi F, Reaven GM. 2001. Insulin resistance as a predictor of age-related diseases. *J Clin Endocrinol Metab* **86**: 3574-3578.
Fadista J, Manning AK, Florez JC, Groop L. 2016. The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. *Eur J Hum Genet* **24**: 1202-1205.
Fadista J, Vikman P, Laakso EO, Mollet IG, Esguerra JL, Taneera J, Storm P, Osmark P, Ladenvall C, Prasad RB et al. 2014. Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism. *Proc Natl Acad Sci U S A* **111**: 13924-13929.
Fajans SS, Bell GI, Polonsky KS. 2001. Molecular mechanisms and clinical pathophysiology of maturity-onset diabetes of the young. *N Engl J Med* **345**: 971-980.
Fisher RA. 1918. The correlation between relatives on the supposition of Mendelian inheritance. *Trans R Soc Edinb* **52**: 35.
Fisher RA. 1930. *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford.
Flannick J, Beer NL, Bick AG, Agarwala V, Molnes J, Gupta N, Burtt NP, Florez JC, Meigs JB, Taylor H et al. 2013. Assessing the phenotypic effects in the general population of rare variants in genes for a dominant Mendelian form of diabetes. *Nat Genet* **45**: 1380-1385.
Flannick J, Florez JC. 2016. Type 2 diabetes: genetic data sharing to advance complex disease research. *Nat Rev Genet* doi:10.1038/nrg.2016.56.
Flannick J, Johansson S, Njolstad PR. 2016. Common and rare forms of diabetes mellitus: towards a continuum of diabetes subtypes. *Nat Rev Endocrinol* **12**: 394-406.
Flannick J, Thorleifsson G, Beer NL, Jacobs SB, Grarup N, Burtt NP, Mahajan A, Fuchsberger C, Atzmon G, Benediktsson R et al. 2014. Loss-of-function mutations in SLC30A8 protect against type 2 diabetes. *Nat Genet* **46**: 357-363.
Flemming W. 1965. Contributions to the Knowledge of the Cell and Its Vital Processes. *J Cell Biol* **25**: 3-69.
Forbes JM, Cooper ME. 2013. Mechanisms of diabetic complications. *Physiol Rev* **93**: 137-188.
Fox CS, Liu Y, White CC, Feitosa M, Smith AV, Heard-Costa N, Lohman K, Consortium G, Consortium M, Consortium G et al. 2012. Genome-wide association for abdominal subcutaneous and visceral adipose reveals a novel locus for visceral fat in women. *PLoS Genet* **8**: e1002695.
Frayling TM, Walker M, McCarthy MI, Evans JC, Allen LI, Lynn S, Ayres S, Millauer B, Turner C, Turner RC et al. 1999. Parent-offspring trios: a resource to facilitate the identification of type 2 diabetes genes. *Diabetes* **48**: 2475-2479.
Frazer KA, Murray SS, Schork NJ, Topol EJ. 2009. Human genetic variation and its contribution to complex traits. *Nat Rev Genet* **10**: 241-251.
Frey BJ, Dueck D. 2007. Clustering by passing messages between data points. *Science* **315**: 972-976.
Frojdo S, Vidal H, Pirola L. 2009. Alterations of insulin signaling in type 2 diabetes: a review of the current evidence from humans. *Biochim Biophys Acta* **1792**: 83-92.
Fuchsberger C Flannick J Teslovich TM Mahajan A Agarwala V Gaulton KJ Ma C Fontanillas P Moutsianas L McCarthy DJ et al. 2016. The genetic architecture of type 2 diabetes. *Nature* doi:10.1038/nature18642.
Garcia ML, Kaczorowski GJ. 2014. Targeting the inward-rectifier potassium channel ROMK in cardiovascular disease. *Curr Opin Pharmacol* **15**: 1-6.
Gelernter J, Kranzler HR, Sherva R, Almasy L, Koesterer R, Smith AH, Anton R, Preuss UW, Ridinger M, Rujescu D et al. 2014. Genome-wide association study of alcohol dependence:significant findings in African- and European-Americans including novel risk loci. *Mol Psychiatry* **19**: 41-49.
Genome of the Netherlands C. 2014. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet* **46**: 818-825.
Gibson G. 2011. Rare and common variants: twenty arguments. *Nat Rev Genet* **13**: 135-145.
Gloyn AL, Weedon MN, Owen KR, Turner MJ, Knight BA, Hitman G, Walker M, Levy JC, Sampson M, Halford S et al. 2003. Large-scale association studies of variants in genes encoding the pancreatic beta-cell KATP channel subunits Kir6.2 (KCNJ11) and SUR1 (ABCC8) confirm that the KCNJ11 E23K variant is associated with type 2 diabetes. *Diabetes* **52**: 568-572.
Gradwohl G, Dierich A, LeMeur M, Guillemot F. 2000. neurogenin3 is required for the development of the four endocrine cell lineages of the pancreas. *Proc Natl Acad Sci U S A* **97**: 1607-1611.
Grant RW, Moore AF, Florez JC. 2009. Genetic architecture of type 2 diabetes: recent progress and clinical implications. *Diabetes Care* **32**: 1107-1114.
Grant SF, Thorleifsson G, Reynisdottir I, Benediktsson R, Manolescu A, Sainz J, Helgason A, Stefansson H, Emilsson V, Helgadottir A et al. 2006. Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nat Genet* **38**: 320-323.
Griffiths AJF, Miller JH, Suzuki DT, al. e. 2000a. The discovery of linkage. In *Introduction to Genetic Analysis*. W. H. Freeman, New York.
Griffiths AJF, Miller JH, Suzuki DT, al. e. 2000b. Introduction to Genetic Analysis. In *DNA: The genetic material*. W. H. Freeman, New York.
Griffiths AJF, Miller JH, Suzuki DT, al. e. 2000c. Structure of DNA. In *Introduction to Genetic Analysis*. W. H. Freeman, New York.
Gritti M, Wurth R, Angelini M, Barbieri F, Peretti M, Pizzi E, Pattarozzi A, Carra E, Sirito R, Daga A et al. 2014. Metformin repositioning as antitumoral agent: selective antiproliferative effects in human glioblastoma stem cells, via inhibition of CLIC1-mediated ion current. *Oncotarget* **5**: 11252-11268.

Groves CJ, Zeggini E, Minton J, Frayling TM, Weedon MN, Rayner NW, Hitman GA, Walker M, Wiltshire S, Hattersley AT et al. 2006. Association analysis of 6,736 U.K. subjects provides replication and confirms TCF7L2 as a type 2 diabetes susceptibility gene with a substantial effect on individual risk. *Diabetes* **55**: 2640-2644.
Grundberg E, Small KS, Hedman AK, Nica AC, Buil A, Keildson S, Bell JT, Yang TP, Meduri E, Barrett A et al. 2012. Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat Genet* **44**: 1084-1089.
GTEx Consortium. 2013. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**: 580-585.
Gudbjartsson DF, Helgason H, Gudjonsson SA, Zink F, Oddson A, Gylfason A, Besenbacher S, Magnusson G, Halldorsson BV, Hjartarson E et al. 2015. Large-scale whole-genome sequencing of the Icelandic population. *Nat Genet* **47**: 435-444.
Gusella JF, Wexler NS, Conneally PM, Naylor SL, Anderson MA, Tanzi RE, Watkins PC, Ottina K, Wallace MR, Sakaguchi AY et al. 1983. A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* **306**: 234-238.
Gyorffy B, Hatzis C, Sanft T, Hofstatter E, Aktas B, Pusztai L. 2015. Multigene prognostic tests in breast cancer: past, present, future. *Breast Cancer Res* **17**: 11.
Hakonarson H, Grant SF, Bradfield JP, Marchand L, Kim CE, Glessner JT, Grabs R, Casalunovo T, Taback SP, Frackelton EC et al. 2007. A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene. *Nature* **448**: 591-594.
Haldane JBS. 1932. *The Causes of Natural Selection*. Longmans Green, London.
Hamza TH, Zabetian CP, Tenesa A, Laederach A, Montimurro J, Yearout D, Kay DM, Doheny KF, Paschall J, Pugh E et al. 2010. Common genetic variation in the HLA region is associated with late-onset sporadic Parkinson's disease. *Nat Genet* **42**: 781-785.
Hara K, Fujita H, Johnson TA, Yamauchi T, Yasuda K, Horikoshi M, Peng C, Hu C, Ma RC, Imamura M et al. 2014. Genome-wide association study identifies three novel loci for type 2 diabetes. *Hum Mol Genet* **23**: 239-246.
Hardy GH. 1908. Mendelian Proportions in a Mixed Population. *Science* **28**: 49-50.
Hayden EC. 2014. Is the $1,000 genome for real? , doi:10.1038/nature.2014.14530. Nature Publishing Group, Nature News.
Haynes RH. 1998. Heritable variation and mutagenesis at early International Congresses of Genetics. *Genetics* **148**: 1419-1431.
Health GADoCDo. 2010. *Understanding Genetics: A District of Columbia Guide for Patients and Health Professionals*. Genetic Alliance, Washington (DC).
Hehir-Kwa JY, Marschall T, Kloosterman WP, Francioli LC, Baaijens JA, Dijkstra LJ, Abdellaoui A, Koval V, Thung DT, Wardenaar R et al. 2016. A high-quality human reference panel reveals the complexity and distribution of genomic structural variants. *Nat Commun* **7**: 12989.
Hershey AD, Chase M. 1952. Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J Gen Physiol* **36**: 39-56.
Higgins JP, Thompson SG. 2002. Quantifying heterogeneity in a meta-analysis. *Stat Med* **21**: 1539-1558.
Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* **106**: 9362-9367.
Hinrichs AL, Larkin EK, Suarez BK. 2009. Population stratification and patterns of linkage disequilibrium. *Genet Epidemiol* **33 Suppl 1**: S88-92.
Hirschhorn JN, Daly MJ. 2005. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* **6**: 95-108.
Ho LA, Lange EM. 2010. Using public control genotype data to increase power and decrease cost of case-control genetic association studies. *Hum Genet* **128**: 597-608.
Hofker MH, Fu J, Wijmenga C. 2014. The genome revolution and its role in understanding complex diseases. *Biochim Biophys Acta* **1842**: 1889-1895.
Holland WL, Knotts TA, Chavez JA, Wang LP, Hoehn KL, Summers SA. 2007. Lipid mediators of insulin resistance. *Nutr Rev* **65**: S39-46.
Hood L, Rowen L. 2013. The Human Genome Project: big science transforms biology and medicine. *Genome Med* **5**: 79.
Horikoshi M Beaumont RN Day FR Warrington NM Kooijman MN Fernandez-Tajes J Feenstra B van Zuydam NR Gaulton KJ Grarup N et al. 2016. Genome-wide associations for birth weight and correlations with adult disease. *Nature* **538**: 248-252.
Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. 2012. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet* **44**: 955-959.
Howie B, Marchini J, Stephens M. 2011. Genotype imputation with thousands of genomes. *G3 (Bethesda)* **1**: 457-470.
Howie BN, Donnelly P, Marchini J. 2009. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**: e1000529.
Hu FB. 2011. Globalization of diabetes: the role of diet, lifestyle, and genes. *Diabetes Care* **34**: 1249-1257.
Huang GH, Tseng YC. 2014. Genotype imputation accuracy with different reference panels in admixed populations. *BMC Proc* **8**: S64.
Huang J, Howie B, McCarthy S, Memari Y, Walter K, Min JL, Danecek P, Malerba G, Trabetti E, Zheng HF et al. 2015. Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat Commun* **6**: 8111.

Huang S, Czech MP. 2007. The GLUT4 glucose transporter. *Cell Metab* **5**: 237-252.
Hutchison CA, 3rd. 2007. DNA sequencing: bench to bedside and beyond. *Nucleic Acids Res* **35**: 6227-6237.
Innocenti F, Cox NJ, Dolan ME. 2011. The use of genomic information to optimize cancer chemotherapy. *Semin Oncol* **38**: 186-195.
International Diabetes Federation. 2015. IDF Diabetes Atlas, 7th edn. International Diabetes Federation, Brussels, Belgium.
International Expert C. 2009. International Expert Committee report on the role of the A1C assay in the diagnosis of diabetes. *Diabetes Care* **32**: 1327-1334.
International HapMap C. 2003. The International HapMap Project. *Nature* **426**: 789-796.
International HapMap C Frazer KA Ballinger DG Cox DR Hinds DA Stuve LL Gibbs RA Belmont JW Boudreau A Hardenbol P et al. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**: 851-861.
Ismail-Beigi F, Craven T, Banerji MA, Basile J, Calles J, Cohen RM, Cuddihy R, Cushman WC, Genuth S, Grimm RH, Jr. et al. 2010. Effect of intensive treatment of hyperglycaemia on microvascular outcomes in type 2 diabetes: an analysis of the ACCORD randomised trial. *Lancet* **376**: 419-430.
Janssens AC, van Duijn CM. 2008. Genome-based prediction of common diseases: advances and prospects. *Hum Mol Genet* **17**: R166-173.
Jeffreys AJ. 1979. DNA sequence variants in the G gamma-, A gamma-, delta- and beta-globin genes of man. *Cell* **18**: 1-10.
Jette M, Yoo A, Grondona M. 2002. SLURM: Simple Linux Utility for Resource Management. In *In Lecture Notes in Computer Science: Proceedings of Job Scheduling Strategies for Parallel Processing (JSSPP) 2003*, doi:citeulike-article-id:10601753, pp. 44-60. Springer-Verlag.
Jimenez-Chillaron JC, Ramon-Krauel M, Ribo S, Diaz R. 2016. Transgenerational epigenetic inheritance of diabetes risk as a consequence of early nutritional imbalances. *Proceedings of the Nutrition Society* **75**: 78-89.
Johnson EO, Hancock DB, Levy JL, Gaddis NC, Saccone NL, Bierut LJ, Page GP. 2013. Imputation across genotyping arrays for genome-wide association studies: assessment of bias and a correction strategy. *Hum Genet* **132**: 509-522.
Jorde LB, Wooding SP. 2004. Genetic variation, classification and 'race'. *Nat Genet* **36**: S28-33.
Jorgensen T, Borch-Johnsen K, Thomsen TF, Ibsen H, Glumer C, Pisinger C. 2003. A randomized non-pharmacological intervention study for prevention of ischaemic heart disease: baseline results Inter99. *Eur J Cardiovasc Prev Rehabil* **10**: 377-386.
Kahn SE. 2001. Clinical review 135: The importance of beta-cell failure in the development and progression of type 2 diabetes. *J Clin Endocrinol Metab* **86**: 4047-4058.
Kashyap S, Belfort R, Gastaldelli A, Pratipanawatr T, Berria R, Pratipanawatr W, Bajaj M, Mandarino L, DeFronzo R, Cusi K. 2003. A sustained increase in plasma free fatty acids impairs insulin secretion in nondiabetic subjects genetically predisposed to develop type 2 diabetes. *Diabetes* **52**: 2461-2474.
Kaye J, Heeney C, Hawkins N, de Vries J, Boddington P. 2009. Data sharing in genomics--re-shaping scientific practice. *Nat Rev Genet* **10**: 331-335.
Kerem B, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, Chakravarti A, Buchwald M, Tsui LC. 1989. Identification of the cystic fibrosis gene: genetic analysis. *Science* **245**: 1073-1080.
Kim S, Soltani-Bejnood M, Quignard-Boulange A, Massiera F, Teboul M, Ailhaud G, Kim JH, Moustaid-Moussa N, Voy BH. 2006. The adipose renin-angiotensin system modulates systemic markers of insulin sensitivity and activates the intrarenal renin-angiotensin system. *J Biomed Biotechnol* **2006**: 27012.
Kim YJ, Lee J, Kim BJ, Consortium TD-G, Park T. 2015. A new strategy for enhancing imputation quality of rare variants from next-generation sequencing data via combining SNP and exome chip data. *BMC Genomics* **16**: 1109.
Kimura M, Crow JF. 1964. The Number of Alleles That Can Be Maintained in a Finite Population. *Genetics* **49**: 725-738.
Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**: 310-315.
Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST et al. 2005. Complement factor H polymorphism in age-related macular degeneration. *Science* **308**: 385-389.
Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER. 2013. The next-generation sequencing revolution and its impact on genomics. *Cell* **155**: 27-38.
Konig IR, Loley C, Erdmann J, Ziegler A. 2014. How to include chromosome X in your genome-wide association study. *Genet Epidemiol* **38**: 97-103.
Koves TR, Ussher JR, Noland RC, Slentz D, Mosedale M, Ilkayeva O, Bain J, Stevens R, Dyck JR, Newgard CB et al. 2008. Mitochondrial overload and incomplete fatty acid oxidation contribute to skeletal muscle insulin resistance. *Cell Metab* **7**: 45-56.
Krssak M, Falk Petersen K, Dresner A, DiPietro L, Vogel SM, Rothman DL, Roden M, Shulman GI. 1999. Intramyocellular lipid concentrations are correlated with insulin sensitivity in humans: a 1H NMR spectroscopy study. *Diabetologia* **42**: 113-116.
Ku CS, Loy EY, Salim A, Pawitan Y, Chia KS. 2010. The discovery of human genetic variations and their use as disease markers: past, present and future. *J Hum Genet* **55**: 403-415.

Kukurba KR, Parsana P, Balliu B, Smith KS, Zappala Z, Knowles DA, Fave MJ, Davis JR, Li X, Zhu X et al. 2016. Impact of the X Chromosome and sex on regulatory variation. *Genome Res* **26**: 768-777.
Kvale MN, Hesselson S, Hoffmann TJ, Cao Y, Chan D, Connell S, Croen LA, Dispensa BP, Eshragh J, Finn A et al. 2015. Genotyping Informatics and Quality Control for 100,000 Subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort. *Genetics* **200**: 1051-1060.
Lachance J, Tishkoff SA. 2013. Population Genomics of Human Adaptation. *Annu Rev Ecol Evol Syst* **44**: 123-143.
Lander ES Linton LM Birren B Nusbaum C Zody MC Baldwin J Devon K Dewar K Doyle M FitzHugh W et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860-921.
Lane JM, Vlasac I, Anderson SG, Kyle SD, Dixon WG, Bechtold DA, Gill S, Little MA, Luik A, Loudon A et al. 2016. Genome-wide association analysis identifies novel loci for chronotype in 100,420 individuals from the UK Biobank. *Nature Communications* **7**: 10889.
Langenberg C, Sharp SJ, Franks PW, Scott RA, Deloukas P, Forouhi NG, Froguel P, Groop LC, Hansen T, Palla L et al. 2014. Gene-lifestyle interaction and type 2 diabetes: the EPIC interact case-cohort study. *PLoS Med* **11**: e1001647.
Lappalainen I, Almeida-King J, Kumanduri V, Senf A, Spalding JD, Ur-Rehman S, Saunders G, Kandasamy J, Caccamo M, Leinonen R et al. 2015. The European Genome-phenome Archive of human data consented for biomedical research. *Nat Genet* **47**: 692-695.
Lappalainen T, Montgomery Stephen B, Nica Alexandra C, Dermitzakis Emmanouil T. 2011. Epistatic Selection between Coding and Regulatory Variation in Human Evolution and Disease. *American Journal of Human Genetics* **89**: 459-463.
Lee S, Abecasis GR, Boehnke M, Lin X. 2014. Rare-variant association analysis: study designs and statistical tests. *Am J Hum Genet* **95**: 5-23.
Lei X, Callaway M, Zhou H, Yang Y, Chen W. 2015. Obesity associated Lyplal1 gene is regulated in diet induced obesity but not required for adipocyte differentiation. *Molecular and cellular endocrinology* **411**: 207-213.
Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB et al. 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**: 285-291.
Lenay C. 2000. Hugo De Vries: from the theory of intracellular pangenesis to the rediscovery of Mendel. *C R Acad Sci III* **323**: 1053-1060.
Leon-Latre M, Moreno-Franco B, Andres-Esteban EM, Ledesma M, Laclaustra M, Alcalde V, Penalvo JL, Ordovas JM, Casasnovas JA, Aragon Workers' Health Study i. 2014. Sedentary lifestyle and its relation to cardiovascular risk factors, insulin resistance and inflammatory profile. *Rev Esp Cardiol (Engl Ed)* **67**: 449-455.
Levy MA, Lovly CM, Pao W. 2012. Translating genomic information into clinical medicine: lung cancer as a paradigm. *Genome Res* **22**: 2101-2108.
Lewontin RC. 1964. The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. *Genetics* **49**: 49-67.
Li J, Guo YF, Pei Y, Deng HW. 2012. The impact of imputation on meta-analysis of genome-wide association studies. *PLoS One* **7**: e34486.
Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. 2010. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol* **34**: 816-834.
Lin DY, Sullivan PF. 2009. Meta-analysis of genome-wide association studies with overlapping subjects. *Am J Hum Genet* **85**: 862-872.
Liu Y. 2007. Like father like son. A fresh review of the inheritance of acquired characteristics. *EMBO Rep* **8**: 798-803.
Liu Y, Li X. 2012. Darwin's Pangenesis and molecular medicine. *Trends Mol Med* **18**: 506-508.
Lobo I, Shaw K. 2008. Discovery and Types of Genetic Linkage. *Nature Education* **1**.
Loh PR, Palamara PF, Price AL. 2016. Fast and accurate long-range phasing in a UK Biobank cohort. *Nat Genet* **48**: 811-816.
Lordan F, Tejedor E, Ejarque J, Rafanell R, Álvarez J, Marozzo F, Lezzi D, Sirvent R, Talia D, Badia RM. 2014. ServiceSs: An Interoperable Programming Framework for the Cloud. *Journal of Grid Computing* **12**: 67-91.
Lowe WL, Jr., Reddy TE. 2015. Genomic approaches for understanding the genetics of complex disease. *Genome Res* **25**: 1432-1441.
Lupski JR, Belmont JW, Boerwinkle E, Gibbs RA. 2011. Clan genomics and the complex architecture of human disease. *Cell* **147**: 32-43.
Lyssenko V, Eliasson L, Kotova O, Pilgaard K, Wierup N, Salehi A, Wendt A, Jonsson A, De Marinis YZ, Berglund LM et al. 2011. Pleiotropic effects of GIP on islet function involve osteopontin. *Diabetes* **60**: 2424-2433.
Lyssenko V, Nagorny CL, Erdos MR, Wierup N, Jonsson A, Spegel P, Bugliani M, Saxena R, Fex M, Pulizzi N et al. 2009. Common variant in MTNR1B associated with increased risk of type 2 diabetes and impaired early insulin secretion. *Nat Genet* **41**: 82-88.
MacDonald ME, Ambrose CM, Duyao MP, Myers RH, Lin C, Srinidhi L, Barnes G, Taylor SA, James M, Groot N et al. 1993. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72**: 971-983.
Mahajan A, Sim X, Ng HJ, Manning A, Rivas MA, Highland HM, Locke AE, Grarup N, Im HK, Cingolani P et al. 2015. Identification and functional characterization of G6PC2 coding variants influencing glycemic traits define an effector transcript at the G6PC2-ABCB11 locus. *PLoS Genet* **11**: e1004876.

Majithia AR, Flannick J, Shahinian P, Guo M, Bray MA, Fontanillas P, Gabriel SB, Go TDC, Project NJFAS, Consortium STD et al. 2014. Rare variants in PPARG with decreased activity in adipocyte differentiation are associated with increased risk of type 2 diabetes. *Proc Natl Acad Sci U S A* **111**: 13127-13132.
Manning AK Hivert MF Scott RA Grimsby JL Bouatia-Naji N Chen H Rybin D Liu CT Bielak LF Prokopenko I et al. 2012. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nat Genet* **44**: 659-669.
Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A et al. 2009. Finding the missing heritability of complex diseases. *Nature* **461**: 747-753.
Marchini J, Howie B. 2010. Genotype imputation for genome-wide association studies. *Nat Rev Genet* **11**: 499-511.
Marchini J, Howie B, Myers S, McVean G, Donnelly P. 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* **39**: 906-913.
Marigorta UM, Navarro A. 2013. High trans-ethnic replicability of GWAS results implies common causal variants. *PLoS Genet* **9**: e1003566.
McCarthy S Das S Kretzschmar W Delaneau O Wood AR Teumer A Kang HM Fuchsberger C Danecek P Sharp K et al. 2016. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* doi:10.1038/ng.3643.
McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. 2010. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**: 2069-2070.
Mele M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, Young TR, Goldmann JM, Pervouchine DD, Sullivan TJ et al. 2015. Human genomics. The human transcriptome across tissues and individuals. *Science* **348**: 660-665.
Mercader JM, Puiggros M, Segre AV, Planet E, Sorianello E, Sebastian D, Rodriguez-Cuenca S, Ribas V, Bonas-Guarch S, Draghici S et al. 2012. Identification of novel type 2 diabetes candidate genes involved in the crosstalk between the mitochondrial and the insulin signaling systems. *PLoS Genet* **8**: e1003046.
Meyer C, Stumvoll M, Nadkarni V, Dostou J, Mitrakou A, Gerich J. 1998. Abnormal renal and hepatic glucose metabolism in type 2 diabetes mellitus. *J Clin Invest* **102**: 619-624.
Michalakis K, Goulis DG, Vazaiou A, Mintziori G, Polymeris A, Abrahamian-Michalakis A. 2013. Obesity in the ageing man. *Metabolism* **62**: 1341-1349.
Mihalik SJ, Goodpaster BH, Kelley DE, Chace DH, Vockley J, Toledo FG, DeLany JP. 2010. Increased levels of plasma acylcarnitines in obesity and type 2 diabetes and identification of a marker of glucolipotoxicity. *Obesity (Silver Spring)* **18**: 1695-1700.
Mohlke KL, Boehnke M. 2015. Recent advances in understanding the genetic architecture of type 2 diabetes. *Hum Mol Genet* **24**: R85-92.
Moltke I, Grarup N, Jorgensen ME, Bjerregaard P, Treebak JT, Fumagalli M, Korneliussen TS, Andersen MA, Nielsen TS, Krarup NT et al. 2014. A common Greenlandic TBC1D4 variant confers muscle insulin resistance and type 2 diabetes. *Nature* **512**: 190-193.
Morgan TH. 1910. Sex Limited Inheritance in Drosophila. *Science* **32**: 120-122.
Morgan TH. 1911. Random Segregation Versus Coupling in Mendelian Inheritance. *Science* **34**: 384.
Morley JE. 2010. Nutrition and the aging male. *Clin Geriatr Med* **26**: 287-299.
Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, Cheung VG. 2004. Genetic analysis of genome-wide variation in human gene expression. *Nature* **430**: 743-747.
Morris AP. 2014. Fine mapping of type 2 diabetes susceptibility loci. *Curr Diab Rep* **14**: 549.
Morris AP Voight BF Teslovich TM Ferreira T Segre AV Steinthorsdottir V Strawbridge RJ Khan H Grallert H Mahajan A et al. 2012. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat Genet* **44**: 981-990.
Mosca R, Tenorio-Laranga J, Olivella R, Alcalde V, Ceol A, Soler-Lopez M, Aloy P. 2015. dSysMap: exploring the edgetic role of disease mutations. *Nat Methods* **12**: 167-168.
Muoio DM, Neufer PD. 2012. Lipid-induced mitochondrial stress and insulin action in muscle. *Cell Metab* **15**: 595-605.
National Cholesterol Education Program Expert Panel on Detection E, Treatment of High Blood Cholesterol in A. 2002. Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report. *Circulation* **106**: 3143-3421.
National Institutes of Health (US). 2007. Understanding Human Genetic Variation. In *NIH Curriculum Supplement Series [Internet]*, Vol Biological Sciences Curriculum Study. Bethesda (MD): National Institutes of Health (US).
Nauck MA. 2014. Update on developments with SGLT2 inhibitors in the management of type 2 diabetes. *Drug Des Devel Ther* **8**: 1335-1380.
Nauck MA, Baller B, Meier JJ. 2004. Gastric inhibitory polypeptide and glucagon-like peptide-1 in the pathogenesis of type 2 diabetes. *Diabetes* **53 Suppl 3**: S190-196.
Naylor RN, Greeley SA, Bell GI, Philipson LH. 2011. Genetics and pathophysiology of neonatal diabetes mellitus. *J Diabetes Investig* **2**: 158-169.
Ng MC, Shriner D, Chen BH, Li J, Chen WM, Guo X, Liu J, Bielinski SJ, Yanek LR, Nalls MA et al. 2014. Meta-analysis of genome-wide association studies in African Americans provides insights into the genetic architecture of type 2 diabetes. *PLoS Genet* **10**: e1004517.
Nica AC, Dermitzakis ET. 2013. Expression quantitative trait loci: present and future. *Philos Trans R Soc Lond B Biol Sci* **368**: 20120362.

O'Connell J, Sharp K, Shrine N, Wain L, Hall I, Tobin M, Zagury JF, Delaneau O, Marchini J. 2016. Haplotype estimation for biobank-scale data sets. *Nat Genet* **48**: 817-820.
Ocana K, de Oliveira D. 2015. Parallel computing in genomic research: advances and applications. *Adv Appl Bioinform Chem* **8**: 23-35.
Ohshima K, Mogi M, Jing F, Iwanami J, Tsukuda K, Min LJ, Ogimoto A, Dahlof B, Steckelings UM, Unger T et al. 2012. Direct angiotensin II type 2 receptor stimulation ameliorates insulin resistance in type 2 diabetes mice with PPARgamma activation. *PLoS One* **7**: e48387.
Onengut-Gumuscu S, Chen WM, Burren O, Cooper NJ, Quinlan AR, Mychaleckyj JC, Farber E, Bonnie JK, Szpak M, Schofield E et al. 2015. Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. *Nat Genet* **47**: 381-386.
Pagotto U. 2009. Where does insulin resistance start? The brain. *Diabetes Care* **32 Suppl 2**: S174-177.
Paltoo DN, Rodriguez LL, Feolo M, Gillanders E, Ramos EM, Rutter JL, Sherry S, Wang VO, Bailey A, Baker R et al. 2014. Data use under the NIH GWAS data sharing policy and future directions. *Nat Genet* **46**: 934-938.
Paul DS, Soranzo N, Beck S. 2014. Functional interpretation of non-coding sequence variation: concepts and challenges. *Bioessays* **36**: 191-199.
Perry JR, Voight BF, Yengo L, Amin N, Dupuis J, Ganser M, Grallert H, Navarro P, Li M, Qi L et al. 2012. Stratifying type 2 diabetes cases by BMI identifies genetic risk variants in LAMA1 and enrichment for risk variants in lean compared to obese cases. *PLoS Genet* **8**: e1002741.
Pers TH, Karjalainen JM, Chan Y, Westra HJ, Wood AR, Yang J, Lui JC, Vedantam S, Gustafsson S, Esko T et al. 2015. Biological interpretation of genome-wide association studies using predicted gene functions. *Nat Commun* **6**: 5890.
Petersen KF, Befroy D, Dufour S, Dziura J, Ariyan C, Rothman DL, DiPietro L, Cline GW, Shulman GI. 2003. Mitochondrial dysfunction in the elderly: possible role in insulin resistance. *Science* **300**: 1140-1142.
Pistis G, Porcu E, Vrieze SI, Sidore C, Steri M, Danjou F, Busonero F, Mulas A, Zoledziewska M, Maschio A et al. 2015. Rare variant genotype imputation with thousands of study-specific whole-genome sequences: implications for cost-effective study designs. *Eur J Hum Genet* **23**: 975-983.
Poitout V, Robertson RP. 2002. Minireview: Secondary beta-cell failure in type 2 diabetes--a convergence of glucotoxicity and lipotoxicity. *Endocrinology* **143**: 339-342.
Portales-Casamar E, Arenillas D, Lim J, Swanson MI, Jiang S, McCallum A, Kirov S, Wasserman WW. 2009. The PAZAR database of gene regulatory information coupled to the ORCA toolkit for the study of regulatory sequences. *Nucleic Acids Res* **37**: D54-60.
Poulsen P, Kyvik KO, Vaag A, Beck-Nielsen H. 1999. Heritability of type II (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance--a population-based twin study. *Diabetologia* **42**: 139-145.
Poulsen P, Levin K, Petersen I, Christensen K, Beck-Nielsen H, Vaag A. 2005. Heritability of insulin secretion, peripheral and hepatic insulin action, and intracellular glucose partitioning in young and old Danish twins. *Diabetes* **54**: 275-283.
Pound LD, Hang Y, Sarkar SA, Wang Y, Milam LA, Oeser JK, Printz RL, Lee CE, Stein R, Hutton JC et al. 2011. The Pancreatic Islet Beta Cell-Enriched Transcription Factor Pdx-1 Regulates Slc30a8 Gene Transcription Through an Intronic Enhancer. *The Biochemical journal* **433**: 95-105.
Price AL, Spencer CC, Donnelly P. 2015. Progress and promise in understanding the genetic basis of common diseases. *Proc Biol Sci* **282**: 20151684.
Pritchard JK. 2001. Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet* **69**: 124-137.
Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**: 559-575.
Randall JC Winkler TW Kutalik Z Berndt SI Jackson AU Monda KL Kilpelainen TO Esko T Magi R Li S et al. 2013. Sex-stratified genome-wide association studies including 270,000 individuals show sexual dimorphism in genetic loci for anthropometric traits. *PLoS Genet* **9**: e1003500.
Rees SD, Hydrie MZ, O'Hare JP, Kumar S, Shera AS, Basit A, Barnett AH, Kelly MA. 2011. Effects of 16 genetic variants on fasting glucose and type 2 diabetes in South Asians: ADCY5 and GLIS3 variants may predispose to type 2 diabetes. *PLoS One* **6**: e24710.
Reich DE, Lander ES. 2001. On the allelic spectrum of human disease. *Trends Genet* **17**: 502-510.
Relling MV, Evans WE. 2015. Pharmacogenomics in the clinic. *Nature* **526**: 343-350.
Replication DIG Meta-analysis C Asian Genetic Epidemiology Network Type 2 Diabetes C South Asian Type 2 Diabetes C Mexican American Type 2 Diabetes C Type 2 Diabetes Genetic Exploration by Nex-generation sequencing in muylti-Ethnic Samples C Mahajan A Go MJ Zhang W Below JE et al. 2014. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat Genet* **46**: 234-244.
Rice WR. 2014. The synthesis paradigm in genetics. *Genetics* **196**: 367-371.
Riordan JR, Rommens JM, Kerem B, Alon N, Rozmahel R, Grzelczak Z, Zielenski J, Lok S, Plavsic N, Chou JL et al. 1989a. Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science* **245**: 1066.

Riordan JR, Rommens JM, Kerem B, Alon N, Rozmahel R, Grzelczak Z, Zielenski J, Lok S, Plavsic N, Chou JL et al. 1989b. Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science* **245**: 1066-1073.
Roadmap Epigenomics C, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J et al. 2015. Integrative analysis of 111 reference human epigenomes. *Nature* **518**: 317-330.
Rommens JM, Iannuzzi MC, Kerem B, Drumm ML, Melmer G, Dean M, Rozmahel R, Cole JL, Kennedy D, Hidaka N et al. 1989. Identification of the cystic fibrosis gene: chromosome walking and jumping. *Science* **245**: 1059.
Rubio-Cabezas O, Jensen JN, Hodgson MI, Codner E, Ellard S, Serup P, Hattersley AT. 2011. Permanent Neonatal Diabetes and Enteric Anendocrinosis Associated With Biallelic Mutations in NEUROG3. *Diabetes* **60**: 1349-1353.
Sabeti PC Varilly P Fry B Lohmueller J Hostetter E Cotsapas C Xie X Byrne EH McCarroll SA Gaudet R et al. 2007. Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**: 913-918.
Salanti G, Southam L, Altshuler D, Ardlie K, Barroso I, Boehnke M, Cornelis MC, Frayling TM, Grallert H, Grarup N et al. 2009. Underlying genetic models of inheritance in established type 2 diabetes associations. *Am J Epidemiol* **170**: 537-545.
Sanchez F, Guindo-Martínez M, Bonàs-Guarch S, Puiggròs M, Ejarque J, Díaz C, Tejedor E, Badia R, Mercader JM, Torrents D. 2016. GUIDANCE: An Integrated Framework for Large-scale Genome and Phenome-Wide Association Studies on Parallel Computing Platforms. *Under review*.
Sandhu MS, Weedon MN, Fawcett KA, Wasson J, Debenham SL, Daly A, Lango H, Frayling TM, Neumann RJ, Sherva R et al. 2007. Common variants in WFS1 confer risk of type 2 diabetes. *Nat Genet* **39**: 951-953.
Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**: 5463-5467.
Santer R, Kinner M, Lassen CL, Schneppenheim R, Eggert P, Bald M, Brodehl J, Daschner M, Ehrich JH, Kemper M et al. 2003. Molecular analysis of the SGLT2 gene in patients with renal glucosuria. *J Am Soc Nephrol* **14**: 2873-2882.
Saxena R Elbers CC Guo Y Peter I Gaunt TR Mega JL Lanktree MB Tare A Castillo BA Li YR et al. 2012. Large-scale gene-centric meta-analysis across 39 studies identifies type 2 diabetes loci. *Am J Hum Genet* **90**: 410-425.
Saxena R Hivert MF Langenberg C Tanaka T Pankow JS Vollenweider P Lyssenko V Bouatia-Naji N Dupuis J Jackson AU et al. 2010. Genetic variation in GIPR influences the glucose and insulin responses to an oral glucose challenge. *Nat Genet* **42**: 142-148.
Saxena R, Saleheen D, Been LF, Garavito ML, Braun T, Bjonnes A, Young R, Ho WK, Rasheed A, Frossard P et al. 2013. Genome-wide association study identifies a novel locus contributing to type 2 diabetes susceptibility in Sikhs of Punjabi origin from India. *Diabetes* **62**: 1746-1755.
Schadt EE, Linderman MD, Sorenson J, Lee L, Nolan GP. 2010. Computational solutions to large-scale data management and analysis. *Nat Rev Genet* **11**: 647-657.
Scheet P, Stephens M. 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet* **78**: 629-644.
Schizophrenia Working Group of the Psychiatric Genomics C. 2014. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**: 421-427.
Schonherr S, Forer L, Weissensteiner H, Kronenberg F, Specht G, Kloss-Brandstatter A. 2012. Cloudgene: a graphical execution platform for MapReduce programs on private and public clouds. *BMC Bioinformatics* **13**: 200.
Schooneman MG, Vaz FM, Houten SM, Soeters MR. 2013. Acylcarnitines: reflecting or inflicting insulin resistance? *Diabetes* **62**: 1-8.
Schork NJ. 1997. Genetics of complex disease: approaches, problems, and solutions. *Am J Respir Crit Care Med* **156**: S103-109.
Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, Erdos MR, Stringham HM, Chines PS, Jackson AU et al. 2007. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* **316**: 1341-1345.
Segurel L, Austerlitz F, Toupance B, Gautier M, Kelley JL, Pasquet P, Lonjou C, Georges M, Voisin S, Cruaud C et al. 2013. Positive selection of protective variants for type 2 diabetes from the Neolithic onward: a case study in Central Asia. *Eur J Hum Genet* **21**: 1146-1151.
Sham PC, Cherny SS. 2011. Chapter 1 - Genetic Architecture of Complex Diseases. In *Analysis of Complex Disease Association Studies*, doi:http://dx.doi.org/10.1016/B978-0-12-375142-3.10001-X (ed. E Zeggini, Morris, Andrew), pp. 1-13. Academic Press, San Diego.
Shum M, Pinard S, Guimond MO, Labbe SM, Roberge C, Baillargeon JP, Langlois MF, Alterman M, Wallinder C, Hallberg A et al. 2013. Angiotensin II type 2 receptor promotes adipocyte differentiation and restores adipocyte size in high-fat/high-fructose diet-induced insulin resistance in rats. *Am J Physiol Endocrinol Metab* **304**: E197-210.
Sidransky E. 2006. Heterozygosity for a Mendelian disorder as a risk factor for complex disease. *Clin Genet* **70**: 275-282.

Sigma Type 2 Diabetes Consortium, Estrada K, Aukrust I, Bjorkhaug L, Burtt NP, Mercader JM, Garcia-Ortiz H, Huerta-Chagoya A, Moreno-Macias H, Walford G et al. 2014a. Association of a low-frequency variant in HNF1A with type 2 diabetes in a Latino population. *JAMA* **311**: 2305-2314.
Sigma Type 2 Diabetes Consortium, Williams AL, Jacobs SB, Moreno-Macias H, Huerta-Chagoya A, Churchhouse C, Marquez-Luna C, Garcia-Ortiz H, Gomez-Vazquez MJ, Burtt NP et al. 2014c. Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico. *Nature* **506**: 97-101.
Simonis-Bik AM, Nijpels G, van Haeften TW, Houwing-Duistermaat JJ, Boomsma DI, Reiling E, van Hove EC, Diamant M, Kramer MHH, Heine RJ et al. 2010. Gene Variants in the Novel Type 2 Diabetes Loci CDC123/CAMK1D, THADA, ADAMTS9, BCL11A, and MTNR1B Affect Different Aspects of Pancreatic β-Cell Function. *Diabetes* **59**: 293-301.
Sinnott JA, Kraft P. 2012. Artifact due to differential error when cases and controls are imputed from different platforms. *Hum Genet* **131**: 111-119.
Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, Boutin P, Vincent D, Belisle A, Hadjadj S et al. 2007. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**: 881-885.
Slatkin M. 2008. Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nat Rev Genet* **9**: 477-485.
Spjuth O, Bongcam-Rudloff E, Hernandez GC, Forer L, Giovacchini M, Guimera RV, Kallio A, Korpelainen E, Kandula MM, Krachunov M et al. 2015. Experiences with workflows for automating data-intensive bioinformatics. *Biol Direct* **10**: 43.
Srinivasan S, Florez JC. 2015. Therapeutic Challenges in Diabetes Prevention: We Have Not Found the "Exercise Pill". *Clin Pharmacol Ther* **98**: 162-169.
Steinthorsdottir V, Thorleifsson G, Reynisdottir I, Benediktsson R, Jonsdottir T, Walters GB, Styrkarsdottir U, Gretarsdottir S, Emilsson V, Ghosh S et al. 2007. A variant in CDKAL1 influences insulin response and risk of type 2 diabetes. *Nat Genet* **39**: 770-775.
Steinthorsdottir V, Thorleifsson G, Sulem P, Helgason H, Grarup N, Sigurdsson A, Helgadottir HT, Johannsdottir H, Magnusson OT, Gudjonsson SA et al. 2014. Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. *Nat Genet* **46**: 294-298.
Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, Iyer R, Schatz MC, Sinha S, Robinson GE. 2015. Big Data: Astronomical or Genomical? *PLoS Biol* **13**: e1002195.
Strachan DP, Rudnicka AR, Power C, Shepherd P, Fuller E, Davis A, Gibb I, Kumari M, Rumley A, Macfarlane GJ et al. 2007. Lifecourse influences on health among British adults: effects of region of residence in childhood and adulthood. *Int J Epidemiol* **36**: 522-531.
Stranger BE, Stahl EA, Raj T. 2011. Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics* **187**: 367-383.
Stringer CB, Andrews P. 1988. Genetic and fossil evidence for the origin of modern humans. *Science* **239**: 1263-1268.
Studies N-NWGoRiA, Chanock SJ, Manolio T, Boehnke M, Boerwinkle E, Hunter DJ, Thomas G, Hirschhorn JN, Abecasis G, Altshuler D et al. 2007. Replicating genotype-phenotype associations. *Nature* **447**: 655-660.
Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M et al. 2015. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* **12**: e1001779.
Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Hsi-Yang Fritz M et al. 2015. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**: 75-81.
Sulem P, Helgason H, Oddson A, Stefansson H, Gudjonsson SA, Zink F, Hjartarson E, Sigurdsson GT, Jonasdottir A, Jonasdottir A et al. 2015. Identification of a large set of rare complete human knockouts. *Nat Genet* **47**: 448-452.
Surakka I, Sarin A-P, Ruotsalainen SE, Durbin R, Salomaa V, Daly M, Palotie A, Ripatti S. 2016. The rate of false polymorphisms introduced when imputing genotypes from global imputation panels. *bioRxiv* doi:10.1101/080770.
Swanton C, Govindan R. 2016. Clinical Implications of Genomic Discoveries in Lung Cancer. *N Engl J Med* **374**: 1864-1873.
Szumilas M. 2010. Explaining odds ratios. *J Can Acad Child Adolesc Psychiatry* **19**: 227-229.
Tabassum R, Chauhan G, Dwivedi OP, Mahajan A, Jaiswal A, Kaur I, Bandesh K, Singh T, Mathai BJ, Pandey Y et al. 2013. Genome-wide association study for type 2 diabetes in Indians identifies a new susceptibility locus at 2q21. *Diabetes* **62**: 977-986.
Tabor HK, Risch NJ, Myers RM. 2002. Candidate-gene approaches for studying complex genetic traits: practical considerations. *Nat Rev Genet* **3**: 391-397.
Tallapragada DS, Bhaskar S, Chandak GR. 2015. New insights from monogenic diabetes for "common" type 2 diabetes. *Front Genet* **6**: 251.
Tejedor E, Farreras M, Grove D, Badia RM, Almasi G, Labarta J. 2012. A high-productivity task-based programming model for clusters. *Concurrency and Computation: Practice and Experience* **24**: 2421-2448.
Tenesa A, Haley CS. 2013. The heritability of human disease: estimation, uses and abuses. *Nat Rev Genet* **14**: 139-149.
The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467**: 1061-1073.

The 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467**: 1061-1073.
The 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**: 56-65.
The 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA et al. 2015. A global reference for human genetic variation. *Nature* **526**: 68-74.
The American Diabetes Association. 2015. Genetics portal for type 2 diabetes debuts. Vol 2016. Diabetes Dispatch.
The CARDIoGRAMplusC4D Consortium. 2015. A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat Genet* doi:10.1038/ng.3396.
Thompson SG, Sharp SJ. 1999. Explaining heterogeneity in meta-analysis: a comparison of methods. *Stat Med* **18**: 2693-2708.
Thuesen BH, Cerqueira C, Aadahl M, Ebstrup JF, Toft U, Thyssen JP, Fenger RV, Hersoug LG, Elberling J, Pedersen O et al. 2014. Cohort Profile: the Health2006 cohort, research centre for prevention and health. *Int J Epidemiol* **43**: 568-575.
Tienari PJ, Tuomilehto-Wolf E, Tuomilehto J, Peltonen L. 1992. HLA haplotypes in type 1 (insulin-dependent) diabetes mellitus: molecular analysis of the HLA-DQ locus. The DIME Study Group. *Diabetologia* **35**: 254-260.
Todd JN, Poon W, Lyssenko V, Groop L, Nichols B, Wilmot M, Robson S, Enjyoji K, Herman MA, Hu C et al. 2015. Variation in glucose homeostasis traits associated with P2RX7 polymorphisms in mice and humans. *J Clin Endocrinol Metab* **100**: E688-696.
Torres JM, Cox NJ, Philipson LH. 2013. Genome wide association studies for diabetes: perspective on results and challenges. *Pediatr Diabetes* **14**: 90-96.
Tryka KA, Hao L, Sturcke A, Jin Y, Wang ZY, Ziyabari L, Lee M, Popova N, Sharopova N, Kimura M et al. 2014. NCBI's Database of Genotypes and Phenotypes: dbGaP. *Nucleic Acids Res* **42**: D975-979.
Tukiainen T, Pirinen M, Sarin AP, Ladenvall C, Kettunen J, Lehtimaki T, Lokki ML, Perola M, Sinisalo J, Vlachopoulou E et al. 2014. Chromosome X-wide association study identifies Loci for fasting insulin and height and evidence for incomplete dosage compensation. *PLoS Genet* **10**: e1004127.
U.S. Department of Energy & Human Genome Project program. Human Genome Project Budget. In *"About the Human Genome Project"*, Vol 2016. Human Genome Information Archive.
UK10K Consortium, Walter K, Min JL, Huang J, Crooks L, Memari Y, McCarthy S, Perry JR, Xu C, Futema M et al. 2015. The UK10K project identifies rare variants in health and disease. *Nature* **526**: 82-90.
UK Prospective Diabetes Study (UKPDS) Group. 1998. Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33). . *Lancet* **352**: 837-853.
Underwood PC, Adler GK. 2013. The renin angiotensin aldosterone system and insulin resistance in humans. *Curr Hypertens Rep* **15**: 59-70.
Unoki H, Takahashi A, Kawaguchi T, Hara K, Horikoshi M, Andersen G, Ng DP, Holmkvist J, Borch-Johnsen K, Jorgensen T et al. 2008. SNPs in KCNQ1 are associated with susceptibility to type 2 diabetes in East Asian and European populations. *Nat Genet* **40**: 1098-1102.
Vaxillaire M, Froguel P. 2006. Genetic basis of maturity-onset diabetes of the young. *Endocrinol Metab Clin North Am* **35**: 371-384, x.
Vaxillaire M, Yengo L, Lobbens S, Rocheleau G, Eury E, Lantieri O, Marre M, Balkau B, Bonnefond A, Froguel P. 2014. Type 2 diabetes-related genetic risk scores associated with variations in fasting plasma glucose and development of impaired glucose homeostasis in the prospective DESIR study. *Diabetologia* **57**: 1601-1610.
Vinkhuyzen AA, Wray NR, Yang J, Goddard ME, Visscher PM. 2013. Estimation and partition of heritability in human populations using whole-genome analysis methods. *Annu Rev Genet* **47**: 75-95.
Visscher PM, Brown MA, McCarthy MI, Yang J. 2012. Five years of GWAS discovery. *Am J Hum Genet* **90**: 7-24.
Visscher PM, Hill WG, Wray NR. 2008. Heritability in the genomics era--concepts and misconceptions. *Nat Rev Genet* **9**: 255-266.
Voight BF, Kang HM, Ding J, Palmer CD, Sidore C, Chines PS, Burtt NP, Fuchsberger C, Li Y, Erdmann J et al. 2012. The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet* **8**: e1002793.
Voight BF Scott LJ Steinthorsdottir V Morris AP Dina C Welch RP Zeggini E Huth C Aulchenko YS Thorleifsson G et al. 2010. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat Genet* **42**: 579-589.
Vukcevic D, Hechter E, Spencer C, Donnelly P. 2011. Disease model distortion in association studies. *Genet Epidemiol* **35**: 278-290.
Wain LV, Shrine N, Miller S, Jackson VE, Ntalla I, Artigas MS, Billington CK, Kheirallah AK, Allen R, Cook JP et al. 2015. Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank. *The Lancet Respiratory Medicine* **3**: 769-781.
Wakeley J. 2005. The limits of theoretical population genetics. *Genetics* **169**: 1-7.

Wall JD, Pritchard JK. 2003. Haplotype blocks and linkage disequilibrium in the human genome. *Nat Rev Genet* **4**: 587-597.
Wang HM, Hsiao CL, Hsieh AR, Lin YC, Fann CS. 2012a. Constructing endophenotypes of complex diseases using non-negative matrix factorization and adjusted rand index. *PLoS One* **7**: e40996.
Wang J, Cortina G, Wu SV, Tran R, Cho JH, Tsai MJ, Bailey TJ, Jamrich M, Ament ME, Treem WR et al. 2006. Mutant neurogenin-3 in congenital malabsorptive diarrhea. *N Engl J Med* **355**: 270-280.
Wang Q, Lu Q, Zhao H. 2015. A review of study designs and statistical methods for genomic epidemiology studies using next generation sequencing. *Front Genet* **6**: 149.
Wang X, Wei X, Thijssen B, Das J, Lipkin SM, Yu H. 2012b. Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nat Biotechnol* **30**: 159-164.
Ward LD, Kellis M. 2012a. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res* **40**: D930-934.
Ward LD, Kellis M. 2012d. Interpreting noncoding genetic variation in complex traits and human disease. *Nat Biotechnol* **30**: 1095-1106.
Ward LD, Kellis M. 2016. HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. *Nucleic Acids Res* **44**: D877-881.
Watson JD, Crick FH. 1953. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**: 737-738.
Weinberg W. 1908. On the demonstration of heredity in man. In *Papers on human genetics*, (ed. S Boyer). Prentice Hall, Englewood Cliffs.
Wellcome Trust Case Control C. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**: 661-678.
Wellcome Trust Case Control C, Maller JB, McVean G, Byrnes J, Vukcevic D, Palin K, Su Z, Howson JM, Auton A, Myers S et al. 2012. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat Genet* **44**: 1294-1301.
Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L et al. 2014. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* **42**: D1001-1006.
Wessel J Chu AY Willems SM Wang S Yaghootkar H Brody JA Dauriz M Hivert MF Raghavan S Lipovich L et al. 2015. Low-frequency and rare exome chip variants associate with fasting glucose and type 2 diabetes susceptibility. *Nat Commun* **6**: 5897.
Westra HJ, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, Christiansen MW, Fairfax BP, Schramm K, Powell JE et al. 2013. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat Genet* **45**: 1238-1243.
WHO. 2014. Global status report on noncommunicable diseases 2014. p. 298. WHO.
Wigginton JE, Cutler DJ, Abecasis GR. 2005. A note on exact tests of Hardy-Weinberg equilibrium. *Am J Hum Genet* **76**: 887-893.
Wijesekara N, Chimienti F, Wheeler MB. 2009. Zinc, a regulator of islet function and glucose homeostasis. *Diabetes, Obesity and Metabolism* **11**: 202-214.
Willemsen G, Ward KJ, Bell CG, Christensen K, Bowden J, Dalgard C, Harris JR, Kaprio J, Lyle R, Magnusson PK et al. 2015. The Concordance and Heritability of Type 2 Diabetes in 34,166 Twin Pairs From International Twin Registers: The Discordant Twin (DISCOTWIN) Consortium. *Twin Res Hum Genet* **18**: 762-771.
Willer CJ, Li Y, Abecasis GR. 2010. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**: 2190-2191.
Williams AL, Patterson N, Glessner J, Hakonarson H, Reich D. 2012. Phasing of many thousands of genotyped samples. *Am J Hum Genet* **91**: 238-251.
Wilson AC, Cann RL. 1992. The recent African genesis of humans. *Sci Am* **266**: 68-73.
Wilson BJ, Nicholls SG. 2015. The Human Genome Project, and recent advances in personalized genomics. *Risk Manag Healthc Policy* **8**: 9-20.
Wiltshire S, Hattersley AT, Hitman GA, Walker M, Levy JC, Sampson M, O'Rahilly S, Frayling TM, Bell JI, Lathrop GM et al. 2001. A genomewide scan for loci predisposing to type 2 diabetes in a U.K. population (the Diabetes UK Warren 2 Repository): analysis of 573 pedigrees provides independent replication of a susceptibility locus on chromosome 1q. *Am J Hum Genet* **69**: 553-569.
Wise AL, Gyi L, Manolio TA. 2013. eXclusion: toward integrating the X chromosome in genome-wide association analyses. *Am J Hum Genet* **92**: 643-647.
Wray N, Visscher P. 2008. Estimating trait heritability. *Nature Education* **1**: 29.
Wu JJ, Roth RJ, Anderson EJ, Hong EG, Lee MK, Choi CS, Neufer PD, Shulman GI, Kim JK, Bennett AM. 2006. Mice lacking MAP kinase phosphatase-1 have enhanced MAP kinase activity and resistance to diet-induced obesity. *Cell Metab* **4**: 61-73.
Xiong Y, Collins QF, An J, Lupo E, Jr., Liu HY, Liu D, Robidoux J, Liu Z, Cao W. 2007. p38 mitogen-activated protein kinase plays an inhibitory role in hepatic lipogenesis. *J Biol Chem* **282**: 4975-4982.
Yamamoto F, Clausen H, White T, Marken J, Hakomori S. 1990. Molecular genetic basis of the histo-blood group ABO system. *Nature* **345**: 229-233.

Yang J, Ferreira T, Morris AP, Medland SE, Genetic Investigation of ATC, Replication DIG, Meta-analysis C, Madden PA, Heath AC, Martin NG et al. 2012. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet* **44**: 369-375.
Yang J, Weedon MN, Purcell S, Lettre G, Estrada K, Willer CJ, Smith AV, Ingelsson E, O'Connell JR, Mangino M et al. 2011. Genomic inflation factors under polygenic inheritance. *Eur J Hum Genet* **19**: 807-812.
Yang L, Chiu SC, Liao WK, Thomas MA. 2014. High Performance Data Clustering: A Comparative Analysis of Performance for GPU, RASC, MPI, and OpenMP Implementations. *J Supercomput* **70**: 284-300.
Yasuda K, Miyake K, Horikawa Y, Hara K, Osawa H, Furuta H, Hirota Y, Mori H, Jonsson A, Sato Y et al. 2008. Variants in KCNQ1 are associated with susceptibility to type 2 diabetes mellitus. *Nat Genet* **40**: 1092-1097.
Yvan-Charvet L, Even P, Bloch-Faure M, Guerre-Millo M, Moustaid-Moussa N, Ferre P, Quignard-Boulange A. 2005. Deletion of the angiotensin type 2 receptor (AT2R) reduces adipose cell size and protects from diet-induced obesity and insulin resistance. *Diabetes* **54**: 991-999.
Zaitlen N, Pasaniuc B, Sankararaman S, Bhatia G, Zhang J, Gusev A, Young T, Tandon A, Pollack S, Vilhjalmsson BJ et al. 2014. Leveraging population admixture to characterize the heritability of complex traits. *Nat Genet* **46**: 1356-1362.
Zavattari P, Lampis R, Mulargia A, Loddo M, Angius E, Todd JA, Cucca F. 2000. Confirmation of the DRB1-DQB1 loci as the major component of IDDM1 in the isolated founder population of Sardinia. *Hum Mol Genet* **9**: 2967-2972.
Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, Hu T, de Bakker PI, Abecasis GR, Almgren P, Andersen G et al. 2008. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet* **40**: 638-645.
Zeggini E, Weedon MN, Lindgren CM, Frayling TM, Elliott KS, Lango H, Timpson NJ, Perry JR, Rayner NW, Freathy RM et al. 2007. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* **316**: 1336-1341.
Zeng P, Zhao Y, Qian C, Zhang L, Zhang R, Gou J, Liu J, Liu L, Chen F. 2015. Statistical analysis for genome-wide association study. *J Biomed Res* **29**: 285-297.
Zhao M, Wang Q, Wang Q, Jia P, Zhao Z. 2013. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics* **14 Suppl 11**: S1.
Zheng HF Forgetta V Hsu YH Estrada K Rosello-Diez A Leo PJ Dahia CL Park-Min KH Tobias JH Kooperberg C et al. 2015. Whole-genome sequencing identifies EN1 as a determinant of bone density and fracture. *Nature* doi:10.1038/nature14878.
Zhou S, Zheng X, Wang J, Delisle P. 1993. Utopia: A Load Sharing Facility for Large, Heterogeneous Distributed Computer Systems. *Softw Pract Exper* **23**: 1305-1336.
Zhou X, Li D, Zhang B, Lowdon RF, Rockweiler NB, Sears RL, Madden PA, Smirnov I, Costello JF, Wang T. 2015. Epigenomic annotation of genetic variants using the Roadmap Epigenome Browser. *Nat Biotechnol* **33**: 345-346.
Zhou X, Lowdon RF, Li D, Lawson HA, Madden PA, Costello JF, Wang T. 2013. Exploring long-range genome interactions using the WashU Epigenome Browser. *Nat Methods* **10**: 375-376.
Zhou X, Maricque B, Xie M, Li D, Sundaram V, Martin EA, Koebbe BC, Nielsen C, Hirst M, Farnham P et al. 2011. The Human Epigenome Browser at Washington University. *Nat Methods* **8**: 989-990.
Zinman B, Wanner C, Lachin JM, Fitchett D, Bluhmki E, Hantel S, Mattheus M, Devins T, Johansen OE, Woerle HJ et al. 2015. Empagliflozin, Cardiovascular Outcomes, and Mortality in Type 2 Diabetes. *N Engl J Med* **373**: 2117-2128.
Zoungas S, de Galan BE, Ninomiya T, Grobbee D, Hamet P, Heller S, MacMahon S, Marre M, Neal B, Patel A et al. 2009. Combined effects of routine blood pressure lowering and intensive glucose control on macrovascular and microvascular outcomes in patients with type 2 diabetes: New results from the ADVANCE trial. *Diabetes Care* **32**: 2068-2074.

Appendix

Appendix 1. Bonnelykke K, Sleiman P, Nielsen K, Kreiner-Moller E, Mercader JM, Belgrave D, den, Dekker HT, Husby A, Sevelsted A, Faura-Tellez G … Bonàs-Guarch S … Bisgaard H 2014. A genome-wide association study identifies CDHR3 as a susceptibility locus for early childhood asthma with severe exacerbations. *Nat Genet* **46**: 51-55.

Contribution of the PhD candidate

- Genotype imputation.
- Bioinformatics and statistical analyses.
- Interpretation and costumed data analyses.

A genome-wide association study identifies *CDHR3* as a susceptibility locus for early childhood asthma with severe exacerbations

Klaus Bønnelykke[1,24,25], Patrick Sleiman[2,24], Kasper Nielsen[3,24], Eskil Kreiner-Møller[1], Josep M Mercader[4], Danielle Belgrave[5,6], Herman T den Dekker[7–9], Anders Husby[1,10], Astrid Sevelsted[1], Grissel Faura-Tellez[11,12], Li Juel Mortensen[1], Lavinia Paternoster[13], Richard Flaaten[1], Anne Mølgaard[1], David E Smart[10], Philip F Thomsen[14], Morten A Rasmussen[15], Silvia Bonàs-Guarch[4], Claus Holst[16], Ellen A Nohr[17,18], Rachita Yadav[3], Michael E March[2], Thomas Blicher[19], Peter M Lackie[11], Vincent W V Jaddoe[7,9,20], Angela Simpson[5], John W Holloway[11], Liesbeth Duijts[8,9,21], Adnan Custovic[5], Donna E Davies[10], David Torrents[4,22], Ramneek Gupta[3], Mads V Hollegaard[23], David M Hougaard[23], Hakon Hakonarson[2,25] & Hans Bisgaard[1,25]

Asthma exacerbations are among the most frequent causes of hospitalization during childhood, but the underlying mechanisms are poorly understood. We performed a genome-wide association study of a specific asthma phenotype characterized by recurrent, severe exacerbations occurring between 2 and 6 years of age in a total of 1,173 cases and 2,522 controls. Cases were identified from national health registries of hospitalization, and DNA was obtained from the Danish Neonatal Screening Biobank. We identified five loci with genome-wide significant association. Four of these, *GSDMB*, *IL33*, *RAD50* and *IL1RL1*, were previously reported as asthma susceptibility loci, but the effect sizes for these loci in our cohort were considerably larger than in the previous genome-wide association studies of asthma. We also obtained strong evidence for a new susceptibility gene, *CDHR3* (encoding cadherin-related family member 3), which is highly expressed in airway epithelium. These results demonstrate the strength of applying specific phenotyping in the search for asthma susceptibility genes.

Acute asthma exacerbations are among the most frequent causes of hospitalization during childhood and are responsible for large healthcare expenditures[1–4]. Available treatment options for prevention and treatment of asthma exacerbations are inadequate[5], suggesting that asthma with severe exacerbations may represent a distinct subtype of disease and demonstrating a need for improved understanding of its pathogenesis.

Asthma heritability is estimated to be 70–90% (refs. 6,7), but only a limited number of susceptibility loci have been verified in genome-wide association studies (GWAS)[8–13]. Larger GWAS may identify new susceptibility loci with smaller effects, but, owing to the large heterogeneity in asthma[14], an alternative strategy is to increase phenotype specificity in genome-wide analyses. A specific phenotype is likely to be more closely related to a specific pathogenetic mechanism, and focusing on a particular phenotype may increase the power of genetic studies.

We aimed to increase understanding of the genetic background of early childhood asthma with severe exacerbations by conducting a

[1]Copenhagen Prospective Studies on Asthma in Childhood, Health Sciences, University of Copenhagen & Danish Pediatric Asthma Center, Copenhagen University Hospital, Gentofte, Denmark. [2]Center for Applied Genomics, Children's Hospital of Philadelphia (CHOP), Philadelphia, Pennsylvania, USA. [3]Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Lyngby, Denmark. [4]Joint Institute for Research in Biomedicine and Barcelona Supercomputing Center (IRB-BSC) Program on Computational Biology, Barcelona Supercomputing Center, Barcelona, Spain. [5]Centre for Respiratory Medicine and Allergy, Institute of Inflammation and Repair, University of Manchester and University Hospital of South Manchester, Manchester, UK. [6]Centre for Health Informatics, Institute of Population Health, University of Manchester, Manchester, UK. [7]Generation R Study Group, Erasmus Medical Center, Rotterdam, The Netherlands. [8]Department of Pediatrics, Division of Respiratory Medicine, Erasmus Medical Center, Rotterdam, The Netherlands. [9]Department of Epidemiology, Erasmus Medical Center, Rotterdam, The Netherlands. [10]Brooke Laboratory, Clinical and Experimental Sciences, Faculty of Medicine, University of Southampton, University Hospital Southampton, Southampton, UK. [11]Faculty of Medicine, University of Southampton, Southampton General Hospital, Southampton, UK. [12]Pediatric Pulmonology and Pediatric Allergology, University of Groningen, University Medical Center Groningen, Beatrix Children's Hospital, Groningen Research Institute for Asthma and COPD, Groningen, The Netherlands. [13]Integrative Epidemiology Unit, School of Social & Community Medicine, University of Bristol, Bristol, UK. [14]Center for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark. [15]Department of Food Science, University of Copenhagen, Copenhagen, Denmark. [16]Institute of Preventive Medicine, Copenhagen University Hospital, Copenhagen, Denmark. [17]Institute of Clinical Research, University of Southern Denmark, Aarhus, Denmark. [18]Department of Public Health, Section for Epidemiology, Aarhus University, Aarhus, Denmark. [19]Novo Nordisk Foundation Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark. [20]Department of Pediatrics, Erasmus Medical Center, Rotterdam, The Netherlands. [21]Department of Pediatrics, Division of Neonatology, Erasmus Medical Center, Rotterdam, The Netherlands. [22]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain. [23]Danish Centre for Neonatal Screening, Department of Clinical Biochemistry and Immunology, Statens Serum Institut (SSI), Copenhagen, Denmark. [24]These authors contributed equally to this work. [25]These authors jointly directed this work. Correspondence should be addressed to K.B. (kb@copsac.com).

Received 27 May; accepted 28 October; published online 17 November 2013; doi:10.1038/ng.2830

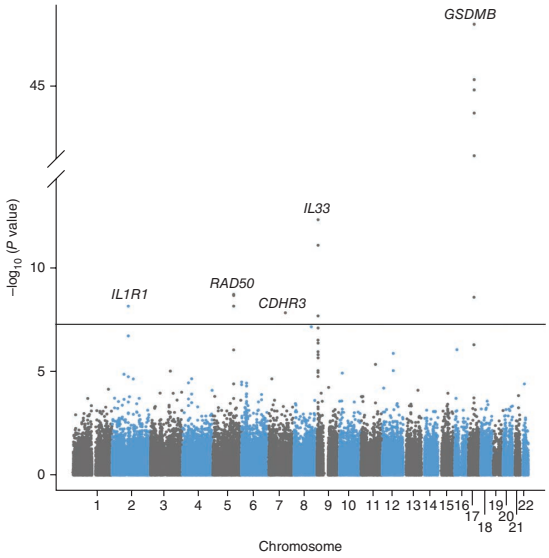


Figure 1 Manhattan plot for the discovery genome-wide association analysis. The horizontal line indicates the genome-wide significance threshold ($P < 5 \times 10^{-8}$).

GWAS of this particular asthma phenotype. We identified children with recurrent acute hospitalizations for asthma occurring between 2 and 6 years of age (cases) from the Danish National Patient Register. We then extracted and amplified DNA from dried blood spot samples isolated from the Danish Neonatal Screening Biobank, as previously described[15,16], before genome-wide array genotyping (Affymetrix Axiom CEU array).

Case criteria were fulfilled for 2,029 of 1.7 million children born in Denmark between 1982 and 1995 (1.1/1,000 children). The final case cohort (Copenhagen Prospective Studies on Asthma in Childhood exacerbation cohort, $COPSAC_{exacerbation}$) after genotyping and quality control comprised 1,173 children (**Supplementary Fig. 1**). Compared to the general population, cases were more often boys (67 versus 51%) and more often had mothers who smoked during pregnancy (32 versus 15%) (**Supplementary Tables 1** and **2**). Controls consisted of 2,511 individuals of Danish descent without asthma who were previously genotyped (Illumina Human610-Quad v1.0 BeadChip). We analyzed association between disease and 124,514 SNPs genotyped in both cases and controls, and we accounted for population stratification by multidimensional scaling. The genomic inflation factor was 1.04. The genome-wide association analysis detected an excess of association signals beyond those expected by chance (**Supplementary Fig. 2**), and SNPs from five regions reached genome-wide significance ($P < 5 \times 10^{-8}$; **Fig. 1** and **Supplementary Fig. 3**). The top SNPs from the five loci were rs2305480 in *GSDMB* (odds ratio (OR) = 2.28, $P = 1.3 \times 10^{-48}$), rs928413 near *IL33* (OR = 1.50, $P = 4.2 \times 10^{-13}$), rs6871536 in *RAD50* (OR = 1.44, $P = 1.7 \times 10^{-9}$), rs1558641 in *IL1RL1* (OR = 1.56, $P = 6.6 \times 10^{-9}$) and rs6967330 in *CDHR3* (OR = 1.45, $P = 1.4 \times 10^{-8}$) (**Table 1**). Validation of results for the top SNPs by regenotyping of cases and use of an alternative control population gave similar results (**Supplementary Tables 3** and **4**).

Association analyses in the discovery cohort stratified on number of asthma-related hospitalizations showed higher OR with increasing number of hospitalizations for all five SNPs (**Table 2**). There was no significant interaction between the top SNPs and no effect modification by sex.

We first sought replication in the childhood-onset stratum (with onset before 16 years of age) from a previous GWAS of asthma including 14,503 individuals conducted by the GABRIEL Consortium[11] (**Supplementary Table 5**), which showed evidence of association for all 5 of the genome-wide significant loci reported here (**Table 1**). The *CDHR3* locus was the only locus that had not previously been associated with asthma or any other atopic trait. We therefore followed up the top SNP from this locus (rs6967330) by further replication in a total of 3,975 children from 2 birth cohorts of European ancestry ($COPSAC_{2000}$ and the Manchester Asthma and Allergy Study (MAAS)) and in 1 cohort with a population of mixed ancestry (Generation R). There was evidence for association with asthma before the age of 6 years in combined analyses of the three birth cohorts and in the combined replication sets (**Table 1**, **Supplementary Fig. 4** and **Supplementary Table 6**), as well as in a subsample including the 980 individuals with non-European ancestry (**Supplementary Table 6**).

Phenotype-specific replication was possible in the $COPSAC_{2000}$ and MAAS birth cohorts with prospective registration of acute asthma hospitalizations and exacerbations from birth to 6 years of age in a

Table 1 Discovery and replication results for the five genome-wide significant loci in the discovery analyses

Chr.	SNP effect allele	Nearest gene	Distance to gene (bp)	Effect allele frequency	Stage	OR (95% CI)	*P* value (fixed-effects model)[a]	*P* value (random effects model)	*P* heterogeneity
17	rs2305480[G]	*GSDMB*	0	0.60	Discovery	2.28 (2.04–2.55)	1.3×10^{-48}	–	–
					Replication 1	1.32 (1.23–1.39)	**6.4×10^{-23}**	**6.4×10^{-23}**	0.86
9	rs928413[G]	*IL33*	2,418	0.28	Discovery	1.50 (1.34–1.67)	4.2×10^{-13}	–	–
					Replication 1	1.24 (1.17–1.32)	**8.8×10^{-13}**	**2.5×10^{-6}**	0.007
5	rs6871536[C]	*RAD50*	0	0.22	Discovery	1.44 (1.28–1.62)	1.8×10^{-9}	–	–
					Replication 1	1.17 (1.10–1.25)	**7.6×10^{-7}**	**7.6×10^{-7}**	0.54
2	rs1558641[G]	*IL1R1*	0	0.85	Discovery	1.56 (1.34–1.81)	6.6×10^{-9}	–	–
					Replication 1	1.11 (1.04–1.19)	**0.003**	**0.003**	0.75
7	rs6967330[A]	*CDHR3*	0	0.19	Discovery	1.45 (1.28–1.66)	1.4×10^{-8}	–	–
					Replication 1	1.18 (1.10–1.27)	**3.0×10^{-6}**	**1.3×10^{-4}**	0.04
					Replication 2	1.40 (1.16–1.67)	**3.2×10^{-4}**	**3.2×10^{-4}**	0.87
					Replications 1 + 2	1.21 (1.13–1.29)	**1.6×10^{-8}**	**2.6×10^{-6}**	0.05
					Discovery + replications 1 + 2	1.26 (1.18–1.33)	**2.7×10^{-14}**	**2.7×10^{-7}**	0.02

Replication *P* values are shown in bold if significant after Bonferroni correction for the five loci tested ($P < 0.01$). Replication 1 results are from a previously published large-scale GWAS of asthma (asthma onset before 16 years; subanalysis of ref. 11). Replication 2 results are from the $COPSAC_{2000}$, MAAS and Generation R cohorts (asthma onset before 6 years). Chr., chromosome.
[a]A fixed-effects model was not applied in the discovery analysis.

Table 2 Association results for the five genome-wide significant and replicated top SNPs stratified on number of hospitalizations for asthma or acute bronchitis from 0–6 years of age in the discovery cohort

		Number of asthma-related hospitalizations				
		2 *n* = 272	3 *n* = 228	4–5 *n* = 277	6 or more *n* = 358	Association between number of hospitalizations and genotype
SNP effect allele	Nearest gene	OR (95% CI) *P* value	OR (95% CI) *P* value	OR (95% CI) *P* value	OR (95% CI) *P* value	*P* value[a]
rs2305480[G]	*GSDMB*	1.87 (1.54–2.26) 1.5×10^{-10}	2.24 (1.81–2.78) 2.1×10^{-13}	2.24 (1.83–2.73) 1.7×10^{-15}	2.72 (2.26–3.28) 3.5×10^{-27}	0.002
rs928413[G]	*IL33*	1.32 (1.09–1.61) 0.005	1.22 (0.98–1.50) 0.07	1.47 (1.21–1.79) 8.5×10^{-5}	1.91 (1.61–2.26) 6.2×10^{-14}	2.4×10^{-4}
rs6871536[C]	*RAD50*	1.31 (1.06–1.61) 0.01	1.26 (1.00–1.59) 0.05	1.45 (1.18–1.78) 3.6×10^{-4}	1.58 (1.31–1.89) 1.3×10^{-6}	0.09
rs1558641[G]	*IL1R1*	1.53 (1.16–2.02) 0.002	1.20 (0.91–1.57) 0.20	1.32 (1.02–1.71) 0.04	2.19 (1.66–2.90) 3.2×10^{-8}	0.02
rs6967330[A]	*CDHR3*	1.23 (0.98–1,56) 0.07	1.37 (1.07–1.75) 0.01	1.42 (1.13–1.78) 0.003	1.63 (1.33–1.97) 1.6×10^{-6}	0.04

Only the 1,135 children with full follow-up were included. The number of controls was 2,511 for all analyses.
[a]Mantel-Haenszel test for linear association.

total of 1,091 children. The rs6967330 risk allele (A) was associated with greater risk of asthma hospitalizations (hazards ratio (HR) = 1.7 (95% confidence interval (CI) = 1.2–2.4), $P = 0.002$) and severe exacerbations (HR = 1.4 (95% CI = 1.1–1.9), $P = 0.007$) in combined analyses (**Fig. 2**, **Supplementary Fig. 5** and **Supplementary Table 6**).

In $COPSAC_{2000}$, we observed a trend in the direction of increased neonatal bronchial responsiveness associated with the rs6967330 risk allele ($P = 0.10$) (**Supplementary Table 7**). There was no association with eczema in any of the three birth cohorts, and data on allergic sensitization were inconsistent (**Supplementary Table 6**).

The top SNP at the *CDHR3* locus (rs6967330) is a nonsynonymous coding SNP, where the risk allele (A), corresponding to the minor allele, results in an amino acid change from cysteine to tyrosine at position 529. This SNP is the only known nonsynonymous variant in this linkage disequilibrium (LD) region, but there are other variants located within Encyclopedia of DNA Elements (ENCODE)-predicted regulatory regions that are in moderate to high LD ($r^2 > 0.5$) with the sentinel SNP (**Supplementary Table 8**). Two SNPs with partial LD ($r^2 = 0.71$ and 0.58) were also associated with asthma in the discovery analysis but with less statistical significance. A similar association pattern with rs6967330 as the top SNP was observed in the GABRIEL (replication) study (**Supplementary Fig. 6**) and in the Generation R (replication) subsample of individuals with non-European ancestry (**Supplementary Fig. 7**), suggesting that rs6967330 might be the causal gene variant at this locus.

We investigated the potential functional consequences of the top variant in *CDHR3* (rs6967330; p.Cys529Tyr) by generating an expression construct encoding tagged human CDHR3 and introducing the mutation encoding p.Cys529Tyr (A allele at rs6967330 resulting in mutation of cysteine 529 to tyrosine) by site-directed mutagenesis. We transfected the constructs for wild-type and mutant CDHR3 into 293T cells. Consistent results from six independent experiments involving flow cytometry ($n = 3$) (**Supplementary Fig. 8**) and immunofluorescence staining ($n = 3$) (**Supplementary Fig. 9**) showed that the wild-type protein was expressed at very low levels at the cell surface, whereas the Cys529Tyr mutant showed a marked increase in cell surface expression (**Supplementary Note**). These results support the possibility that rs6967330 represents the causal variant at this locus. A recent study[17] reported that a SNP (rs17152490) in high LD ($r^2 = 0.69$) with our top SNP was associated with lung expression of *CDHR3*, further supporting a functional role for this locus.

CDHR3 is a transmembrane protein with six extracellular cadherin domains. Protein structure modeling showed that the risk-associated alteration (p.Cys529Tyr) was located at the interface between two membrane-proximal cadherin domains, D5 and D6 (**Fig. 3**). Interestingly, Cys592 and Cys566, which are expected to form a disulfide bridge within D6, are close to Cys529 in D5, and the short distance between them could allow disulfide rearrangement (for the wild-type, non-risk cysteine variant). The location of the variant residue at the domain interface suggests that the variant residue may interfere with interdomain stabilization, overall protein stability, folding or conformation, in agreement with the observation in our experimental studies of altered cell surface expression.

The biological function of CDHR3 is unknown, but it belongs to the cadherin family of transmembrane proteins involved in homologous cell adhesion and important for several cellular processes, including epithelial polarity, cell-cell interaction and differentiation[18]. Other members of the cadherin family have been associated with asthma

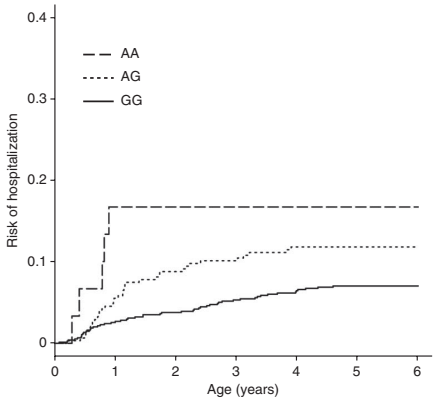


Figure 2 Cumulative risk of asthma hospitalization during the first 6 years of life stratified on *CDHR3* (rs6967330) genotype. Data are from combined analysis of the $COPSAC_{2000}$ and MAAS birth cohorts (replication), including a total of 1,091 children, of whom 92 were hospitalized for asthma. Genotype distribution was as follows: AA, 30 individuals; AG, 312 individuals; GG, 749 individuals. The *P* value for the association between genotype and risk of hospitalization was 0.002 (Cox regression analysis using an additive genetic model).

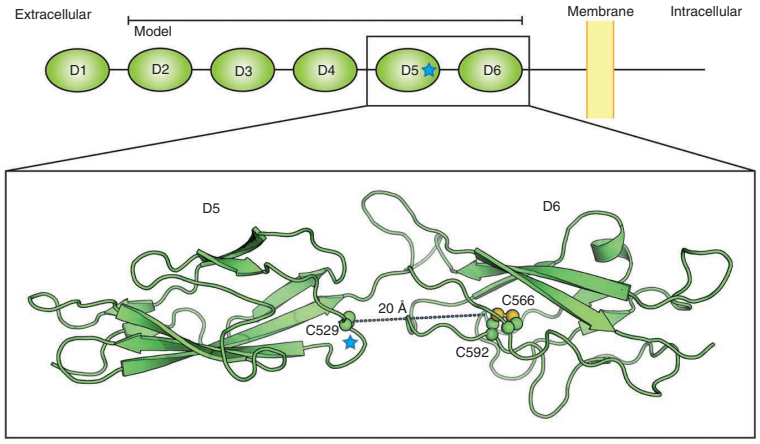


Figure 3 Overview of the CDHR3 protein model. The model covers cadherin domains 2–6 (D2–D6) and is based on the structure of the entire mouse N-cadherin ectodomain (Protein Data Bank (PDB) 3Q2W; domains 1–5). The location of the alteration at position 529 is indicated with a blue star. The distance between residue 529 and the disulfide bridge in D6 (between residues 566 and 592) is approximately 20 Å.

and related traits, including E-cadherin[19] and protocadherin-1 (ref. 20).

We demonstrated protein expression of CDHR3 in bronchial epithelium from adults and in fetal lung tissue (**Supplementary Fig. 10**). *CDHR3* was previously found to be highly expressed in normal human lung tissue[21] and specifically in the bronchial epithelium[22]. *CDHR3* (probe 235650_at) was upregulated by tenfold in differentiating epithelial cells (with a rank of 123 out of more than 47,000 transcripts ranked by magnitude of upregulation)[23] and seems to be highly expressed in the developing human lung[24].

There is an increasing focus on the role of the airway epithelium in asthma pathogenesis. Structural or functional abnormalities in the epithelium may increase susceptibility to environmental stimuli by exaggerating immune responses and structural changes in underlying tissues and increasing airway reactivity[25]. Epithelial integrity is dependent on the interaction of proteins in cell-cell junction complexes, including adhesion molecules. Studies have shown impaired tight junction function[26] and reduced E-cadherin expression[27] in the airway epithelium of individuals with asthma. *CDHR3* is a plausible candidate gene for asthma because of its high level of expression in the airway epithelium and the known role of cadherins in cell adhesion and interaction. Most asthma exacerbations in children are caused by respiratory infections, predominantly common viral infections such as rhinovirus[28], but bacterial infection may also have a role[29], as well as exposure to air pollution[30]. It is therefore plausible that *CDHR3* variation increases susceptibility to respiratory infections or other airway irritants through impaired epithelial integrity and/or disordered repair processes.

Interestingly, the *CDHR3* asthma risk allele is the ancestral allele. Public data from protein databases suggest that humans are unique among 36 other vertebrate species in having the derived (non-risk) allele resulting in a cysteine at position 529 (**Supplementary Table 9**), which is now the wild-type allele in most human populations (Human Genome Diversity Project (HGDP) selection browser; see URLs). This finding suggests that the risk (ancestral) allele, associated with increased surface expression of CDHR3, may have been advantageous during early human evolution. This phenomenon in which the ancestral allele is the risk allele is known for other common diseases and may reflect a shift from a beneficial to a deleterious effect for a particular allele as a result of a changing environment[31].

The *CDHR3* variant seems to be associated with an asthma phenotype of early onset, as demonstrated by the strongest replication of association in the GABRIEL stratum with asthma onset before 16 years of age (**Supplementary Table 10**) and in the second replication including children with asthma onset before 6 years of age (**Table 1**). Increased risk was already demonstrated in the first year of life (**Fig. 2**), particularly in children who were homozygous for the risk allele (A). This finding is in line with the tendency toward association of increased airway reactivity in neonates with the risk allele and findings of *CDHR3* expression in the fetal lung. *CDHR3* variation also seems to be more strongly associated with an asthma phenotype with exacerbations (**Supplementary Table 6**), particularly with recurrent exacerbations (**Table 2** and **Supplementary Table 6**).

The top locus in this study, on chromosome 17q12-21, has consistently been associated with childhood-onset asthma[11,13]. The effect size in the present study is remarkably high, with an OR of 2.3 that increases to 2.7 for the children with the highest number of exacerbations. This finding suggests a key role for this locus in severe exacerbations in early childhood, in line with a previous report from the $COPSAC_{2000}$ birth cohort study[32].

Genome-wide significant association with asthma has previously been shown for variants in or near *IL33*, *RAD50-IL13* and *IL1RL1* (refs. 11,33). The fact that the top loci in our study were generally shared with previous GWAS of asthma suggests that early-onset asthma with severe exacerbations is at least partly driven by multiple common variants in the same genes that contribute to asthma without severe exacerbations.

The sample size of the present GWAS was less than one-fifth that of the largest published GWAS of asthma (GABRIEL)[11], and, yet, we found a similar number of genome-wide significant loci, similar statistical significance and considerably larger effect estimates. Further increasing phenotypic specificity by stratified analysis in the 358 children with the highest number of exacerbations resulted in an additional increase in effect estimates, with ORs between 1.6 and 2.7 per risk allele, and strong statistical significance. Effect estimates were also higher than previously reported when replicating the exact top SNP from the GABRIEL study (**Supplementary Table 11**). This finding demonstrates that specific phenotyping is a helpful approach in the search for asthma susceptibility genes. The narrow age criteria (2–6 years) for disease may be an important phenotypic characteristic, as heritability has been demonstrated to be higher for early-onset asthma[34].

The method of case identification through national registries allowed us to define a specific and rare phenotype of repeated acute hospitalizations in young children from 2 to 6 years of age, which, to our knowledge, has not previously been done in a GWAS. One limitation of this study is that we had relatively poor genome-wide coverage (approximately 125,000 SNPs).

In conclusion, our results demonstrate the strength of specific phenotyping in genetic studies of asthma. Future research focusing on understanding the role of *CDHR3* variants in the development of asthma and severe exacerbations may increase understanding and improve treatment of this clinically important disease entity.

URLs. HGDP selection browser data for rs6967330, http://hgdp.uchicago.edu/cgi-bin/alfreqs.cgi?pos=105445687&chr=chr7&rs=rs6967330&imp=false.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

A full list of acknowledgments for each study is given in the **Supplementary Note**.

AUTHOR CONTRIBUTIONS

K.B. was the main author responsible for designing the study, analyzing and interpreting data, writing the manuscript and directing the work. He had full access to the data and final responsibility for the decision to submit the work for publication. H.B. contributed to design of the study, analysis of data and writing of the manuscript. P.S. and H.H. contributed to design of the study and analysis of data in relation to whole-genome genotyping. K.N. performed the GWAS analysis and contributed to regional imputation. E.K.-M., A. Sevelsted, M.A.R., R.Y. and R.G. contributed to data analysis. J.M.M., S.B.-G. and D.T. directed and contributed to regional imputation and data analyses. M.V.H. and D.M.H. were responsible for subject identification, collection of dried blood spots and DNA extraction and amplification. K.B., E.K.-M., L.J.M., R.F. and A.M. contributed to data acquisition. T.B. performed modeling of the CDHR3 protein structure. L.P., C.H. and E.A.N. were responsible for data from the discovery control cohort. H.H. and M.E.M. were responsible for the functional studies of the CDHR3 variant involving flow cytometry. A.H., D.E.S. and D.E.D. were responsible for the experimental studies involving immunofluorescence staining. A. Simpson, A.C. and D.B. were responsible for data from the MAAS cohort. H.T.d.D., L.D. and V.W.V.J. were responsible for data from the Generation R cohort. G.F.-T., P.M.L. and J.W.H. were responsible for the studies of lung tissue. P.F.T. studied the evolutionary aspects of the *CDHR3* risk variant (rs6967330). All coauthors provided important intellectual input to the study and approved the final version of the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Kocevar, V.S. *et al.* Variations in pediatric asthma hospitalization rates and costs between and within Nordic countries. *Chest* **125**, 1680–1684 (2004).
2. Lozano, P., Sullivan, S.D., Smith, D.H. & Weiss, K.B. The economic burden of asthma in US children: estimates from the National Medical Expenditure Survey. *J. Allergy Clin. Immunol.* **104**, 957–963 (1999).
3. Matterne, U., Schmitt, J., Diepgen, T.L. & Apfelbacher, C. Children and adolescents' health-related quality of life in relation to eczema, asthma and hay fever: results from a population-based cross-sectional study. *Qual. Life Res.* **20**, 1295–1305 (2011).
4. Smith, D.H. *et al.* A national estimate of the economic costs of asthma. *Am. J. Respir. Crit. Care Med.* **156**, 787–793 (1997).
5. Bush, A. Practice imperfect—treatment for wheezing in preschoolers. *N. Engl. J. Med.* **360**, 409–410 (2009).
6. Duffy, D.L., Martin, N.G., Battistutta, D., Hopper, J.L. & Mathews, J.D. Genetics of asthma and hay fever in Australian twins. *Am. Rev. Respir. Dis.* **142**, 1351–1358 (1990).
7. van Beijsterveldt, C.E. & Boomsma, D.I. Genetics of parentally reported asthma, eczema and rhinitis in 5-yr-old twins. *Eur. Respir. J.* **29**, 516–521 (2007).
8. Ferreira, M.A. *et al.* Identification of *IL6R* and chromosome 11q13.5 as risk loci for asthma. *Lancet* **378**, 1006–1014 (2011).
9. Gudbjartsson, D.F. *et al.* Sequence variants affecting eosinophil numbers associate with asthma and myocardial infarction. *Nat. Genet.* **41**, 342–347 (2009).
10. Himes, B.E. *et al.* Genome-wide association analysis identifies *PDE4D* as an asthma-susceptibility gene. *Am. J. Hum. Genet.* **84**, 581–593 (2009).
11. Moffatt, M.F. *et al.* A large-scale, consortium-based genomewide association study of asthma. *N. Engl. J. Med.* **363**, 1211–1221 (2010).
12. Sleiman, P.M. *et al.* Variants of *DENND1B* associated with asthma in children. *N. Engl. J. Med.* **362**, 36–44 (2010).
13. Torgerson, D.G. *et al.* Meta-analysis of genome-wide association studies of asthma in ethnically diverse North American populations. *Nat. Genet.* **43**, 887–892 (2011).
14. Anderson, G.P. Endotyping asthma: new insights into key pathogenic mechanisms in a complex, heterogeneous disease. *Lancet* **372**, 1107–1119 (2008).
15. Hollegaard, M.V. *et al.* Genome-wide scans using archived neonatal dried blood spot samples. *BMC Genomics* **10**, 297 (2009).
16. Hollegaard, M.V. *et al.* Robustness of genome-wide scanning using archived dried blood spot samples as a DNA source. *BMC Genet.* **12**, 58 (2011).
17. Hao, K. *et al.* Lung eQTLs to help reveal the molecular underpinnings of asthma. *PLoS Genet.* **8**, e1003029 (2012).
18. Hulpiau, P. & van Roy, F. Molecular evolution of the cadherin superfamily. *Int. J. Biochem. Cell Biol.* **41**, 349–369 (2009).
19. Nawijn, M.C., Hackett, T.L., Postma, D.S., van Oosterhout, A.J. & Heijink, I.H. E-cadherin: gatekeeper of airway mucosa and allergic sensitization. *Trends Immunol.* **32**, 248–255 (2011).
20. Koppelman, G.H. *et al.* Identification of *PCDH1* as a novel susceptibility gene for bronchial hyperresponsiveness. *Am. J. Respir. Crit. Care Med.* **180**, 929–935 (2009).
21. Yanai, I. *et al.* Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21**, 650–659 (2005).
22. McCall, M.N., Uppal, K., Jaffee, H.A., Zilliox, M.J. & Irizarry, R.A. The Gene Expression Barcode: leveraging public data repositories to begin cataloging the human and murine transcriptomes. *Nucleic Acids Res.* **39**, D1011–D1015 (2011).
23. Ross, A.J., Dailey, L.A., Brighton, L.E. & Devlin, R.B. Transcriptional profiling of mucociliary differentiation in human airway epithelial cells. *Am. J. Respir. Cell Mol. Biol.* **37**, 169–185 (2007).
24. Kho, A.T. *et al.* Transcriptomic analysis of human lung development. *Am. J. Respir. Crit. Care Med.* **181**, 54–63 (2010).
25. Holgate, S.T. The sentinel role of the airway epithelium in asthma pathogenesis. *Immunol. Rev.* **242**, 205–219 (2011).
26. Xiao, C. *et al.* Defective epithelial barrier function in asthma. *J. Allergy Clin. Immunol.* **128**, 549–556 (2011).
27. de Boer, W.I. *et al.* Altered expression of epithelial junctional proteins in atopic asthma: possible role in inflammation. *Can. J. Physiol. Pharmacol.* **86**, 105–112 (2008).
28. Johnston, S.L. *et al.* Community study of role of viral infections in exacerbations of asthma in 9–11 year old children. *Br. Med. J.* **310**, 1225–1229 (1995).
29. Bisgaard, H. *et al.* Association of bacteria and viruses with wheezy episodes in young children: prospective birth cohort study. *Br. Med. J.* **341**, c4978 (2010).
30. Iskandar, A. *et al.* Coarse and fine particles but not ultrafine particles in urban air trigger hospital admission for asthma in children. *Thorax* **67**, 252–257 (2012).
31. Di Rienzo, A. & Hudson, R.R. An evolutionary framework for common diseases: the ancestral-susceptibility model. *Trends Genet.* **21**, 596–601 (2005).
32. Bisgaard, H. *et al.* Chromosome 17q21 gene variants are associated with asthma and exacerbations but not atopy in early childhood. *Am. J. Respir. Crit. Care Med.* **179**, 179–185 (2009).
33. Li, X. *et al.* Genome-wide association study of asthma identifies *RAD50-IL13* and *HLA-DR/DQ* regions. *J. Allergy Clin. Immunol.* **125**, 328–335 (2010).
34. Thomsen, S.F., Duffy, D.L., Kyvik, K.O. & Backer, V. Genetic influence on the age at onset of asthma: a twin study. *J. Allergy Clin. Immunol.* **126**, 626–630 (2010).

ONLINE METHODS

The individual studies are described in further detail in the **Supplementary Note**.

$COPSAC_{exacerbation}$ cohort (GWAS). This is a register-based cohort of children with asthma who were identified and characterized from national health registries. The study was approved by the Ethics Committee for Copenhagen (H-B-2998-103) and the Danish Data Protection Agency (2008-41-2622). According to Danish law, research ethics committees can grant exemption from obtaining informed consent for research projects based on biobank material under certain circumstances. For this study, such an exemption was granted (H-B-2998-103).

Case selection. Children with repeated acute hospitalizations (cases) were identified in the Danish National Patient Register covering all diagnoses of discharges from Danish hospitals[35]. Information on birth-related events was obtained from the national birth register. Inclusion criteria were at least two acute hospitalizations for asthma (ICD8-codes 493, ICD-10 codes J45-46) from 2 to 6 years of age (both years included). Duration of hospitalization had to be more than 1 d, and two hospitalizations had to be separated by at least 6 months. Exclusion criteria were side diagnosis during hospitalization, registered chronic diagnosis considered to affect risk of hospitalization for asthma, low birth weight (<2.5 kg) or gestational age of under 36 weeks at birth. Cases were further characterized with respect to the number of hospitalizations from asthma and acute bronchitis and for concurrent atopy.

DNA sampling and genotyping of cases. DNA was obtained from blood spots sampled as part of the Danish neonatal screening program and stored in the Danish Neonatal Screening Biobank[36]. Two disks, each 3.2 mm in diameter, were punched from each blood spot. DNA was extracted, and the whole genome for each individual sample was amplified in triplicate as previously described[15,16]. Cases were genotyped on the Affymetrix Axiom CEU array (567,090 SNPs). Top SNPs from the five genome-wide significant loci were regenotyped with the PCR KASPar genotyping system (KBiosciences) to validate the results (**Supplementary Table 3**). Two additional SNPs in the proximity of the newly discovered *CDHR3* variant were genotyped for further exploration of the region encompassing it.

Controls. The control population was randomly drawn from two large Danish cohorts: the Danish National Birth Cohort (females) and the Copenhagen draft board examinations (males). Individuals who indicated in a questionnaire that they had physician-diagnosed asthma were excluded. Genome-wide genotyping had previously been performed as part of the Genomics of Overweight in Young Adults (GOYA) study[37] on the Illumina Human610-Quad v1.0 BeadChip (545,350 SNPs). Potential bias introduced by differences in chemistry between the different platforms used for cases and controls (Affymetrix and Illumina, respectively) was investigated by also using control data from the Wellcome Trust Case Control Consortium 2 (WTCCC2) project that performed genotyping on an Affymetrix platform (Affymetrix 6.0) (**Supplementary Table 4**).

Replication in a previously published GWAS. Replication of the five genome-wide significant loci from the discovery analysis was sought in publically available data from a GWAS performed by the GABRIEL Consortium[11]. This replication included 19 studies of childhood-onset asthma (onset before 16 years of age) with a total of 6,783 cases and 7,720 controls.

Replication in birth cohorts for the *CDHR3* top SNP. *The $COPSAC_{2000}$ replication cohort*. Replication and phenotypic characterization of the *CDHR3* risk locus were sought in the $COPSAC_{2000}$ cohort, a prospective clinical study of a birth cohort of 411 children. This cohort is not overlapping with the $COPSAC_{exacerbation}$ discovery study. The $COPSAC_{2000}$ cohort study was approved by the Ethics Committee for Copenhagen (KF 01-289/96) and the Danish Data Protection Agency (2008-41-1754), and informed consent was obtained from both parents of each child. All mothers had a history of a doctor's diagnosis of asthma after 7 years of age. Newborns were enrolled in the first month of life, as previously described in detail[38–40]. This cohort is characterized by deep phenotyping during close clinical follow-up. Doctors employed in the clinical research unit were acting primary physicians for the children from the cohort and diagnosed and treated respiratory and skin symptoms, and asthmatic symptoms were recorded in daily diaries[41].

Acute, severe exacerbations from birth to 6 years of age were defined as requiring the use of oral prednisolone or high-dose inhaled corticosteroid for wheezy symptoms, prescribed at the discretion of the doctor in the clinical research unit, or by acute hospitalization at a local hospital for such symptoms[32]. Asthma from birth to 7 years of age was diagnosed on the basis of predefined algorithms of symptoms and response to treatment, as previously described[40].

Neonatal spirometry and analysis of neonatal bronchial responsiveness to methacholine were carried out by 4 weeks of age, applying the raised volume, rapid thoracic compression technique. Lung function was measured by spirometry in the child's seventh year of life. Specific airway resistance (sR_{aw}) was measured at 4 and 6 years by whole-body plethysmography. Bronchial responsiveness at ages 4 and 6 years was determined as the relative change in sR_{aw} after hyperventilation of cold, dry air.

Allergic sensitization against common inhalant allergens was determined at 6 years of age by measurement of serum-specific IgE levels. Atopic dermatitis was diagnosed using the Hanifin-Rajka criteria[42] from birth to 7 years of age.

High-throughput genome-wide SNP genotyping was performed using the Illumina Infinium II HumanHap550 v1, v3 or Quad BeadChip platform at the Children's Hospital of Philadelphia's Center for Applied Genomics. We excluded SNPs with call rate of <95%, minor allele frequency (MAF) of <1% or Hardy-Weinberg equilibrium P value of $<1 \times 10^{-5}$. rs6967330 was a genotyped SNP on this array.

MAAS replication cohort. The Manchester Asthma and Allergy Study is a population-based birth cohort described in detail elsewhere[43]. Subjects were recruited prenatally and were followed prospectively. The study was approved by the local research ethics committee (South Manchester, reference 03/SM/400). Parents gave written informed consent. Participants attended follow-up at ages 1, 3 and 5 years of age.

For asthma, validated questionnaires were administered by interviewers to collect information on parentally reported symptoms, physician-diagnosed asthma and treatments received. 'Current wheeze and asthma treatment' was defined as parentally reported wheeze in the past 12 months. 'Asthma ever' was defined as positive if, at any given time point, two of three responses were positive to the following questions: "Has your child wheezed within the past 12 months?", "Does your child currently take asthma medication?" or "Has a doctor ever told you that your child has asthma?" Controls were defined as children with none of these symptoms.

For exacerbations, a pediatrician extracted data from primary-care medical records, including information on diagnosis with wheeze and/or asthma, all prescriptions (including inhaled corticosteroids (ICS) and β2 agonists), unscheduled visits and hospital admissions for asthma and/or wheeze during the first 8 years of life. Following American Thoracic Society guidelines, we defined asthma exacerbations by either admission to a hospital or an emergency department visit and/or by receipt of oral corticosteroids for at least 3 d[44].

DNA samples were genotyped on the Illumina Human610-Quad BeadChip. Genotypes were called using the Illumina GenCall application, following the manufacturer's instructions. Quality control criteria for samples included call rate of greater than 97%, exclusion of samples with outlier autosomal heterozygosity and sex validation. We excluded SNPs with call rate of <95%, Hardy-Weinberg equilibrium P value of $>5.9 \times 10^{-7}$ and MAF of <0.005. We then performed a look-up for SNP rs6967330, which showed a genotyping success rate of 100% and a Hardy-Weinberg equilibrium P value of 0.4164.

Generation R replication cohort. The Generation R Study is a population-based prospective cohort study of pregnant women and their children from fetal life onward in Rotterdam, The Netherlands[45]. The study protocol was approved by the Medical Ethical Committee of the Erasmus Medical Center, Rotterdam (MEC 217.595/2002/20). Written informed consent was obtained from all mothers and biological fathers or legal guardians. Information on wheezing, asthma and eczema was collected for the children by questionnaires at the ages of 1 to 4 and 6 years[46]. Questions about wheezing included: "Has your child had problems with a wheezing chest during the last year? (never, 1–3 times, >4 times) (age 1 to 4 years)" and "Did your child ever suffer from chest wheezing? (never, 1–3 times,

>4 times) (age 6 years)." Questions about asthma included: "Has a doctor diagnosed your child as having asthma during the past year? (yes, no) (age 2 and 4 years)" and "Was your child ever diagnosed with asthma by a doctor? (yes, no) (age 3 and 6 years)." On the basis of the last obtained questionnaire, we grouped children as having 'asthma ever before 6 years of age'. Reported asthma at 2, 3 or 4 years of age was used to reclassify children included in this group where appropriate. We then recategorized children as those with an asthma diagnosis before 3 years of age and at 3 years of age or older. Reported numbers of wheezing episodes at 1 and 2 years of age and at 3 to 6 years of age, respectively, were used to reclassify asthma diagnosis before and at 3 years of age into 'asthma diagnosis or ≥3 episodes of wheezing before 3 years of age'. Questions about eczema included: "Has a doctor diagnosed your child as having eczema during the past year? (yes, no) (age 1 to 4 years)" and "Was your child ever diagnosed with eczema by a doctor? (yes, no) (6 years)." As with asthma, we grouped children into those with 'eczema ever before 6 years of age' on the basis of the last obtained questionnaire and used reported eczema at 1 or 4 years of age to reclassify children included in this group where appropriate.

Samples were genotyped using Illumina Infinium II HumanHap610 Quad arrays, following standard manufacturer's protocols. Intensity files were analyzed using BeadStudio Genotyping Module software v.3.2.32, and genotypes were called using default cluster files. Any sample with a call rate of less than 97.5%, excess autosomal heterozygosity (F < mean − 4 s.d.) or mismatch between called and phenotypic sex was excluded. rs6967330 was a genotyped SNP in this set. Individuals identified as genetic outliers by identity-by-state (IBS) clustering analysis (>3 s.d. away from the mean for the HapMap CEU population (Utah residents of Northern and Western European ancestry)) were considered to have non-European ancestry. Ancestry determination analysis included genomic data from all Generation R individuals merged with data for three reference panels from Phase 2 of the HapMap Project (YRI (Yoruba from Ibadan, Nigeria), CHB + JPT (Han Chinese in Beijing, China, and Japanese in Tokyo, Japan) and CEU). Analysis of association between an asthma or eczema phenotype and GWAS SNPs was carried out using a regression framework, adjusting for population stratification in the Generation R cohort using MACH2QTL, as implemented in GRIMP. Ten genomic principal components obtained after the application of SNP quality exclusion criteria and LD pruning were used to adjust for population substructure in the combined population, four principal components were used for the European subpopulation and eight principal components were used for the non-European subpopulation. Individuals were grouped as having European (n = 1,962; 64.5%) or non-European (n = 1,078; 35.5%) ancestry on the basis of genetic ancestry. On the basis of information on the country of birth of parents and grandparents obtained by questionnaires, the largest non-European ancestry groups included individuals of Turkish (5.4%), Surinamese (4.6%), Dutch Antillean (4.0%), Moroccan (2.9%) and Cape Verdean (2.3%) origin.

Statistical analyses. *Genome-wide association analysis.* Quality control was carried out separately on cases and controls. This included filtering on SNP call rate (>99%) and sample call rate (>98%) and tests for excess heterozygosity, deviation from Hardy-Weinberg equilibrium, sex mismatch and familial relatedness. Non-European individuals were excluded on the basis of deviation from the HapMap CEU reference panel (release 22). Indication of population stratification or genotyping bias was tested by multidimensional scaling (MDS) after quality control. This analysis showed evidence of association with disease status for the first seven MDS components, and these were therefore included as covariates in the association analysis. Additional analyses including the first 100 MDS components did not materially alter the results. Merged data for SNPs present on both arrays after quality control were used for association testing with PLINK (v. 1.07) using a logistic additive model, adjusting for the first seven MDS components. Additional quality control was performed for genome-wide significant SNPs after association analysis, including a test for genotyping batch effects, resulting in the removal of one genome-wide significant SNP with strong evidence of batch-related genotyping error.

Functional annotation for the SNPs in LD (r^2 > 0.5) with the *CDHR3* top SNP (rs6967330) was obtained from the RefSeq track downloaded from the UCSC Genome Browser. SNPs were associated with regulatory elements by HaploReg[47] in terms of predicted ENCODE chromatin state (promoter and enhancer histone modification signals) and DNase I hypersensitivity (**Supplementary Table 8**).

Regional imputation was performed to describe the identified loci from the discovery analysis (**Supplementary Fig. 3**) as well as reported loci from the previous largest published GWAS (GABRIEL)[11] (**Supplementary Table 11**). We used two-step genotype imputation as described[48]. We used the SHAPEIT algorithm to prephase the haplotypes[49] and then used IMPUEv2 software for the imputation of unknown genotypes[50] separately in cases and controls. We used the 1000 Genomes Project reference panel[51] (April 2012 version). We used a strict cutoff (info of 0.88), which, according to our analyses, provides an allelic dosage R^2 correlation between real and imputed genotypes of greater than 0.8 and shows an optimal balance between sufficient accuracy and power[52]. We then compared the resulting allelic frequencies using SNPTEST 2.4.1 (ref. 53).

CDHR3 protein expression in experimental models. The top SNP at the *CDHR3* locus is a nonsynonymous SNP (encoding p.Cys529Tyr). To determine the functional consequences of the p.Cys529Tyr variant, we generated expression constructs encoding tagged human CDHR3 protein, and the mutation encoding the p.Cys529Tyr alteration was introduced by site-directed mutagenesis. Plasmids encoding wild-type or mutant *CDHR3* or empty vector were transfected into 293T cells, and cells were monitored for surface and intracellular expression of CDHR3 by flow cytometry. 293T cells were from the American Type Culture Collection (ATCC), catalog number CRL-3216. They were recently tested for mycoplasma contamination but were not authenticated. For protein blotting, cells expressing CDHR3 proteins were lysed, and whole-cell lysates were separated by SDS-PAGE under reducing or non-reducing conditions, transferred to PVDF membranes and blotted for Flag (anti-Flag antibody, clone M2 (Agilent Technologies, 200470-21) at a dilution of 1:2,000). For immunofluorescence and confocal microscopy, 293T cells were grown on glass coverslips in DMEM with 3 mM glutamine and 10% heat-inactivated FBS at 37 °C and 5% CO_2 before and for 2 d after transfection with expression constructs for Flag-tagged wild-type CDHR3 and CDHR3 Cys529Tyr using TransIT 2020 reagent according to a standard protocol (Mirus Bio). Cells were obtained and used at a low passage from ATCC and had recently been tested for mycoplasma. Cells were incubated in 10% serum-containing culture medium plus primary anti-Flag mouse antibodies (F3165, Sigma; 1:300 dilution) for 1 h at 37 °C before being washed briefly with culture medium. Cells were then stained with secondary rabbit anti-mouse antibodies (F0261, Daco; 1:600 dilution) conjugated with fluorescein isothiocyanate (FITC) with incubation at 37 °C for 30 min and washed with culture medium before PBS. Afterward, cells were fixed in 2% paraformaldehyde for 15 min, washed with PBS and permeabilized in 0.2% Triton X-100 in PBS for 5 min, washed and incubated with Cy3-conjugated mouse anti-Flag antibody (Cy3-labeled F3165, Sigma; 1:300 dilution). Finally, cells were mounted with ProLong Gold antifade reagent with DAPI (Invitrogen). Images were acquired using a Leica DMI 6000-B confocal microscope (Leica Microsystems) with 40× magnification and were processed in Photoshop (Adobe Systems). Experiments were performed in triplicate (independent transfections) for both flow cytometry and immunofluorescence staining. Data presented (**Supplementary Figs. 8** and **9**) were chosen as being representative of the repeated experiments.

CDHR3 protein structure modeling. A homology model of CDHR3 domains 2–6 (residues 141–681) was generated using the HHpred server[54]. The model was based on the structure of mouse N-cadherin (PDB 3Q2W) domains 1–5. A disulfide bridge was manually introduced in the final model between the structurally adjacent residues Cys566 and Cys592, as this corresponds to a disulfide bridge commonly observed in cadherin domains.

35. Lynge, E., Sandegaard, J.L. & Rebolj, M. The Danish National Patient Register. *Scand. J. Public Health* **39**, 30–33 (2011).
36. Nørgaard-Pedersen, B. & Hougaard, D.M. Storage policies and use of the Danish Newborn Screening Biobank. *J. Inherit. Metab. Dis.* **30**, 530–536 (2007).
37. Paternoster, L. *et al.* Genome-wide population-based association study of extremely overweight young adults—the GOYA study. *PLoS One* **6**, e24303 (2011).
38. Bisgaard, H. The Copenhagen Prospective Study on Asthma in Childhood (COPSAC): design, rationale, and baseline data from a longitudinal birth cohort study. *Ann. Allergy Asthma Immunol.* **93**, 381–389 (2004).

39. Bisgaard, H., Hermansen, M.N., Loland, L., Halkjaer, L.B. & Buchvald, F. Intermittent inhaled corticosteroids in infants with episodic wheezing. *N. Engl. J. Med.* **354**, 1998–2005 (2006).
40. Bisgaard, H. *et al.* Childhood asthma after bacterial colonization of the airway in neonates. *N. Engl. J. Med.* **357**, 1487–1495 (2007).
41. Bisgaard, H., Pipper, C.B. & Bonnelykke, K. Endotyping early childhood asthma by quantitative symptom assessment. *J. Allergy Clin. Immunol.* **127**, 1155–1164 (2011).
42. Hanifin, J.M. & Rajka, G. Diagnostic features of atopic dermatitis. *Acta Derm. Venereol.* **92**, 44–47 (1980).
43. Lowe, L. *et al.* Specific airway resistance in 3-year-old children: a prospective cohort study. *Lancet* **359**, 1904–1908 (2002).
44. Reddel, H.K. *et al.* An official American Thoracic Society/European Respiratory Society statement: asthma control and exacerbations: standardizing endpoints for clinical asthma trials and clinical practice. *Am. J. Respir. Crit. Care Med.* **180**, 59–99 (2009).
45. Jaddoe, V.W. *et al.* The Generation R Study Biobank: a resource for epidemiological studies in children and their parents. *Eur. J. Epidemiol.* **22**, 917–923 (2007).
46. Jaddoe, V.W. *et al.* The Generation R Study: design and cohort update 2012. *Eur. J. Epidemiol.* **27**, 739–756 (2012).
47. Ward, L.D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* **40**, D930–D934 (2012).
48. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G.R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **44**, 955–959 (2012).
49. Delaneau, O., Marchini, J. & Zagury, J.F. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179–181 (2012).
50. Howie, B., Marchini, J. & Stephens, M. Genotype imputation with thousands of genomes. *G3* **1**, 457–470 (2011).
51. Abecasis, G.R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
52. Auer, P.L. *et al.* Imputation of exome sequence variants into population- based samples and blood-cell-trait-associated loci in African Americans: NHLBI GO Exome Sequencing Project. *Am. J. Hum. Genet.* **91**, 794–808 (2012).
53. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11**, 499–511 (2010).
54. Söding, J., Biegert, A. & Lupas, A.N. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* **33**, W244–W248 (2005).

Appendix 2. Horikoshi M, Beaumont RN, Day FR, Warrington NM, Kooijman MN, Fernandez-Tajes J, Feenstra B, van Zuydam NR, Gaulton KJ, Grarup N … Bonàs-Guarch S … Freathy RM 2016. Genome-wide associations for birth weight and correlations with adult disease. *Nature* **538**: 248-252

Contribution of the PhD candidate

- Quality Control of genotyped data.
- Genotype imputations.
- Statistical analyses.

LETTER

doi:10.1038/nature19806

Genome-wide associations for birth weight and correlations with adult disease

Momoko Horikoshi[1,2]*, Robin N. Beaumont[3]*, Felix R. Day[4]*, Nicole M. Warrington[5,6]*, Marjolein N. Kooijman[7,8,9]*, Juan Fernandez-Tajes[1]*, Bjarke Feenstra[10], Natalie R. van Zuydam[1,2], Kyle J. Gaulton[1,11], Niels Grarup[12], Jonathan P. Bradfield[13], David P. Strachan[14], Ruifang Li-Gao[15], Tarunveer S. Ahluwalia[12,16,17], Eskil Kreiner[16], Rico Rueedi[18,19], Leo-Pekka Lyytikäinen[20,21], Diana L. Cousminer[22,23,24], Ying Wu[25], Elisabeth Thiering[26,27], Carol A. Wang[6], Christian T. Have[12], Jouke-Jan Hottenga[28], Natalia Vilor-Tejedor[29,30,31], Peter K. Joshi[32], Eileen Tai Hui Boh[33], Ioanna Ntalla[34,35], Niina Pitkänen[36], Anubha Mahajan[1], Elisabeth M. van Leeuwen[8], Raimo Joro[37], Vasiliki Lagou[1,38,39], Michael Nodzenski[40], Louise A. Diver[41], Krina T. Zondervan[1,42], Mariona Bustamante[29,30,31,43], Pedro Marques-Vidal[44], Josep M. Mercader[45], Amanda J. Bennett[2], Nilufer Rahmioglu[1], Dale R. Nyholt[46], Ronald C. W. Ma[47,48,49], Claudia H. T. Tam[47], Wing Hung Tam[50], CHARGE Consortium Hematology Working Group†, Santhi K. Ganesh[51], Frank J. A. van Rooij[8], Samuel E. Jones[3], Po-Ru Loh[52,53], Katherine S. Ruth[3], Marcus A. Tuke[3], Jessica Tyrrell[3,54], Andrew R. Wood[3], Hanieh Yaghootkar[3], Denise M. Scholtens[40], Lavinia Paternoster[55,56], Inga Prokopenko[1,57], Peter Kovacs[58], Mustafa Atalay[37], Sara M. Willems[8], Kalliope Panoutsopoulou[59], Xu Wang[33], Lisbeth Carstensen[10], Frank Geller[10], Katharina E. Schraut[32], Mario Murcia[31,60], Catharina E. M. van Beijsterveldt[28], Gonneke Willemsen[28], Emil V. R. Appel[12], Cilius E. Fonvig[12,61], Caecilie Trier[12,61], Carla M. T. Tiesler[26,27], Marie Standl[26], Zoltán Kutalik[19,62], Sílvia Bonàs-Guarch[45], David M. Hougaard[63,64], Friman Sánchez[45,65], David Torrents[45,66], Johannes Waage[16], Mads V. Hollegaard[63,64]‡, Hugoline G. de Haan[15], Frits R. Rosendaal[15], Carolina Medina-Gomez[7,8,67], Susan M. Ring[55,56], Gibran Hemani[55,56], George McMahon[56], Neil R. Robertson[1,2], Christopher J. Groves[2], Claudia Langenberg[4], Jian'an Luan[4], Robert A. Scott[4], Jing Hua Zhao[4], Frank D. Mentch[13], Scott M. MacKenzie[41], Rebecca M. Reynolds[68], Early Growth Genetics (EGG) Consortium†, William L. Lowe Jr[69], Anke Tönjes[70], Michael Stumvoll[58,70], Virpi Lindi[37], Timo A. Lakka[37,71,72], Cornelia M. van Duijn[8], Wieland Kiess[73], Antje Körner[58,73], Thorkild I. A. Sørensen[55,56,74,75], Harri Niinikoski[76,77], Katja Pahkala[36,78], Olli T. Raitakari[36,79], Eleftheria Zeggini[59], George V. Dedoussis[35], Yik-Ying Teo[33,80,81], Seang-Mei Saw[33,82], Mads Melbye[10,83,84], Harry Campbell[32], James F. Wilson[32,85], Martine Vrijheid[29,30,31], Eco J. C. N. de Geus[28,86], Dorret I. Boomsma[28], Haja N. Kadarmideen[87], Jens-Christian Holm[12,61], Torben Hansen[12], Sylvain Sebert[57,88,89], Andrew T. Hattersley[3], Lawrence J. Beilin[90], John P. Newnham[6], Craig E. Pennell[6], Joachim Heinrich[26,91], Linda S. Adair[92], Judith B. Borja[93,94], Karen L. Mohlke[25], Johan G. Eriksson[95,96,97], Elisabeth Widén[22], Mika Kähönen[98,99], Jorma S. Viikari[100,101], Terho Lehtimäki[20,21], Peter Vollenweider[44], Klaus Bønnelykke[16], Hans Bisgaard[16], Dennis O. Mook-Kanamori[15,102,103], Albert Hofman[7,8], Fernando Rivadeneira[7,8,67], André G. Uitterlinden[7,8,67], Charlotta Pisinger[104], Oluf Pedersen[12], Christine Power[105], Elina Hyppönen[105,106,107], Nicholas J. Wareham[4], Hakon Hakonarson[13,23,108], Eleanor Davies[41], Brian R. Walker[68], Vincent W. V. Jaddoe[7,8,9], Marjo-Riitta Järvelin[88,89,109,110], Struan F. A. Grant[13,23,108,111], Allan A. Vaag[83,112,113], Debbie A. Lawlor[55,56], Timothy M. Frayling[3], George Davey Smith[55,56], Andrew P. Morris[1,114,115]§, Ken K. Ong[4,116]§, Janine F. Felix[7,8,9]§, Nicholas J. Timpson[55,56]§, John R. B. Perry[4]§, David M. Evans[5,55,56]§, Mark I. McCarthy[1,2,117]§ & Rachel M. Freathy[3,55]§

Birth weight (BW) has been shown to be influenced by both fetal and maternal factors and in observational studies is reproducibly associated with future risk of adult metabolic diseases including type 2 diabetes (T2D) and cardiovascular disease[1]. These life-course associations have often been attributed to the impact of an adverse early life environment. Here, we performed a multi-ancestry genome-wide association study (GWAS) meta-analysis of BW in 153,781 individuals, identifying 60 loci where fetal genotype was associated with BW ($P < 5 \times 10^{-8}$). Overall, approximately 15% of variance in BW was captured by assays of fetal genetic variation. Using genetic association alone, we found strong inverse genetic correlations between BW and systolic blood pressure ($R_g = -0.22$, $P = 5.5 \times 10^{-13}$), T2D ($R_g = -0.27$, $P = 1.1 \times 10^{-6}$) and coronary artery disease ($R_g = -0.30$, $P = 6.5 \times 10^{-9}$). In addition, using large-cohort datasets, we demonstrated that genetic factors were the major contributor to the negative covariance between BW and future cardiometabolic risk. Pathway analyses indicated that the protein products of genes within BW-associated regions were enriched for diverse processes including insulin signalling, glucose homeostasis, glycogen biosynthesis and chromatin remodelling. There was also enrichment of associations with BW in known imprinted regions ($P = 1.9 \times 10^{-4}$). We demonstrate that life-course associations between early growth phenotypes and adult cardiometabolic disease are in part the result of shared genetic effects and identify some of the pathways through which these causal genetic effects are mediated.

We combined GWAS data for BW from 153,781 individuals representing multiple ancestries from 37 studies across three components (Extended Data Fig. 1 and Supplementary Table 1): (i) 75,891 individuals of European ancestry from 30 studies; (ii) 67,786 individuals of European ancestry from the UK Biobank; and (iii) 10,104 individuals of diverse ancestries (African American, Chinese, Filipino, Surinamese, Turkish and Moroccan) from six studies. Within each study, BW was *Z*-score transformed separately in males and females after excluding non-singletons and premature births and adjusting for gestational age where available. Genotypes were imputed using reference panels from the 1000 Genomes (1000G) Project[2] or combined 1000G and UK10K projects[3] (Supplementary Table 2). We performed quality control assessments to confirm that the distribution of BW was consistent across studies, irrespective of the data collection protocol, and confirmed that self-reported BW in the UK Biobank showed genetic and phenotypic associations consistent with those seen for measured BW in other studies[4] (Methods).

We identified 60 loci (of which 59 were autosomal) associated with BW at genome-wide significance ($P < 5 \times 10^{-8}$) in either the European

A list of affiliations appears in the online version of this paper.

ancestry or trans-ancestry meta-analyses (Extended Data Fig. 2a, Extended Data Table 1a and Supplementary Data; Methods). For lead single nucleotide polymorphisms (SNPs), we observed no heterogeneity in allelic effects between the three study components (Cochran's Q statistic $P > 0.00083$) (Supplementary Table 3). We found that 53 of these loci were novel in that the lead SNP mapped >2 Mb away from, and was independent ($R^2 < 0.05$ in the European (EUR) component of 1000G) of, the seven previously reported BW signals[5], all of which were confirmed in this larger analysis (Supplementary Table 4). Approximate conditional analysis in the European ancestry data indicated that three of these novel loci (near *ZBTB7B, HMGA1* and *PTCH1*) harboured multiple distinct association signals that attained genome-wide significance ($P < 5 \times 10^{-8}$) (Methods, Supplementary Table 5 and Extended Data Fig. 3).

The lead variants for most signals mapped to non-coding sequences, and at only two loci, *ADRB1* (rs7076938; $R^2 = 0.99$ with *ADRB1* G389R) and *NRIP1* (rs2229742, R448G), did the association data point to potential causal non-synonymous coding variants (Supplementary Table 6 and Methods). Lead SNPs for all but two loci (those mapping near *YKT6–GCK* and *SUZ12P1–CRLF3*) were common (minor allele frequency (MAF) $\geq$ 5%) with individually modest effects on BW ($\beta = 0.020$–0.053 standard deviations (s.d.) per allele, equivalent to 10–26 g). This was despite the much-improved coverage of low-frequency variants in this study (compared to previous HapMap 2 imputed meta-analyses, ref. 5) reflecting imputation from larger, and more complete, reference panels (Extended Data Table 1b). Indeed, all but five of the common variant association signals were tagged by variants (EUR $R^2 > 0.6$) in the HapMap 2 reference panel (Supplementary Tables 4, 5), indicating that most of the novel discoveries in the present study were driven by increased sample size[5]. Fine-mapping analysis yielded 14 regions in which fewer than ten variants contributed to the locus-specific credible sets that accounted for >99% of the posterior probability of association (Methods and Supplementary Table 7). The greatest refinement was at *YKT6–GCK*, where the credible set included only the low frequency variant rs138715366, which maps intronic to *YKT6*. These credible-set variants collectively showed enrichment for overlap with DNaseI hypersensitivity sites, particularly those generated, by ENCODE, from fetal (4.2-fold, 95% CI 1.8–10.7) and neonatal tissues (4.9-fold, 1.8–11.0) (Supplementary Fig. 1, Supplementary Table 8 and Methods).

In combination, the 62 distinct genome-wide significant signals at the 59 autosomal loci explained at least $2.0 \pm 1.1\%$ (standard error (s.e.)) of variance in BW (Supplementary Table 9 and Methods), which is similar in magnitude to that attributable to sex or maternal body mass index (BMI)[5]. However, the variance in BW captured collectively by all autosomal genotyped variants on the array was considerably larger, estimated at $15.1 \pm 0.9\%$ in the UK Biobank (Methods). These figures are consistent with a large number of genetic variants with smaller effects contributing to variation in BW.

Associations between fetal genotype and BW could result from indirect effects of the maternal genotype influencing BW via the intrauterine environment, given the correlation ($R \approx 0.5$) between maternal and fetal genotype. However, two lines of evidence indicated that variation in the fetal genome was the predominant driver of BW associations. First, an analysis of the global contribution of maternal versus fetal genetic variation, using a maternal genome-wide complex trait analysis (GCTA) model (ref. 6) (Methods) applied to 4,382 mother–child pairs, estimated that the child's genotype ($\sigma_C^2 = 0.24 \pm 0.11$) made a larger contribution to BW variance than either the mother's genotype ($\sigma_M^2 = 0.04 \pm 0.10$), or the covariance between the two ($\sigma_{CM} = 0.04 \pm 0.08$). Second, when we compared the point estimates of the BW-effect size dependent on maternal genotype at each of the 60 loci (as measured in up to 68,254 women[7]) with those dependent on fetal genotype (using European ancestry data from 143,677 individuals in the present study), fetal variation had a greater impact than maternal variation at 93% of the loci (55 out of 60; binomial $P = 10^{-11}$) (Supplementary Table 10, Extended Data Figs 4, 5 and Methods). The power to further disentangle maternal and fetal contributions using analyses of fetal genotype which were conditional on maternal genotype was constrained by the limited sample size available ($n = 12{,}909$ mother–child pairs) (Supplementary Table 11).

Collectively, these analyses provide evidence that the fetal genotype has a substantial impact on early growth, as measured by BW. We used these genetic associations to understand the causal relationships underlying observed associations between BW and disease, and to characterize the processes responsible.

To quantify the shared genetic contribution to BW and other health-related traits, we estimated their genetic correlations using linkage-disequilibrium score regression[8] (Methods). BW (in European ancestry samples) showed strong positive genetic correlations with anthropometric and obesity-related traits including birth length ($R_g = 0.81$, $P = 2.0 \times 10^{-44}$) and, in adults, height ($R_g = 0.41$, $P = 4.8 \times 10^{-52}$), waist circumference ($R_g = 0.18$, $P = 3.9 \times 10^{-10}$) and BMI ($R_g = 0.11$, $P = 7.3 \times 10^{-6}$). By contrast, BW showed inverse genetic correlations with indicators of adverse metabolic and cardiovascular health including coronary artery disease (CAD, $R_g = -0.30$, $P = 6.5 \times 10^{-9}$), systolic blood pressure (SBP, $R_g = -0.22$, $P = 5.5 \times 10^{-13}$) and T2D ($R_g = -0.27$, $P = 1.1 \times 10^{-6}$) (Fig. 1, Supplementary Table 12). The correlations between BW and adult cardiometabolic phenotypes are of similar magnitude, although directionally opposite, to the reported genetic correlations between adult BMI and those same cardiometabolic outcomes[8]. These findings support observational associations between a history of paternal T2D and lower BW (ref. 4), and establish more generally that the observed life-course associations between early growth and adult disease, at least in part, reflect the impact of shared genetic variants that influence both sets of phenotypes.

In an effort to estimate the extent of genetic contribution to these life-course associations, we first focused on data from the UK Biobank ($n = 57{,}715$). For many of the traits for which data were available, genetic variation contributed substantially to the life-course relationship between BW and adult phenotypes, and in some cases appeared to be the major source of covariance between the traits. For example, we estimated that 85% (95% CI = 70–99%) of the negative covariance between BW and SBP was explained by shared genetic associations captured by directly genotyped SNPs (Supplementary Table 13, Methods and Supplementary Fig. 2). For continuous cardiometabolic measures, including lipids and fasting glycaemia, for which measures are not currently available in the UK Biobank, we used data from the Northern Finland Birth Cohort ($n = 5{,}009$), and obtained similar results (Supplementary Table 13). However, these estimates were limited, not only by wide confidence intervals, but also by the assumption of a linear relationship between BW and each of the phenotypes and by the inability to explicitly model maternal genotypic effects. In other words, the inverse genetic correlations between BW and cardiometabolic traits may not exclusively reflect genetic effects mediated directly through the offspring, but also effects mediated by maternal genotype acting indirectly on the fetus via perturbation of the *in utero* environment. Nevertheless, these estimates indicate that a substantial proportion of the variance in cardiometabolic risk that correlates with BW can be attributed to the effects of common genetic variation.

To elucidate the biological pathways and processes underlying regulation of fetal growth, we first performed gene set enrichment analysis of our BW GWAS analysis using MAGENTA (Meta-Analysis Gene-set Enrichment of variaNT Associations, ref. 9) approach (Methods). Twelve pathways reached study-wide significance (false discovery rate, FDR < 0.05), including pathways involved in metabolism (insulin signalling, glycogen biosynthesis and cholesterol biosynthesis), growth (IGF signalling and growth hormone pathway) and development (chromatin remodelling) (Extended Data Table 2a). Similar pathways were detected in a complementary analysis in which we analysed empirical protein–protein interaction (PPI) data identifying

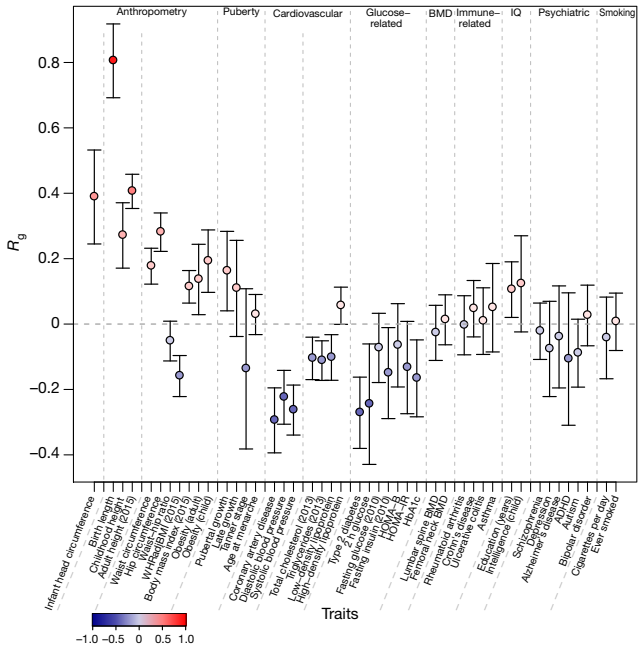


Figure 1 | Genome-wide genetic correlation between BW and a range of traits and diseases in later life. Genetic correlation (R_g) and corresponding s.e. (error bars) between BW and the traits displayed on the *x* axis were estimated using linkage-disequilibrium score regression (ref. 8). The genetic correlation estimates (R_g) are colour coded according to their intensity and direction (red for positive and blue for inverse correlation). WHRadjBMI, waist–hip ratio adjusted for body mass index; HOMA-B/IR, homeostasis model assessment of beta-cell function/insulin resistance; HbA1c, haemoglobin A1c; BMD, bone mineral density; ADHD, attention deficit hyperactivity disorder. See Supplementary Table 12 for references for each of the traits and diseases displayed.

13 PPI network modules with marked (Z score > 5) enrichment for BW-association scores (Extended Data Table 2b, Extended Data Fig. 6a, b and Methods). The proteins within these modules were themselves enriched for diverse processes related to metabolism, growth and development (Extended Data Fig. 6a, b).

We also observed enrichment of BW association signals across the set of 77 imprinted genes defined by the Genotype-Tissue Expression (GTEx) project (ref. 10) ($P = 1.9 \times 10^{-4}$; Extended Data Table 2a and Supplementary Table 14). Such enrichment is consistent with the 'parental conflict' hypothesis regarding the allocation of maternal resources to the fetus[11]. Although the role of imprinted genes in fetal growth has been described in animal models and rare human disorders[12], these data provide a large-scale, systematic indication of their contribution to normal variation in BW. Of the 60 genome-wide significant loci, two (*INS–IGF2* and *RB1*) fall within (or near) imprinted regions (Extended Data Fig. 2b), with a noteworthy third signal at *DLK1* (previously fetal antigen-1; $P = 5.6 \times 10^{-8}$). Parent-of-origin specific analyses to further investigate these individual loci (comparing heterozygote versus homozygote BW variance in 57,715 unrelated individuals, and testing BW associations with paternal versus maternal alleles in 4,908 mother–child pairs; see Methods) proved, despite these sample sizes, to be underpowered (Extended Data Fig. 7 and Supplementary Tables 15, 16).

Many of the genome-wide signals for BW detected here are also established genome-wide association signals for a wide variety of cardiometabolic traits (Fig. 2). These include the BW signals near *CDKAL1, ADCY5, HHEX–IDE* and *ANK1* (also genome-wide significant for T2D), *NT5C2* (for blood pressure, CAD and BMI) and *ADRB1* (for blood pressure). We used two approaches to understand whether this pattern of adult trait association represented a generic property of BW-associated loci or reflected heterogeneous mechanisms linking BW to adult disease.

First, we applied unsupervised hierarchical clustering (Methods) to the non-BW trait association statistics for the 60 significant BW loci. The resultant heat map showed the heterogeneity of locus-specific effect sizes across the range of adult traits (Fig. 2 and Supplementary Table 17). For example, it revealed that the associations between BW-raising alleles and increased adult height are concentrated amongst a subset of loci including *HHIP* and *GNA12*, and highlighted particularly strong associations with lipid traits for variants at the *TRIB1* and *MAFB* loci.

Second, we constructed trait-specific 'point-of-contact' (PoC) PPI networks from proteins represented in both the global BW PPI network and equivalent PPI networks generated for each of the adult traits (Methods and Extended Data Figs 6c–e). We reasoned that these PoC PPI networks would be enriched for the specific proteins mediating the observed links between BW and adult traits, generating hypotheses that are amenable to subsequent empirical validation. To highlight processes implicated in specific BW-trait associations, we overlaid these PoC PPI with the top 50 pathways that were over-represented in the global BW PPI network. These analyses revealed, for example, that proteins in the Wnt canonical signalling pathway were detected in the PoC PPI network only for blood pressure traits. We used these PPI overlaps to highlight the specific transcripts within BW GWAS loci that were likely to mediate the mechanistic links. For example, the overlap between the Wnt signalling pathway and the PoC PPI network for the intersection of BW and blood pressure-related traits implicated *FZD9* as the likely effector gene at the *MLXIPL* BW locus (Extended Data Fig. 6d and Supplementary Table 6).

We focused our more detailed investigation of the mechanistic links between early growth and adult traits on two phenotypic areas: arterial blood pressure and T2D/glycaemia. Across both the overall GWAS and specifically among the 60 significant BW loci, most BW-raising alleles were associated with reduced blood pressure (Figs 1, 2); the strongest inverse associations were seen for the loci near *NT5C2, FES, NRIP1, EBF1* and *PTH1R*. However, we also observed locus-specific heterogeneity in the genetic relationships between blood pressure and BW: the SBP-raising allele at *ADRB1*[13] is associated with higher, rather than lower, BW (Extended Data Fig. 8a). When we considered the reciprocal relationship, that is, the effects on BW of blood-pressure-raising alleles at 30 reported loci for SBP[13,14], there was an excess of associations (5 out of 30 with lower BW at $P < 0.05$; binomial $P = 0.0026$; Extended Data Fig. 8a). To dissect maternal and fetal genotype effects at these loci, we tested the impact on BW of a risk score generated from the 30 SBP SNPs, restricted to the untransmitted maternal haplotype score[15] in a set of 5,201 mother–child pairs. Analysis of these loci indicated that maternal genotype effects on the intrauterine environment probably

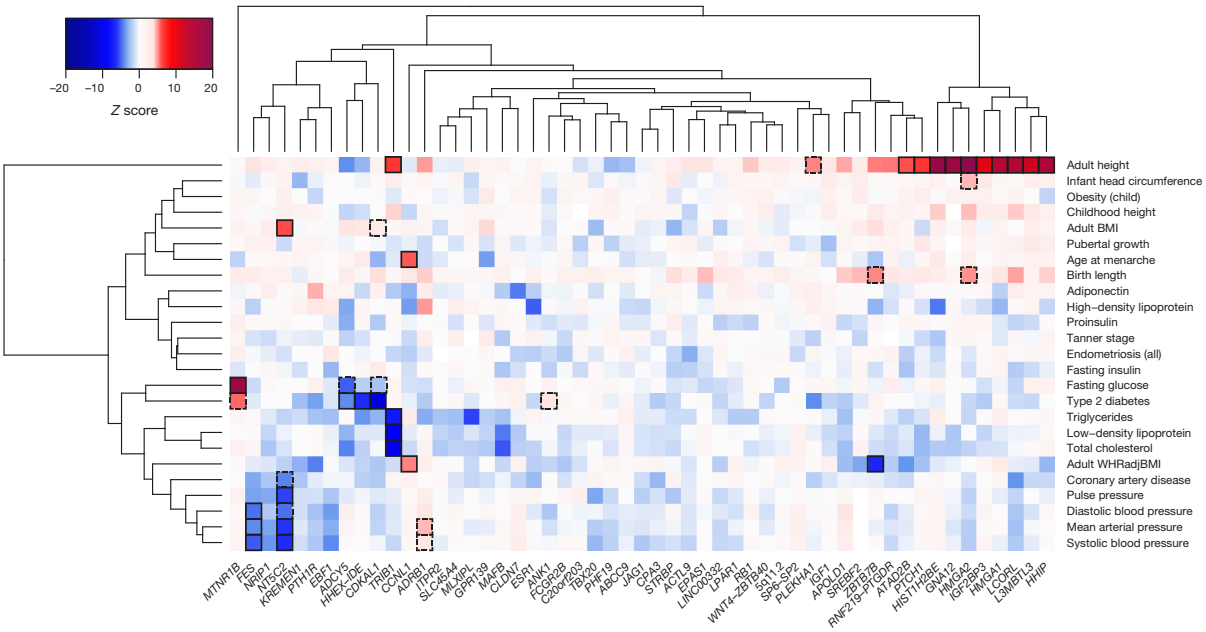


Figure 2 | Hierarchical clustering of BW loci based on similarity of overlap with adult diseases, metabolic and anthropometric traits. For the lead SNP at each BW locus (*x* axis), *Z* scores (aligned to BW-raising allele) were obtained from publicly available GWAS for various traits (*y* axis; see Supplementary Table 17). A positive *Z* score (red) indicates a positive association between the BW-raising allele and the outcome trait, while a negative *Z* score (blue) indicates an inverse association. BW loci and traits were clustered according to the Euclidean distance amongst *Z* scores (see Methods). Squares are outlined with a solid black line if the BW locus is significantly ($P < 5\times 10^{-8}$) associated with the trait in publicly available GWAS, or with a dashed line if reported significant elsewhere.

contribute to the inverse genetic correlation between SBP and BW (Methods and Supplementary Table 18), and was consistent with the results of a wider study of >30,000 women which demonstrated associations between a maternal genetic score for SBP (conditional on fetal genotype) and lower offspring BW[16].

The blood-pressure-raising allele with the largest BW-lowering effect mapped to the *NT5C2* locus (index variant for BW, rs74233809, $R^2 = 0.98$ with index variant for blood pressure, rs11191548; ref. 14) and was also associated with lower adult BMI ($R^2 = 0.99$ with rs11191560; ref. 17). The BW-lowering allele at rs74233809 is a proxy for a recently described[18] functional variant in the nearby *CYP17A1* gene ($R^2 = 0.92$ with rs138009835). The *CYP17A1* gene encodes the cytochrome P450c17α enzyme CYP17 (ref. 19), which catalyses key steps in steroidogenesis that determine the balance between mineralocorticoid, glucocorticoid and androgen synthesis. This variant has been shown to alter transcriptional efficiency *in vitro* and is associated with increased urinary tetrahydroaldosterone excretion[18]. *CYP17A1* is expressed in fetal adrenal glands and testes from early gestation[20] as well as in the placenta[21]. These data suggest that variation in *CYP17A1* expression contributes to the observational association between low BW and adult hypertension[22].

When we analysed 45 loci associated with CAD[23], the inverse genetic correlation between CAD and BW was concentrated amongst the five CAD loci with primary blood pressure associations. This suggests that genetic determinants of blood pressure play a leading role in mediating the life-course associations between BW and CAD (Extended Data Fig. 8b, e).

Linkage-disequilibrium score regression analyses demonstrated overall inverse genetic correlation between lower BW and elevated risk of T2D (Fig. 1). However, the locus-specific heat map indicates a heterogeneous pattern across individual loci (Fig. 2). To explore this further, we tested the 84 reported T2D loci[24] for association with BW. Some T2D risk alleles (such as those at *ADCY5*, *CDKAL1* and *HHEX–IDE*) were strongly associated with lower BW, while others (including *ANK1* and *MTNR1B*) were associated with higher BW (Extended Data Fig. 8c). This was in contrast with the BW effects of 422 known height loci[25] (Extended Data Fig. 8d), which showed a strong positive correlation consistent with the overall genetic correlation between height and BW, indicating that the growth effects of many height loci start prenatally and persist into adulthood.

The contrasting associations of T2D-risk alleles with both higher and lower BW probably reflect the differential impacts, across loci, of variation in the maternal and fetal genomes. Observational data link paternal diabetes with lower offspring BW[4], indicating that the inheritance of T2D risk alleles by the fetus tends, in line with the linkage-disequilibrium score regression analysis, to reduce growth. These relationships are consistent with the precepts of the 'fetal insulin hypothesis'[26] and reflect the potential for reduced insulin secretion and/or signalling to lead to both reduced fetal growth and, many decades later, enhanced predisposition to T2D. In line with this, the inferred paternal transmitted haplotype score generated from the 84 T2D risk variants was associated with lower BW ($P = 0.045$) in 5,201 mother–child pairs (Methods and Supplementary Table 18). In contrast, maternal diabetes is observationally associated with higher offspring BW[4], reflecting the ability of maternal hyperglycaemia to stimulate fetal insulin secretion. The contribution of genotype-dependent maternal hyperglycaemia to BW is in line with the evidence, from a recent study, that maternal genotype scores for fasting glucose and T2D (conditional on fetal genotype) were causally associated with higher offspring BW[16]. It is also consistent with the observation that a subset of glucose-raising alleles is associated with higher BW[7]. For example, the T2D-risk variant at *MTNR1B* (which also has a marked effect on fasting glucose levels in non-diabetic individuals[27,28]) was amongst the subset of BW loci (5 out of 60) for which the BW effect attributable to maternal genotype exceeded that associated with the fetal genotype (maternal: $\beta = 0.048$, $P = 5.1 \times 10^{-15}$; fetal: $\beta = 0.023$, $P = 2.9 \times 10^{-8}$) (Supplementary Table 10 and Extended Data Figs 4, 5). Thus, both maternal and fetal genetic effects connect BW to later T2D risk, albeit acting in opposing

directions. When we categorized T2D loci using a classification of physiological functions derived from their effects on related glycaemic and anthropometric traits[27], we found that T2D-risk alleles associated with lower BW were those typically characterized by reduced insulin processing and secretion without detectable changes in fasting glucose (the 'Beta Cell' cluster in Extended Data Fig. 8f).

The *YTK6* signal at rs138715366 is notable not only because the genetic data indicate that a single low-frequency non-coding variant is driving the association signal (see above) but also because of the proximity of this signal to *GCK*. Rare coding variants in glucokinase are causal for a form of monogenic hyperglycaemia and lead to large reductions in BW when parental alleles are passed on to their offspring[29]. In addition, common non-coding variants nearby are implicated in T2D risk and fasting hyperglycaemia[28]. However, the latter variants are conditionally independent of rs138715366 (Supplementary Table 19) and show no comparable association with lower BW. Either rs138715366 acts through effector transcripts other than *GCK*, or the impact of the low-frequency SNP near *YKT6* on *GCK* expression involves tissue- and/or temporal-specific variation in regulatory impact.

In conclusion, we have identified 60 genetic loci associated with BW and used them to gain insights into the aetiology of fetal growth and into well-established, but until now poorly understood, life-course disease associations. The evidence that the relationship between early growth and later metabolic disease has an appreciable genetic component contrasts with, but is not necessarily incompatible with, the emphasis on adverse early environmental events highlighted by the fetal origins hypothesis[1]. As we have shown, these genetic effects reflect variation in both the fetal and the maternal genome: the impact of the latter on the offspring's predisposition to adult disease could be mediated, at least in part, through perturbation of the antenatal and early life environment. Future mechanistic and genetic studies should support reconciliation between these alternative, but complementary, explanations for the far-reaching life-course associations that exist between events in early life and predisposition to cardiometabolic disease several decades later.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 4 February; accepted 2 September 2016.
Published online 28 September 2016.

1. Barker, D. J. The developmental origins of chronic adult disease. *Acta Paediatr. Suppl.* **93,** 26–33 (2004).
2. The 1000 Genomes Project Consortium An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
3. The UK10K Project Consortium The UK10K project identifies rare variants in health and disease. *Nature* **526,** 82–90 (2015).
4. Tyrrell, J. S., Yaghootkar, H., Freathy, R. M., Hattersley, A. T. & Frayling, T. M. Parental diabetes and birthweight in 236,030 individuals in the UK Biobank study. *Int. J. Epidemiol.* **42,** 1714–1723 (2013).
5. Horikoshi, M. *et al.* New loci associated with birth weight identify genetic links between intrauterine growth and adult height and metabolism. *Nat. Genet.* **45,** 76–82 (2013).
6. Eaves, L. J., Pourcain, B. S., Smith, G. D., York, T. P. & Evans, D. M. Resolving the effects of maternal and offspring genotype on dyadic outcomes in genome wide complex trait analysis ("M-GCTA"). *Behav. Genet.* **44,** 445–455 (2014).
7. Feenstra, B., *et al.* Maternal genome-wide association study identifies a fasting glucose variant associated with offspring birth weight. Preprint at: http://biorxiv.org/content/early/2015/12/11/034207 (2015).
8. Bulik-Sullivan, B. K. *et al.* LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47,** 291–295 (2015).
9. Segrè, A. V. *et al.* Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PLoS Genet.* **6,** e1001058 (2010).
10. Baran, Y. *et al.* The landscape of genomic imprinting across diverse adult human tissues. *Genome Res.* **25,** 927–936 (2015).
11. Haig, D. & Westoby, M. Parent-specific gene expression and the triploid endosperm. *Am. Nat.* **134,** 147–155 (1989).
12. Peters, J. The role of genomic imprinting in biology and disease: an expanding view. *Nat. Rev. Genet.* **15,** 517–530 (2014).
13. Johnson, T. *et al.* Blood pressure loci identified with a gene-centric array. *Am. J. Hum. Genet.* **89,** 688–700 (2011).
14. International Consortium for Blood Pressure Genome-Wide Association Studies *et al.* Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature* **478,** 103–109 (2011).
15. Zhang, G. *et al.* Assessing the causal relationship of maternal height on birth size and gestational age at birth: a Mendelian randomization analysis. *PLoS Med.* **12,** e1001865 (2015).
16. Tyrrell, J. *et al.* Genetic evidence for causal relationships between maternal obesity-related traits and birth weight. *J. Am. Med. Assoc.* **315,** 1129–1140 (2016).
17. Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518,** 197–206 (2015).
18. Diver, L. A. *et al.* Common polymorphisms at the *CYP17A1* locus associate with steroid phenotype: support for blood pressure genome-wide association study signals at this locus. *Hypertension* **67,** 724–732 (2016).
19. Picado-Leonard, J. & Miller, W. L. Cloning and sequence of the human gene for P450c17 (steroid 17 alpha-hydroxylase/17,20 lyase): similarity with the gene for P450c21. *DNA* **6,** 439–448 (1987).
20. Pezzi, V., Mathis, J. M., Rainey, W. E. & Carr, B. R. Profiling transcript levels for steroidogenic enzymes in fetal tissues. *J. Steroid Biochem. Mol. Biol.* **87,** 181–189 (2003).
21. Escobar, J. C., Patel, S. S., Beshay, V. E., Suzuki, T. & Carr, B. R. The human placenta expresses CYP17 and generates androgens *de novo*. *J. Clin. Endocrinol. Metab.* **96,** 1385–1392 (2011).
22. Reynolds, R. M. *et al.* Programming of hypertension: associations of plasma aldosterone in adult men and women with birthweight, cortisol, and blood pressure. *Hypertension* **53,** 932–936 (2009).
23. CARDIoGRAMplusC4D Consortium *et al.* Large-scale association analysis identifies new risk loci for coronary artery disease. *Nat Genet.* **45,** 25–33 (2013).
24. DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium *et al.* Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat. Genet.* **46,** 234–244 (2014).
25. Wood, A. R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **46,** 1173–1186 (2014).
26. Hattersley, A. T. & Tooke, J. E. The fetal insulin hypothesis: an alternative explanation of the association of low birth weight with diabetes and vascular disease. *Lancet* **353,** 1789–1792 (1999).
27. Dimas, A. S. *et al.* Impact of type 2 diabetes susceptibility variants on quantitative glycemic traits reveals mechanistic heterogeneity. *Diabetes* **63,** 2158–2171 (2014).
28. Morris, A. P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat. Genet.* **44,** 981–990 (2012).
29. Hattersley, A. T. *et al.* Mutations in the glucokinase gene of the fetus result in reduced birth weight. *Nat. Genet.* **19,** 268–270 (1998).

Supplementary Information is available in the online version of the paper.

Acknowledgements Full acknowledgements and supporting grant details can be found in the Supplementary Information.

Author Contributions Core analyses and writing: M.H., R.N.B., F.R.D., N.M.W., M.N.K., J.F.-T., N.R.v.Z., K.J.G., A.P.M., K.K.O., J.F.F., N.J.T., J.R.P., D.M.E., M.I.M., R.M.F. Statistical analysis in individual studies: M.H., R.N.B., F.R.D., N.M.W., M.N.K., B.F., N.G., J.P.B., D.P.S., R.L.-G., T.S.A., E.K., R.R., L.-P.L., D.L.C., Y.W., E.T., C.A.W., C.T.H., J.-J.H., N.V.-T., P.K.J., E.T.H.B., I.N., N.P., A.M., E.M.v.L., R.J., V.La., M.N., J.M.M., S.E.J., P.-R.L., K.S.R., M.A.T., J.T., A.R.W., H.Y., D.M.S., I.P., K.Pan., X.W., L.C., F.G., K.E.S., M.Mu., E.V.R.A., Z.K., S.B.-G., F.S., D.T., J.W., C.M.-G., N.R.R., E.Z., G.V.D., Y.-Y.T., H.N.K., A.P.M., J.F.F., N.J.T., J.R.P., D.M.E., R.M.F. GWAS look-up in unpublished datasets: K.T.Z., N.R., D.R.N., R.C.W.M., C.H.T.T., W.H.T., S.K.G., F.J.v.R. Sample collection and data generation in individual studies: F.R.D., M.N.K., B.F., N.G., J.P.B., D.P.S., R.L.-G., R.R., L.-P.L., J.-J.H., I.N., E.M.v.L., M.B., P.M.-V., A.J.B., L.P., P.K., M.A., S.M.W., F.G., C.E.v.B., G.W., E.V.R.A., C.E.F., C.T., C.M.T., M.Sta., Z.K., D.M.H., M.V.H., H.G.d.H., F.R.R., C.M.-G., S.M.R., G.H., G.M., N.R.R., C.J.G., C.L., J.L., R.A.S., J.H.Z., F.D.M., W.L.L.Jr, A.T., M.Stu., V.Li., T.A.L., C.M.v.D., A.K., T.I.S., H.N., K.Pah., O.T.R., E.Z., G.V.D., S.-M.S., M.Me., H.C., J.F.W., M.V., J.-C.H., T.H., S.S., L.J.B., J.P.N., C.E.P., L.S.A., J.B.B., K.L.M., J.G.E., E.E.W., M.K., J.S.V., T.L., P.V., K.B., H.B., D.O.M.-K., F.R., A.G.U., C.Pi., O.P., N.J.W., H.H., V.W.J., S.F.G., A.A.V., D.A.L., G.D.S., K.K.O., J.F.F., N.J.T., J.R.P., M.I.M. Functional follow-up experiment: L.A.D., S.M.M., R.M.R., E.D., B.R.W. Individual study design and principal investigators: J.P.B., I.N., M.A., F.D.M., W.L.L.Jr, A.T., M.Stu., V.Li., T.A.L., C.M.v.D., W.K., A.K., T.I.S., H.N., K.Pah., O.T.R., G.V.D., Y.-Y.T., S.-M.S., M.Me., H.C., J.F.W., M.V., E.J.d.G., D.I.B., H.N.K., J.-C.H., T.H., A.T.H., L.J.B., J.P.N., C.E.P., J.H., L.S.A., J.B.B., K.L.M., J.G.E., E.E.W., M.K., J.S.V., T.L., P.V., K.B., H.B., D.O.M-K., A.H., F.R., A.G.U., C.Pi., O.P., C.Po., E.H., N.J.W., H.H., V.W.J., M.-R.J., S.F.G., A.A.V., T.M.F., A.P.M., K.K.O., N.J.T., J.R.P., M.I.M., R.M.F.

Author Information Summary statistics from the meta-analyses are available at http://egg-consortium.org/. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.I.M (mark.mccarthy@drl.ox.ac.uk) or R.M.F (r.freathy@ex.ac.uk).

Reviewer Information *Nature* thanks J. Whitfield and the other anonymous reviewer(s) for their contribution to the peer review of this work.

[1]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK. [2]Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Oxford OX3 7LE, UK. [3]Institute of Biomedical and Clinical Science, University of Exeter Medical School, University of Exeter, Royal Devon and Exeter Hospital, Exeter EX2 5DW, UK. [4]MRC Epidemiology Unit, University of Cambridge School of Clinical Medicine, Cambridge CB2 0QQ, UK. [5]The University of Queensland Diamantina Institute, Translational Research Institute, Brisbane, Queensland 4102, Australia. [6]School of Women's and Infants' Health, The University of Western Australia, Perth, Western Australia 6009, Australia. [7]The Generation R Study Group, Erasmus MC, University Medical Center Rotterdam, Rotterdam 3015 CE, the Netherlands. [8]Department of Epidemiology, Erasmus MC, University Medical Center Rotterdam, Rotterdam 3015 CE, the Netherlands. [9]Department of Pediatrics, Erasmus MC, University Medical Center Rotterdam, Rotterdam 3015 CE, the Netherlands. [10]Department of Epidemiology Research, Statens Serum Institut, Copenhagen DK-2300, Denmark. [11]Department of Pediatrics, University of California San Diego, La Jolla, San Diego, California 92093, USA. [12]The Novo Nordisk Foundation Center for Basic Metabolic Research, Section of Metabolic Genetics, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen DK-2100, Denmark. [13]Center for Applied Genomics, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania 19104, USA. [14]Population Health Research Institute, St George's University of London, Cranmer Terrace, London SW17 0RE, UK. [15]Department of Clinical Epidemiology, Leiden University Medical Center, Leiden 2333 ZA, the Netherlands. [16]COPSAC, Copenhagen Prospective Studies on Asthma in Childhood, Herlev and Gentofte Hospital, University of Copenhagen, Copenhagen, 2820 Gentofte, Denmark. [17]Steno Diabetes Center, Gentofte DK-2820, Denmark. [18]Department of Computational Biology, University of Lausanne, Lausanne 1011, Switzerland. [19]Swiss Institute of Bioinformatics, Lausanne 1015, Switzerland. [20]Department of Clinical Chemistry, Fimlab Laboratories, Tampere 33520, Finland. [21]Department of Clinical Chemistry, University of Tampere School of Medicine, Tampere 33014, Finland. [22]Institute for Molecular Medicine, Finland (FIMM), University of Helsinki, Helsinki FI-00100, Finland. [23]Division of Human Genetics, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania 19104, USA. [24]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. [25]Department of Genetics, University of North Carolina, Chapel Hill, North Carolina 27599, USA. [26]Institute of Epidemiology I, Helmholtz Zentrum München- German Research Center for Environmental Health, 85764 Neuherberg, Germany. [27]Division of Metabolic and Nutritional Medicine, Dr. von Hauner Children's Hospital, University of Munich Medical Center, 80337 Munich, Germany. [28]Netherlands Twin Register, Department of Biological Psychology, Vrije Universiteit, Amsterdam 1081 BT, the Netherlands. [29]ISGlobal, Centre for Research in Environmental Epidemiology (CREAL), Barcelona 08003, Spain. [30]Universitat Pompeu Fabra (UPF), Barcelona 08002, Spain. [31]CIBER de Epidemiología y Salud Pública (CIBERESP), Madrid 28029, Spain. [32]Usher Institute for Population Health Sciences and Informatics, University of Edinburgh, Edinburgh EH8 9AG, UK. [33]Saw Swee Hock School of Public Health, National University of Singapore, National University Health System, Singapore 119077, Singapore. [34]William Harvey Research Institute, Barts and the London School of Medicine and Dentistry, Queen Mary University of London, London EC1M 6BQ, UK. [35]Department of Nutrition and Dietetics, School of Health Science and Education, Harokopio University, Athens 17671, Greece. [36]Research Centre of Applied and Preventive Cardiovascular Medicine, University of Turku, Turku 20014, Finland. [37]Institute of Biomedicine, Physiology, University of Eastern Finland, Kuopio FI-70211, Finland. [38]KUL – University of Leuven, Department of Neurosciences, Leuven 3000, Belgium. [39]Translational Immunology Laboratory, VIB, Leuven 3000, Belgium. [40]Department of Preventive Medicine, Division of Biostatistics, Feinberg School of Medicine, Northwestern University, Chicago, Illinois 60611, USA. [41]Institute of Cardiovascular & Medical Sciences, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow G12 8TA, UK. [42]Endometriosis CaRe Centre, Nuffield Department of Obstetrics & Gynaecology, University of Oxford, Oxford OX3 9DU, UK. [43]Center for Genomic Regulation (CRG), Barcelona 08003, Spain. [44]Department of Internal Medicine, Internal Medicine, Lausanne University Hospital (CHUV), Lausanne 1011, Switzerland. [45]Joint BSC-CRG-IRB Research Program in Computational Biology, Barcelona Supercomputing Center, Barcelona 08034, Spain. [46]Institute of Health and Biomedical Innovation, Queensland University of Technology, Brisbane, Queensland 4000, Australia. [47]Department of Medicine and Therapeutics, The Chinese University of Hong Kong, Hong Kong, China. [48]Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Hong Kong, China. [49]Hong Kong Institute of Diabetes and Obesity, The Chinese University of Hong Kong, Hong Kong, China. [50]Department of Obstetrics and Gynaecology, The Chinese University of Hong Kong, Hong Kong, China. [51]Department of Human Genetics and Cardiovascular Medicine, Department of Internal Medicine, University of Michigan, Ann Arbor, Michigan 48109, USA. [52]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts 02115, USA. [53]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA. [54]European Centre for Environment and Human Health, University of Exeter, Truro TR1 3HD, UK. [55]Medical Research Council Integrative Epidemiology Unit at the University of Bristol, Bristol BS8 2BN, UK. [56]School of Social and Community Medicine, University of Bristol, Bristol BS8 2BN, UK. [57]Department of Genomics of Common Disease, School of Public Health, Imperial College London, London SW7 2AZ, UK. [58]IFB Adiposity Diseases, University of Leipzig, 04103 Leipzig, Germany. [59]Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1HH, UK. [60]FISABIO–Universitat Jaume I–Universitat de València, Joint Research Unit of Epidemiology and Environmental Health, Valencia 46020, Spain. [61]The Children's Obesity Clinic, Department of Pediatrics, Copenhagen University Hospital Holbæk, Holbæk DK-4300, Denmark. [62]Institute of Social and Preventive Medicine, Lausanne University Hospital (CHUV), Lausanne 1010, Switzerland. [63]Danish Center for Neonatal Screening, Statens Serum Institute, Copenhagen DK-2300, Denmark. [64]Department for Congenital Disorders, Statens Serum Institute, Copenhagen DK-2300, Denmark. [65]Computer Sciences Department, Barcelona Supercomputing Center, Barcelona 08034, Spain. [66]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona 08010, Spain. [67]Department of Internal Medicine, Erasmus MC, University Medical Center Rotterdam, Rotterdam 3015 CE, the Netherlands. [68]BHF Centre for Cardiovascular Science, University of Edinburgh, Queen's Medical Research Institute, Edinburgh EH16 4TJ, UK. [69]Department of Medicine, Division of Endocrinology, Metabolism, and Molecular Medicine, Feinberg School of Medicine, Northwestern University, Chicago, Illinois 60611, USA. [70]Medical Department, University of Leipzig, 04103 Leipzig, Germany. [71]Department of Clinical Physiology and Nuclear Medicine, Kuopio University Hospital, Kuopio FI-70029, Finland. [72]Kuopio Research Institute of Exercise Medicine, Kuopio FI-70100, Finland. [73]Pediatric Research Center, Department of Women´s & Child Health, University of Leipzig, 04103 Leipzig, Germany. [74]Novo Nordisk Foundation Center for Basic Metabolic Research and Department of Public Health, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen DK-2200, Denmark. [75]Institute of Preventive Medicine, Bispebjerg and Frederiksberg Hospital, The Capital Region, Copenhagen DK-2000, Denmark. [76]Department of Pediatrics, Turku University Hospital, Turku 20521, Finland. [77]Department of Physiology, University of Turku, Turku 20014, Finland. [78]Paavo Nurmi Centre, Sports and Exercise Medicine Unit, Department of Physical Activity and Health, Turku 20014, Finland. [79]Department of Clinical Physiology and Nuclear Medicine, Turku University Hospital, Turku 20521, Finland. [80]Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546, Singapore. [81]Life Sciences Institute, National University of Singapore, Singapore 117456, Singapore. [82]Singapore Eye Research Institute, Singapore 168751, Singapore. [83]Department of Clinical Medicine, University of Copenhagen, Copenhagen DK-2200, Denmark. [84]Department of Medicine, Stanford School of Medicine, Stanford, California 94305, USA. [85]MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh EH4 2XU, UK. [86]EMGO Institute for Health and Care Research, VU University and VU University Medical Center, Amsterdam 1081 HV, the Netherlands. [87]Department of Large Animal Sciences, Faculty of Health and Medical Sciences, University of Copenhagen, Frederiksberg C DK-1870, Denmark. [88]Center for Life Course Health Research, Faculty of Medicine, University of Oulu, Oulu 90014, Finland. [89]Biocenter Oulu, University of Oulu, Oulu 90014, Finland. [90]School of Medicine and Pharmacology, Royal Perth Hospital Unit, The University of Western Australia, Perth, Western Australia 6000, Australia. [91]Institute and Outpatient Clinic for Occupational, Social and Environmental Medicine, Inner City Clinic, University Hospital Munich, Ludwig Maximilian University of Munich, 80336 Munich, Germany. [92]Department of Nutrition, University of North Carolina, Chapel Hill, North Carolina 27599, USA. [93]USC-Office of Population Studies Foundation, Inc., University of San Carlos, Cebu City 6000, Philippines. [94]Department of Nutrition and Dietetics, University of San Carlos, Cebu City 6000, Philippines. [95]National Institute for Health and Welfare, Helsinki 00271, Finland. [96]Department of General Practice and Primary Health Care, University of Helsinki and Helsinki University Hospital, Helsinki 00014, Finland. [97]Folkhälsan Research Center, Helsinki 00250, Finland. [98]Department of Clinical Physiology, Tampere University Hopital, Tampere 33521, Finland. [99]Department of Clinical Physiology, University of Tampere School of Medicine, Tampere 33014, Finland. [100]Division of Medicine, Turku University Hospital, Turku 20521, Finland. [101]Department of Medicine, University of Turku, Turku 20014, Finland. [102]Department of Public Health and Primary Care, Leiden University Medical Center, Leiden 2333 ZA, the Netherlands. [103]3Epidemiology Section, BESC Department, King Faisal Specialist Hospital and Research Centre, Riyadh 12713, Saudi Arabia. [104]Research Center for Prevention and Health Capital Region, Center for Sundhed, Rigshospitalet – Glostrup, Copenhagen University, Glostrup DK-2600, Denmark. [105]Population, Policy and Practice, UCL Institute of Child Health, University College London, London WC1N 1EH, UK. [106]Centre for Population Health Research, School of Health Sciences, and Sansom Institute, University of South Australia, Adelaide, South Australia 5001, Australia. [107]South Australian Health and Medical Research Institute, Adelaide, South Australia 5000, Australia. [108]Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. [109]Department of Epidemiology and Biostatistics, MRC–PHE Centre for Environment & Health, School of Public Health, Imperial College London, London SW7 2AZ, UK. [110]Unit of Primary Care, Oulu University Hospital, Oulu 90220, Finland. [111]Division of Endocrinology, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania 19104, USA. [112]Department of Endocrinology, Rigshospitalet, Copenhagen DK-2100, Denmark. [113]AstraZeneca, Innovative Medicines and Early Development | Early Clinical Development, Mölndal 431 83, Sweden. [114]Department of Biostatistics, University of Liverpool, Liverpool L69 3GA, UK. [115]Estonian Genome Center, University of Tartu, Tartu 50090, Estonia. [116]Department of Paediatrics, University of Cambridge, Cambridge CB2 0QQ, UK. [117]Oxford National Institute for Health Research (NIHR) Biomedical Research Centre, Churchill Hospital, Oxford OX3 7LE, UK.

*These authors contributed equally to this work.

†A list of consortium members appears in the Supplementary Information.

‡Deceased.

§These authors jointly supervised this work.

METHODS

Ethics statement. All human research was approved by the relevant institutional review boards and conducted according to the Declaration of Helsinki. All participants provided written informed consent. Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee and the local Research Ethics Committees.

Study-level analyses. No statistical methods were used to predetermine sample size: to maximise power to detect association signals, we set out to collect the largest possible set of samples for which the combination of genome-wide genotyping data and reliable measures of BW could be made available for analysis. Within each study, BW was collected from a variety of sources, including measurements at birth by medical practitioners, obstetric records, medical registers, interviews with the mother and self-report as adults (Supplementary Table 1). BW was *Z*-score transformed separately in males and females. Individuals with extreme BW (>5 s.d. from the sex-specific study mean), monozygotic or polyzygotic siblings, or preterm births (gestational age <37 weeks, where this information was available) were excluded from downstream association analyses (Supplementary Table 1).

Within each study, stringent quality control of the GWAS genotype scaffold was carried out before imputation (Supplementary Table 2). Each scaffold was then pre-phased and imputed[30,31] up to reference panels from the 1000G project[2] or the combined 1000G and UK10K projects[3] (Supplementary Table 2). Association of BW with each variant passing established GWAS quality control filters[32] was tested in a linear regression framework, under an additive model for the allelic effect, after adjustment for study-specific covariates, including gestational age, where available (Supplementary Table 2). Where necessary, population structure was accounted for by adjustment for axes of genetic variation from principal components analysis[33] and subsequent genomic control correction[34], or inclusion of a genetic relationship matrix in a mixed model[35] (Supplementary Table 2). We calculated the genomic control inflation factor (λ) in each study to confirm that study-level population structure was accounted for before meta-analysis.

Preparation, quality control and genetic analysis in UK Biobank samples. UK Biobank phenotype data were available for 502,655 participants[36]. All participants in the UK Biobank were asked to recall their BW, of which 279,971 did so at either the baseline or follow-up assessment visit. Of these, 7,686 participants reported being part of multiple births and were excluded from downstream analyses. Ancestry checks, based on self-reported ancestry, resulted in the exclusion of 8,998 additional participants reported not to be white European. Of those individuals reporting BW at baseline and follow-up assessments, 393 were excluded because the two reported values differed by more than 0.5 kg. For those reporting different values (≤0.5 kg) between baseline and follow-up, we took the baseline measure forward for downstream analyses. We then excluded 36,716 individuals reporting values <2.5 kg or >4.5 kg as implausible for live term births before 1970. In total 226,178 participants had data relating to BW that matched these inclusion criteria.

Genotype data from the May 2015 release were available for a subset of 152,249 participants from UK Biobank. In addition to the quality control metrics performed centrally by UK Biobank, we defined a subset of 'white European' ancestry samples using a *K*-means ($K=4$) clustering approach based on the first four genetically determined principal components. A maximum of 67,786 individuals (40,425 females and 27,361 males) with genotype and valid BW measures were available for downstream analyses. We tested for association with BW, assuming an additive allelic effect, in a linear mixed model implemented in BOLT-LMM (ref. 37) to account for cryptic population structure and relatedness. Genotyping array was included as a binary covariate in all models. Total chip heritability (that is, the variance explained by all autosomal polymorphic genotyped SNPs passing quality control) was calculated using restricted maximum likelihood (REML) implemented in BOLT-LMM (ref. 37). We additionally analysed the association between BW and directly genotyped SNPs on the X chromosome: for this analysis, we used 57,715 unrelated individuals with BW available and identified by UK Biobank as white British. We excluded SNPs with evidence of deviation from Hardy–Weinberg equilibrium ($P<1\times10^{-6}$), MAF < 0.01 or overall missing rate >0.015, resulting in 19,423 SNPs for analysis in Plink v1.07 (http://pngu.mgh.harvard.edu/purcell/plink/)[38], with the first five ancestry principal components as covariates.

In both the full UK Biobank sample and our refined sample, we observed that BW was associated with sex, year of birth and maternal smoking ($P<0.0015$, all in the expected directions), confirming more comprehensive previous validation of self-reported BW[4]. We additionally verified that BW associations with lead SNPs at seven established loci[5] based on self-report in UK Biobank were consistent with those previously published.

European ancestry meta-analysis. The European ancestry meta-analysis consisted of two components: (i) 75,891 individuals from 30 GWAS from Europe, USA and Australia; and (ii) 67,786 individuals of white European origin from the UK Biobank. In the first component, we combined sex-specific BW association summary statistics across studies in a fixed-effects meta-analysis, implemented in GWAMA (ref. 39) and applied a second round of genomic control[34] ($\lambda_{GC}=1.001$). Subsequently, we combined association summary statistics from this component with the UK Biobank in a European ancestry fixed-effects meta-analysis, implemented in GWAMA (ref. 39). Variants failing GWAS quality control filters in the UK Biobank, reported in less than 50% of the total sample size in the first component, or with MAF <0.1%, were excluded from the European ancestry meta-analysis. We aggregated X-chromosome association summary statistics from the UK Biobank (19,423 SNPs) with corresponding statistics from the European GWAS component using fixed effects *P*-value-based meta-analysis in METAL (ref. 40) (max $n=99{,}152$).

We were concerned that self-reported BW as adults in the UK Biobank would not be comparable with that obtained from more stringent collection methods used in other European ancestry GWAS. In addition, the UK Biobank lacked information on gestational age for adjustment, which could have an impact on strength of association compared with the results obtained from other European ancestry GWAS. However, we observed no evidence of heterogeneity in BW allelic effects at lead SNPs between the two components of European ancestry meta-analysis, using Cochran's *Q* statistic[41] implemented in GWAMA (ref. 39) after Bonferroni correction ($P>0.00083$) (Supplementary Table 3). We tested for heterogeneity in allelic effects between studies within the European component using Cochran's *Q*. At loci demonstrating evidence of heterogeneity, we confirmed that association signals were not driven by outlying studies by visual inspection of forest plots. We performed sensitivity analyses to assess the impact of covariate adjustment (gestational age and population structure) on heterogeneity.

We were also concerned that overlap of individuals (duplicated or related) between the two components of the European ancestry meta-analysis might lead to false positive association signals. We performed bivariate linkage-disequilibrium score regression[8] using the two components of the European ancestry meta-analysis and observed a genetic covariance intercept of 0.0156 ± 0.0058 (s.e.), indicating a maximum of 1,119 duplicate individuals. Univariate linkage-disequilibrium score regression[8] of the European ancestry meta-analysis estimated the intercept as 1.0426, which may indicate population structure or relatedness that was not adequately accounted for in the analysis. To assess the impact of this inflation on the European ancestry meta-analysis, we expanded the standard errors of BW allelic effect size estimates and re-calculated association *P* values. On the basis of this adjusted analysis, only the lead SNP at *MTNR1B* dropped below genome-wide significance (rs10830963, $P=5.5\times10^{-8}$).

Trans-ancestry meta-analysis. The trans-ancestry meta-analysis combined the two European ancestry components with an additional 10,104 individuals from six GWAS from diverse ancestry groups: African American, Chinese, Filipino, Surinamese, Turkish and Moroccan. Within each GWAS, we first combined sex-specific BW association summary statistics in a fixed-effects meta-analysis, implemented in GWAMA (ref. 39) and applied a second round of genomic control[34]. Subsequently, we combined association summary statistics from the six non-European GWAS and the two European ancestry components in a trans-ancestry fixed-effects meta-analysis, implemented in GWAMA (ref. 39). Variants failing GWAS quality control filters in the UK Biobank, reported in less than 50% of the total sample size in the first component, or with MAF <0.1%, were excluded from the trans-ancestry meta-analysis. We tested for heterogeneity in allelic effects between ancestries using Cochran's *Q* (ref. 41).

Approximate conditional analysis. We searched for multiple distinct BW association signals in each of the established and novel loci, defined as 1 Mb up- and down-stream of the lead SNP from the trans-ancestry meta-analysis, through approximate conditional analysis. We applied GCTA (ref. 42) to identify 'index SNPs' for distinct association signals attaining genome-wide significance ($P<5\times10^{-8}$) in the European ancestry meta-analysis using a reference sample of 5,000 individuals of white British origin, randomly selected from the UK Biobank, to approximate patterns of linkage disequilibrium between variants in these regions. Note that we performed approximate conditioning on the basis of only the European ancestry meta-analysis because GCTA cannot accommodate linkage-disequilibrium variation between diverse populations.

Prioritizing candidate genes in each BW locus. We combined a number of approaches to prioritize the most likely candidate gene(s) in each BW locus. Expression quantitative trait loci (eQTLs) were obtained from the Genotype Tissue Expression (GTEx) Project[43], the GEUVADIS project[44] and eleven other studies[45–55] using HaploReg v4 (ref. 56). We interrogated coding variants for each BW lead SNP and its proxies (EUR $R^2>0.8$) using Ensembl[57] and HaploReg. Their likely functional consequences were predicted by SIFT (ref. 58) and PolyPhen2 (ref. 59). Biological candidacy was assessed by presence in significantly enriched gene set pathways from MAGENTA analyses (see below for details). We extracted all genes within 300 kb of all lead BW SNPs and searched for connectivity between

any genes using STRING (ref. 60). If two or more genes between two separate BW loci were connected, they were given an increased prior for both being plausible candidates. We also applied protein–protein interaction (PPI) analysis (see below for details) to all genes within 300 kb of each lead BW SNPs and ranked the genes based on the score for connectivity with the surrounding genes.

Evaluation of imputation quality of the low-frequency variant at the *YKT6–GCK* locus. At the *YKT6–GCK* locus, the lead SNP (rs138715366) was found at a low frequency in European ancestry populations (MAF = 0.92%) and was even rarer in other ancestry groups (MAF = 0.23% in African Americans, otherwise monomorphic) and was not present in the HapMap reference panel[61]. To assess the accuracy of imputation for this low-frequency variant, we genotyped rs138715366 in the Northern Finland Birth Cohort (NFBC) 1966 (Supplementary Table 1). Of the 5,009 samples in the study, 4,704 were successfully imputed and genotyped (or sequenced) for rs138715366. The overall concordance rate between imputed and directly assayed genotypes was 99.8% and for directly assayed heterozygote calls was 75.0%.

Fine-mapping analyses. We investigated linkage-disequilibrium differences between populations contributing to the trans-ancestry meta-analysis and to take advantage of the improved coverage of common and low-frequency variation offered by 1000G or 1000G and UK10K combined imputation to localize variants driving each distinct association signal achieving locus-wide significance. For each distinct signal, we used MANTRA (ref. 62) to construct 99% credible sets of variants[63] that together account for 99% of the posterior probability of driving the association. MANTRA incorporates a prior model of relatedness between studies, based on mean pair-wise allele frequency differences across loci, to account for heterogeneity in allelic effects (Supplementary Table 3). MANTRA has been demonstrated, by simulation, to improve localization of causal variants compared with either a fixed- or random-effects trans-ancestry meta-analysis[62,64].

For loci with only one signal of association, we used MANTRA to combine summary statistics from the six non-European GWAS and the two European ancestry components. However, for loci with multiple distinct association signals, we used MANTRA to combine summary statistics from approximate conditioning for the two European components, separately for each signal.

For each distinct signal, we calculated the posterior probability that the *j*th variant, π_{Cj}, is driving the association, given by

$$\pi_{Cj} = \frac{\Lambda_j}{\sum_k \Lambda_k}$$

where the summation is over all variants mapping within the (conditional) meta-analysis across the locus. In this expression, Λ_j is the Bayes' factor in favour of association from the MANTRA analysis. A 99% credible set[63] was then constructed by: (i) ranking all variants according to their Bayes' factor, Λ_j; and (ii) including ranked variants until their cumulative posterior probability exceeds 0.99.

Genomic annotation. We used genomic annotations of DNaseI hypersensitive sites (DHS) from the ENCODE (ref. 65 project and protein coding genes from GENCODE (ref. 66). We filtered cell types that are cancer cell lines (karyotype 'cancer' from https://genome.ucsc.edu/ENCODE/cellTypes.html), and merged data from multiple samples from the same cell type. This resulted in 128 DHS cell-type annotations, as well as 4 gene-based annotations (coding exon, 5′UTR, 3′UTR and 1 kb upstream of the transcription start site (TSS)). First, we tested for the effect of each cell type DHS and gene annotation individually using the Bayes' factors for all variants in the 62 credible sets using fgwas (ref. 67). Second, we categorized the annotations into 'genic', 'fetal DHS', 'embryonic DHS', 'stem cell DHS', 'neonatal DHS' and 'adult DHS' based on the description fields from ENCODE, and tested for the effect of each category individually as described above using fgwas. Third, we then tested the effect of each category by including all categories in a joint model using fgwas. For each of the three analyses, we obtained the estimated effects and 95% confidence intervals (CI) for each annotation, and considered an annotation enriched if the 95% CI did not overlap zero.

Estimation of genetic variance explained. The 'variance explained' statistic was calculated using the REML method implemented in GCTA (ref. 68). We considered the variance explained by two sets of SNPs: (i) lead SNPs of all 62 distinct association signals at the 59 established and novel autosomal BW loci identified in the European-specific or trans-ancestry meta-analyses; (ii) lead SNPs of 55 distinct association signals at the 52 novel autosomal BW loci (Extended Data Table 1a and Supplementary Table 7). The 'variance explained' was calculated in samples of European ancestry in the Hyperglycemia and Adverse Pregnancy Outcome (HAPO) study[69] (independent of the meta-analysis) and two studies that were part of the European ancestry meta-analysis: NFBC1966 and Generation R (Supplementary Table 1). In each study, the genetic relationship matrix was estimated for each set of SNPs and was tested individually against BW (males and females combined) with study specific covariates. These analyses provided an estimate and s.e. for the variance explained by each of the given sets of SNPs.

Examining the relative effects on BW of maternal and fetal genotype at the 60 identified loci. We performed four sets of analyses. First, we used GWAS data from 4,382 mother–child pairs in the Avon Longitudinal Study of Parents and Children (ALSPAC) study to fit a 'maternal-GCTA model'[6] to estimate the extent to which the maternal genome might influence offspring BW independent of the fetal genome. The maternal-GCTA model uses genome-wide genetic similarity between mothers and offspring to partition the phenotypic variance in BW into components due to the maternal genotype, the child's genotype, the covariance between the two and environmental sources of variation.

Second, we compared associations with BW of the fetal versus maternal genotype at each of the 60 BW loci. The maternal allelic effect on offspring BW was obtained from a maternal GWAS meta-analysis of 68,254 European mothers from the EGG Consortium ($n = 19{,}626$)[7] and the UK Biobank ($n = 48{,}628$). In the UK Biobank, mothers were asked to report the BW of their first child. Women of European ancestry with genotype data available in the May 2015 data release were included, and those with reported BW equivalent to <2.5 kg or >4.5 kg were excluded. No information on gestational age or gender of child was available. BW of first child was associated with maternal factors such as smoking status, BMI and height in the expected directions. Of the 68,254 women included in the maternal GWAS, 13% were mothers of individuals included in the current fetal European ancestry GWAS, and a further ~45% were themselves (with their own BW) included in the fetal GWAS.

Third, we additionally conducted analyses in 12,909 mother–child pairs from nine contributing studies: at each of the 60 loci, we compared the effect of the fetal genotype on BW adjusted for sex and gestational age, with and without adjustment for maternal genotype. We reciprocally compared the association between the maternal genotype and BW with and without adjustment for fetal genotype.

Fourth, we used the method of Zhang *et al.*[15] to test associations between BW and the maternal untransmitted, maternal transmitted and inferred paternal transmitted haplotype score of 422 height SNPs[25], 30 SBP SNPs[13,14] and 84 T2D SNPs[24] in 5,201 mother–child pairs from the ALSPAC study.

Linkage-disequilibrium score regression. The use of linkage-disequilibrium score regression to estimate the genetic correlation between two traits/diseases has been described in detail elsewhere[70]. Briefly, the linkage-disequilibrium score is a measure of how much genetic variation each variant tags; if a variant has a high linkage-disequilibrium score then it is in high linkage disequilibrium with many nearby polymorphisms. Variants with high linkage-disequilibrium scores are more likely to contain more true signals and hence provide more chance of overlap with genuine signals between GWAS. The linkage-disequilibrium score regression method uses summary statistics from the GWAS meta-analysis of BW and the other traits of interest, calculates the cross-product of test statistics at each SNP, and then regresses the cross-product on the linkage-disequilibrium score. Bulik-Sullivan *et al.*[70] show that the slope of the regression is a function of the genetic covariance between traits:

$$E(z_{1j}z_{2j}) = \frac{\sqrt{N_1N_2}\,\rho_g}{M} l_j + \frac{\rho N_s}{\sqrt{N_1N_2}}$$

where N_i is the sample size for study i, ρ_g is the genetic covariance, M is the number of SNPs in the reference panel with MAF between 5% and 50%, l_j is the linkage-disequilibrium score for SNP j, N_s quantifies the number of individuals that overlap both studies, and ρ is the phenotypic correlation amongst the N_s overlapping samples. Thus, if there is sample overlap (or cryptic relatedness between samples), it will only affect the intercept from the regression (that is, the term $\frac{\rho N_s}{\sqrt{N_1N_2}}$) and not the slope, and hence estimates of the genetic covariance will not be biased by sample overlap. Likewise, population stratification will affect the intercept but will have minimal impact on the slope (that is, intuitively since population stratification does not correlate with linkage disequilibrium between nearby markers).

Summary statistics from the GWAS meta-analysis for traits and diseases of interest were downloaded from the relevant consortium website. The summary statistics files were reformatted for linkage-disequilibrium score regression analysis using the munge_sumstats.py python script provided on the developer's website (https://github.com/bulik/ldsc). For each trait, we filtered the summary statistics to the subset of HapMap 3 SNPs[71], as advised by the developers, to ensure that no bias was introduced due to poor imputation quality. Summary statistics from the European-specific BW meta-analysis were used because of the variable linkage-disequilibrium structure between ancestry groups. Where the sample size for each SNP was included in the results file this was flagged using N-col; if no sample size was available then the maximum sample size reported in the reference for the GWAS meta-analysis was used. SNPs were excluded for the

following reasons: MAF < 0.01; ambiguous strand; duplicate rsID; non-autosomal SNPs; reported sample size less than 60% of the total available. Once all files were reformatted, we used the ldsc.py python script, also on the developers' website, to calculate the genetic correlation between BW and each of the traits and diseases. The European linkage-disequilibrium score files calculated from the 1000G reference panel and provided by the developers were used for the analysis. Where multiple GWAS meta-analyses had been conducted on the same phenotype (that is, over a period of years), the genetic correlation with BW was estimated using each set of summary statistics and presented in Supplementary Table 12. The phenotypes with multiple GWAS included height, BMI, waist–hip ratio (adjusted for BMI), total cholesterol, triglycerides, high density lipoprotein (HDL) and low density lipoprotein (LDL). The estimate of the genetic correlation between the multiple GWAS meta-analyses on the same phenotype were comparable and the later GWAS had a smaller standard error due to the increased sample size, so only the genetic correlation between BW and the most recent meta-analyses were presented in Fig. 2.

In the published GWAS for blood pressure[14] the phenotype was adjusted for BMI. Caution is needed when interpreting the genetic correlation between BW and BMI-adjusted SBP owing to the potential for collider bias[72]. Since BMI is associated with both blood pressure and BW, it is possible that the use of a blood pressure genetic score adjusted for BMI might bias the genetic correlation estimate towards a more negative value. To verify that the inverse genetic correlation with BW ($r_g = -0.26$, s.e. $= 0.05$, $P = 6.5 \times 10^{-9}$) was not due to collider bias caused by the BMI adjustment of the phenotype, we obtained an alternative estimate using UK Biobank GWAS data for SBP that was unadjusted for BMI and obtained a similar result ($R_g = -0.22$, s.e. $= 0.03$, $P = 5.5 \times 10^{-13}$). The SBP phenotype in the UK Biobank was prepared as follows. Two blood pressure readings were taken at assessment, approximately 5 min apart. We included all individuals with an automated blood pressure reading (taken using an automated Omron blood pressure monitor). Two valid measurements were available for most participants (averaged to create a blood pressure variable, or alternatively a single reading was used if only one was available). Individuals were excluded if the two readings differed by more than 4.56 s.d. Blood pressure measurements more than 4.56 s.d. away from the mean were excluded. We accounted for blood pressure medication use by adding 15 mm Hg to the SBP measure. Blood pressure was adjusted for age, sex and centre location and then inverse rank normalized. We performed the GWAS on 127,698 individuals of British descent using BOLT-LMM (ref. 37), with genotyping array as covariate.

Estimating the proportion of the BW-adult traits covariance attributable to genotyped SNPs. We estimated the phenotypic, genetic and residual correlations as well as the genetic and residual covariance between BW and several quantitative traits and/or disease outcomes in the UK Biobank using directly genotyped SNPs and the REML method implemented in BOLT-LMM (ref. 37). The traits examined included T2D, SBP, diastolic blood pressure, CAD, height, BMI, weight, waist–hip ratio, hip circumference, waist circumference, obesity, overweight, age at menarche, asthma, and smoking. Where phenotypes were not available (for example, serum blood measures are not currently available in the UK Biobank), we obtained estimates using the NFBC1966 study (for correlations/covariance between BW and triglycerides, total cholesterol, HDL, LDL, fasting glucose and fasting insulin). In the UK Biobank analysis, we used 57,715 unrelated individuals with BW available and identified by the UK Biobank as white British. SNPs with evidence of deviation from Hardy–Weinberg equilibrium ($P < 1 \times 10^{-6}$), MAF < 0.05 or overall missing rate >0.015 were excluded, resulting in 328,928 SNPs for analysis. We included the first five ancestry principal components as covariates. In the NFBC1966 analysis, 5,009 individuals with BW were enrolled. Genotyped SNPs that passed quality control (Supplementary Table 2) were included, resulting in 324,895 SNPs for analysis. The first three ancestry principal components and sex were included as covariates.

Gene set enrichment analysis. Meta-analysis gene-set enrichment of variant associations (MAGENTA) was used to explore pathway-based associations using summary statistics from the trans-ancestry meta-analysis. MAGENTA implements a gene set enrichment analysis (GSEA) based approach, as previously described[9]. Briefly, each gene in the genome was mapped to a single index SNP with the lowest *P* value within a 110 kb upstream and 40 kb downstream window. This *P* value, representing a gene score, was then corrected for confounding factors such as gene size, SNP density and linkage-disequilibrium-related properties in a regression model. Genes within the HLA-region were excluded from analysis due to difficulties in accounting for gene density and linkage-disequilibrium patterns. Each mapped gene in the genome was then ranked by its adjusted gene score. At a given significance threshold (95th and 75th percentiles of all gene scores), the observed number of gene scores in a given pathway, with a ranked score above the specified threshold percentile, was calculated. This observed statistic was then compared to 1,000,000 randomly permuted pathways of identical size. This generates an empirical GSEA *P* value for each pathway. Significance was attained when an individual pathway reached a FDR < 0.05 in either analysis. In total, 3,216 pre-defined biological pathways from Gene Ontology, PANTHER, KEGG and Ingenuity were tested for enrichment of multiple modest associations with BW. The MAGENTA software was also used for enrichment testing of custom gene sets.

Protein–protein interaction network analyses. We used the integrative protein-interaction-network-based pathway analysis (iPINBPA) method[73]. Briefly, we generated gene-wise *P* values from the trans-ancestry meta-analysis using VEGAS2 (ref. 74), which mapped the SNPs to genes and accounted for possible confounders, such as linkage-disequilibrium between markers. The empirical gene-wise *P* values were calculated using simulations from the multivariate normal distribution. Those that were nominally significant ($P \leq 0.01$) were selected as 'seed genes', and were collated within a high confidence version of inweb3 (ref. 75) to weight the nodes in the network following a guilt-by-association approach. In a second step, a network score was defined by the combination of the *Z* scores derived from the gene-wise *P* values with node weights using the Liptak-Stouffer method[76]. A heuristic algorithm was then applied to extensively search for modules enriched in genes with low *P* values. The modules were further normalized using a null distribution of 10,000 random networks. Only those modules with *Z* score > 5 were selected. Finally, the union of all modules constructed a BW-overall PPI network. Both the proteins on the individual modules and on the overall BW-PPI were interrogated for enrichment in Gene Ontology terms (biological processes) using a hypergeometric test. Terms were considered as significant when the adjusted *P* value, following the Benjamini–Hochberg procedure, was below 0.05.

Point of contact analyses. The same methodology described above was applied to 16 different adult traits resulting in a number of enriched modules per trait. Different modules for each trait were combined in a single component and the intersection between these trait-specific components and the BW component was calculated. This intersection was defined as the PoC network. We used the resulting PoC networks in downstream analyses to interrogate which set of proteins connected BW variation and adult trait variation via pathways enriched in the overall BW analysis.

Parent-of-origin specific associations. We first searched for evidence of parent-of-origin effects in the UK Biobank samples by comparing variance between heterozygotes and homozygotes using Quicktest (ref. 77). In this analysis, we used only unrelated individuals identified genetically as of white British origin ($n = 57{,}715$). Principal components were generated using these individuals and the first five were used to adjust for population structure as covariates in the analysis, in addition to a binary indicator for genotyping array.

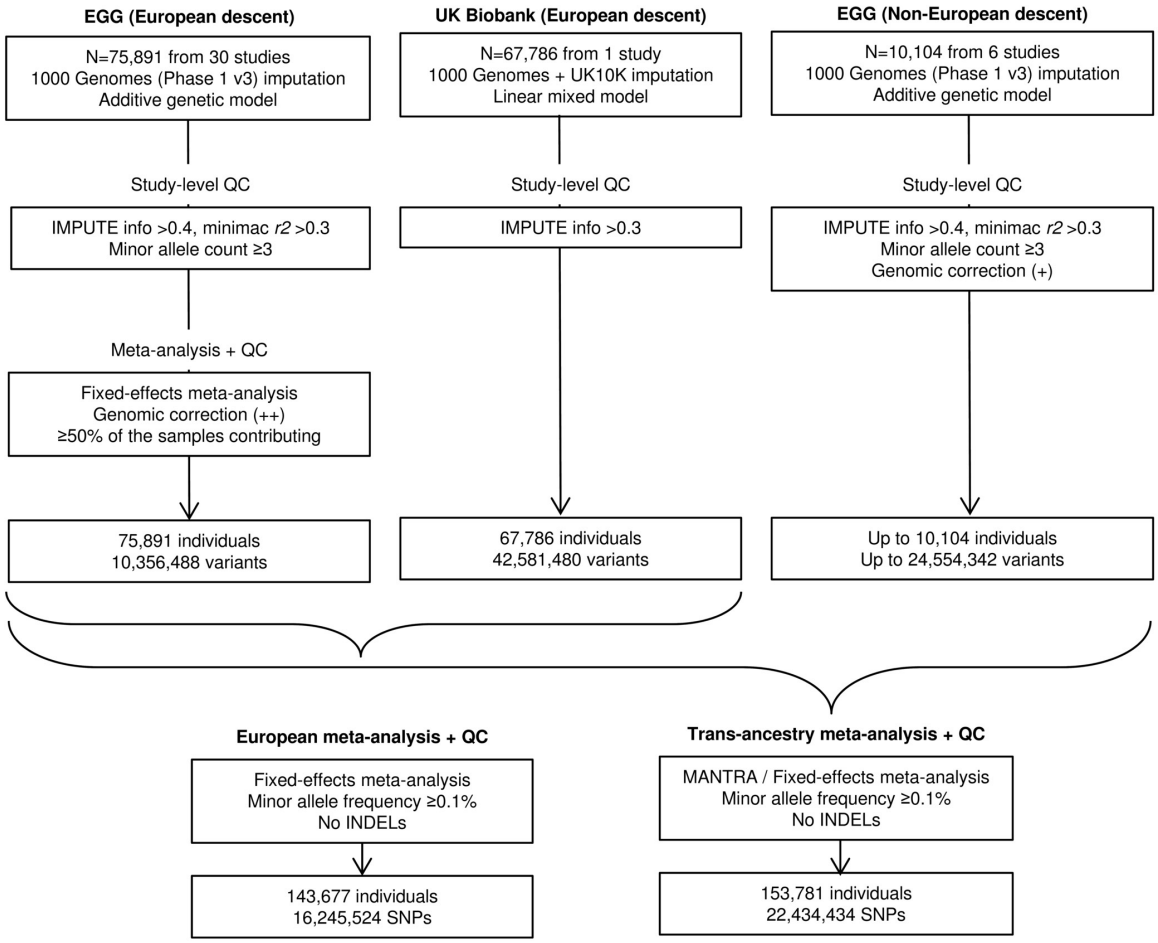
We also examined 4,908 mother–child pairs in ALSPAC and determined the parental origin of the alleles where possible[78]. Briefly, the method used mother–child pairs to determine the parent of origin of each allele. For example, if the mother/child genotypes were AA/Aa, the child's maternal/paternal allele combination was A/a. For the situation where both mother and child were heterozygous, the child's maternal/paternal alleles could not be directly specified. However, the parental origin of the alleles could be determined by phasing the genotype data and comparing maternal and child haplotypes. We then tested these alleles for association with BW adjusting for sex and gestational age.

Statistical power in these currently available sample sizes was insufficient to rule out widespread parent-of-origin effects across the regions tested. Using the mean β (0.034 s.d.) and MAF (0.28) of the identified loci, we estimate that we would need at least 200,000 unrelated individuals or 70,000 mother–child pairs for 80% power to detect parent-of-origin effects at $P < 0.00085$.

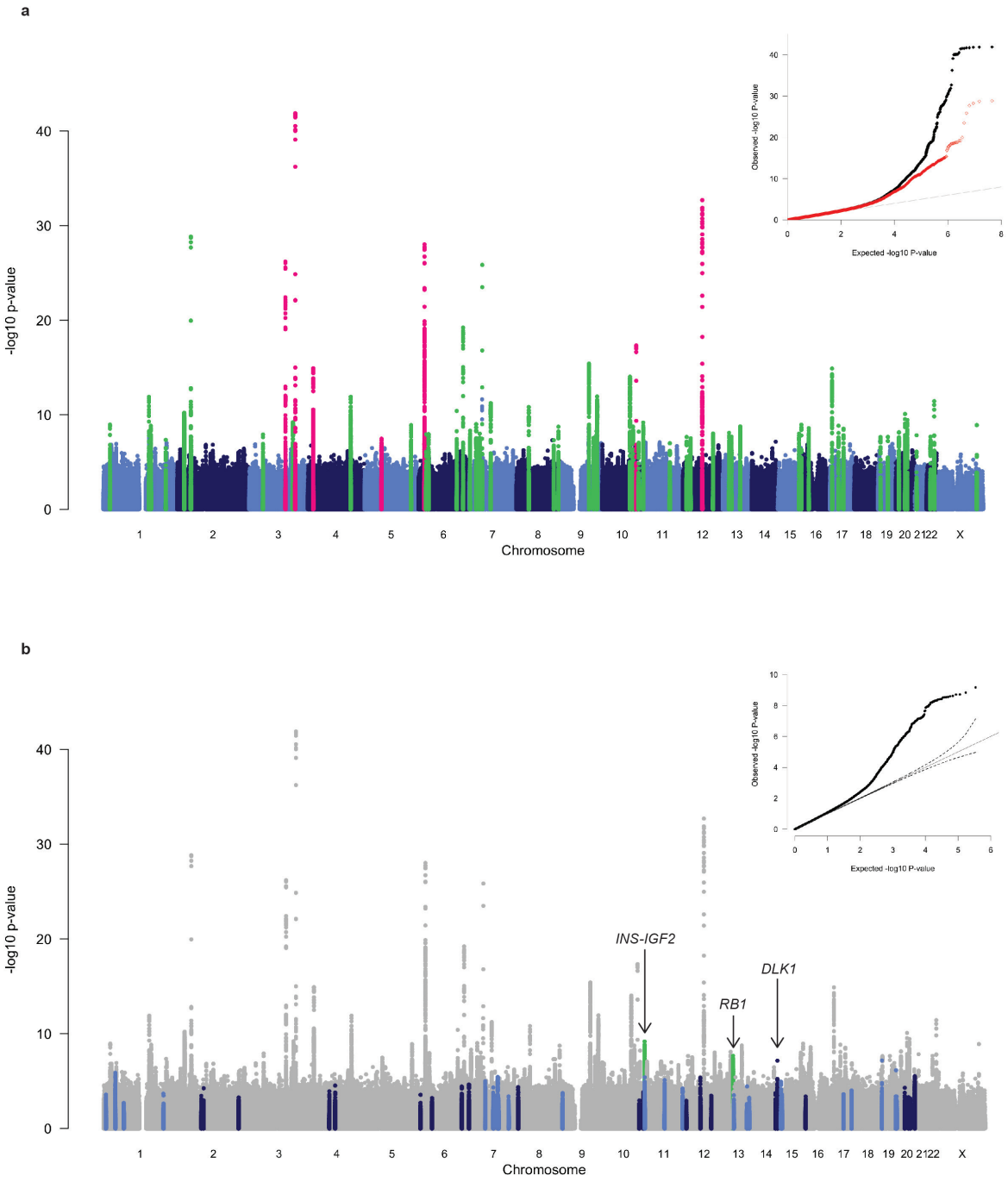
Hierarchical clustering of BW loci. To explore the different patterns of association between BW and other anthropometric/metabolic/endocrine traits and diseases, we performed hierarchical clustering analysis. The lead SNP (or proxy, EUR $R^2 > 0.6$) at the 60 BW loci was queried in publicly available GWAS meta-analysis datasets or in GWAS results obtained through collaboration[79]. Results were available for 53 of those loci and the extracted *Z* score (allelic effect/s.e., Supplementary Table 17) was aligned to the BW-raising allele. We performed two dimensional clustering by trait and by locus. We computed the Euclidean distance amongst *Z* scores of the extracted traits and loci and performed complete hierarchical clustering implemented in the pvclust package (http://www.sigmath.es.osaka-u.ac.jp/shimo-lab/prog/pvclust/) in R v3.2.0 (http://www.R-project.org/). Clustering uncertainty was measured by multiscale bootstrap resampling estimated from 1,000 replicates. We used $\alpha = 0.05$ to define distinct clusters and, based on the bootstrap analysis, calculated the Calinski index to identify the number of well-supported clusters (cascadeKM function, vegan package, http://CRAN.R-project.org/package=vegan). Clustering was visualized by constructing dendrograms and a heat map.

Separately from the hierarchical clustering analysis, we queried the lead SNP at *EPAS1* in a GWAS of haematological traits[80] because variation at that locus has previously been implicated in BW and adaptation to hypoxia at high altitudes in Tibetans[81,82] (Supplementary Table 17).

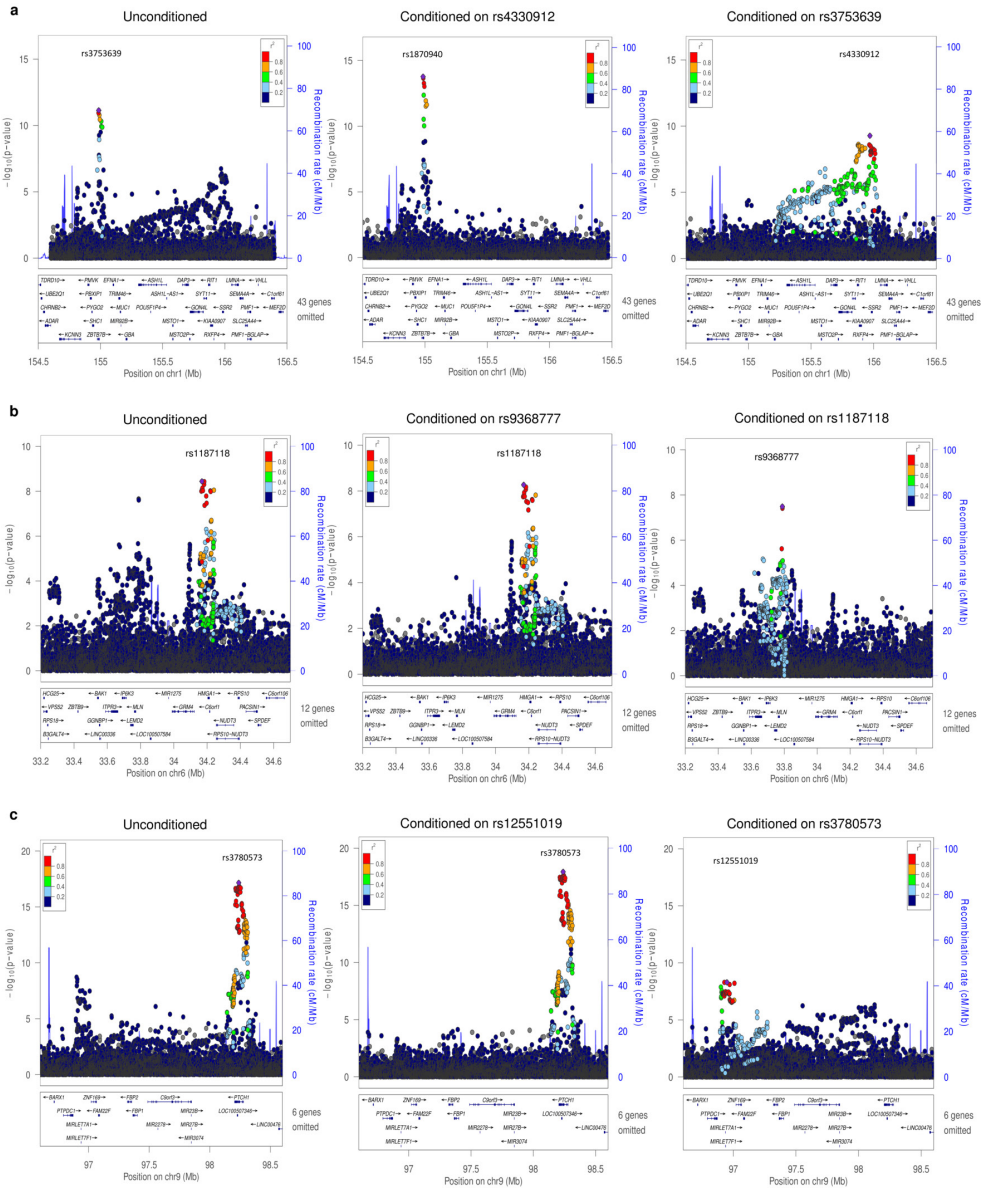
30. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11,** 499–511 (2010).
31. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **44,** 955–959 (2012).
32. Winkler, T. W. *et al.* Quality control and conduct of genome-wide association meta-analyses. *Nat. Protoc.* **9,** 1192–1212 (2014).
33. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38,** 904–909 (2006).
34. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55,** 997–1004 (1999).
35. Kang, H. M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42,** 348–354 (2010).
36. Allen, N. E., Sudlow, C., Peakman, T. & Collins, R. UK Biobank data: come and get it. *Sci. Transl. Med.* **6,** 224ed4 (2014).
37. Loh, P. R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47,** 284–290 (2015).
38. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).
39. Mägi, R. & Morris, A. P. GWAMA: software for genome-wide association meta-analysis. *BMC Bioinformatics* **11,** 288 (2010).
40. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26,** 2190–2191 (2010).
41. Ioannidis, J. P., Patsopoulos, N. A. & Evangelou, E. Heterogeneity in meta-analyses of genome-wide association investigations. *PLoS One* **2,** e841 (2007).
42. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44,** 369–375, S1–S3 (2012).
43. GTEx Consortium. Human genomics. The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348,** 648–660 (2015).
44. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501,** 506–511 (2013).
45. Montgomery, S. B. *et al.* Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464,** 773–777 (2010).
46. Schadt, E. E. *et al.* Mapping the genetic architecture of gene expression in human liver. *PLoS Biol.* **6,** e107 (2008).
47. Gibbs, J. R. *et al.* Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet.* **6,** e1000952 (2010).
48. Stranger, B. E. *et al.* Population genomics of human gene expression. *Nat. Genet.* **39,** 1217–1224 (2007).
49. Li, Q. *et al.* Expression QTL-based analyses reveal candidate causal genes and loci across five tumor types. *Hum. Mol. Genet.* **23,** 5294–5302 (2014).
50. Westra, H. J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45,** 1238–1243 (2013).
51. Zou, F. *et al.* Brain expression genome-wide association study (eGWAS) identifies human disease-associated variants. *PLoS Genet.* **8,** e1002707 (2012).
52. Hao, K. *et al.* Lung eQTLs to help reveal the molecular underpinnings of asthma. *PLoS Genet.* **8,** e1003029 (2012).
53. Koopmann, T. T. *et al.* Genome-wide identification of expression quantitative trait loci (eQTLs) in human heart. *PLoS One* **9,** e97380 (2014).
54. Fairfax, B. P. *et al.* Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science* **343,** 1246949 (2014).
55. Grundberg, E. *et al.* Global analysis of the impact of environmental perturbation on cis-regulation of gene expression. *PLoS Genet.* **7,** e1001279 (2011).
56. Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* **40,** D930–D934 (2012).
57. Flicek, P. *et al.* Ensembl 2014. *Nucleic Acids Res.* **42,** D749–D755 (2014).
58. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4,** 1073–1081 (2009).
59. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7,** 248–249 (2010).
60. Szklarczyk, D. *et al.* STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **43,** D447–D452 (2015).
61. The International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* **467,** 52–58 (2010).
62. Morris, A. P. Transethnic meta-analysis of genome-wide association studies. *Genet. Epidemiol.* **35,** 809–822 (2011).
63. The Wellcome Trust Case Control Consortium. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat. Genet.* **44,** 1294–1301 (2012).
64. Wang, X. *et al.* Comparing methods for performing trans-ethnic meta-analysis of genome-wide association studies. *Hum. Mol. Genet.* **22,** 2303–2311 (2013).
65. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
66. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22,** 1760–1774 (2012).
67. Pickrell, J. K. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.* **94,** 559–573 (2014).
68. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42,** 565–569 (2010).
69. Urbanek, M. *et al.* The chromosome 3q25 genomic region is associated with measures of adiposity in newborns in a multi-ethnic genome-wide association study. *Hum. Mol. Genet.* **22,** 3583–3596 (2013).
70. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47,** 1236–1241 (2015).
71. The International HapMap Consortium *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449,** 851–861 (2007).
72. Aschard, H., Vilhjálmsson, B. J., Joshi, A. D., Price, A. L. & Kraft, P. Adjusting for heritable covariates can bias effect estimates in genome-wide association studies. *Am. J. Hum. Genet.* **96,** 329–339 (2015).
73. Wang, L., Mousavi, P. & Baranzini, S. E. iPINBPA: an integrative network-based functional module discovery tool for genome-wide association studies. *Pac. Symp. Biocomput.* 255–266 (2015).
74. Mishra, A. & Macgregor, S. VEGAS2: software for more flexible gene-based testing. *Twin Res. Hum. Genet.* **18,** 86–91 (2015).
75. Lage, K. *et al.* A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.* **25,** 309–316 (2007).
76. Whitlock, M. C. Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. *J. Evol. Biol.* **18,** 1368–1373 (2005).
77. Hoggart, C. J. *et al.* Novel approach identifies SNPs in SLC2A10 and KCNK9 with evidence for parent-of-origin effect on body mass index. *PLoS Genet.* **10,** e1004508 (2014).
78. Wang, S., Yu, Z., Miller, R. L., Tang, D. & Perera, F. P. Methods for detecting interactions between imprinted genes and environmental exposures using birth cohort designs with mother-offspring pairs. *Hum. Hered.* **71,** 196–208 (2011).
79. Painter, J. N. *et al.* Genome-wide association study identifies a locus at 7p15.2 associated with endometriosis. *Nat. Genet.* **43,** 51–54 (2011).
80. Ganesh, S. K. *et al.* Multiple loci influence erythrocyte phenotypes in the CHARGE Consortium. *Nat. Genet.* **41,** 1191–1198 (2009).
81. Xu, X. H. *et al.* Two functional loci in the promoter of EPAS1 gene involved in high-altitude adaptation of Tibetans. *Sci. Rep.* **4,** 7465 (2014).
82. Huerta-Sánchez, E. *et al.* Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* **512,** 194–197 (2014).



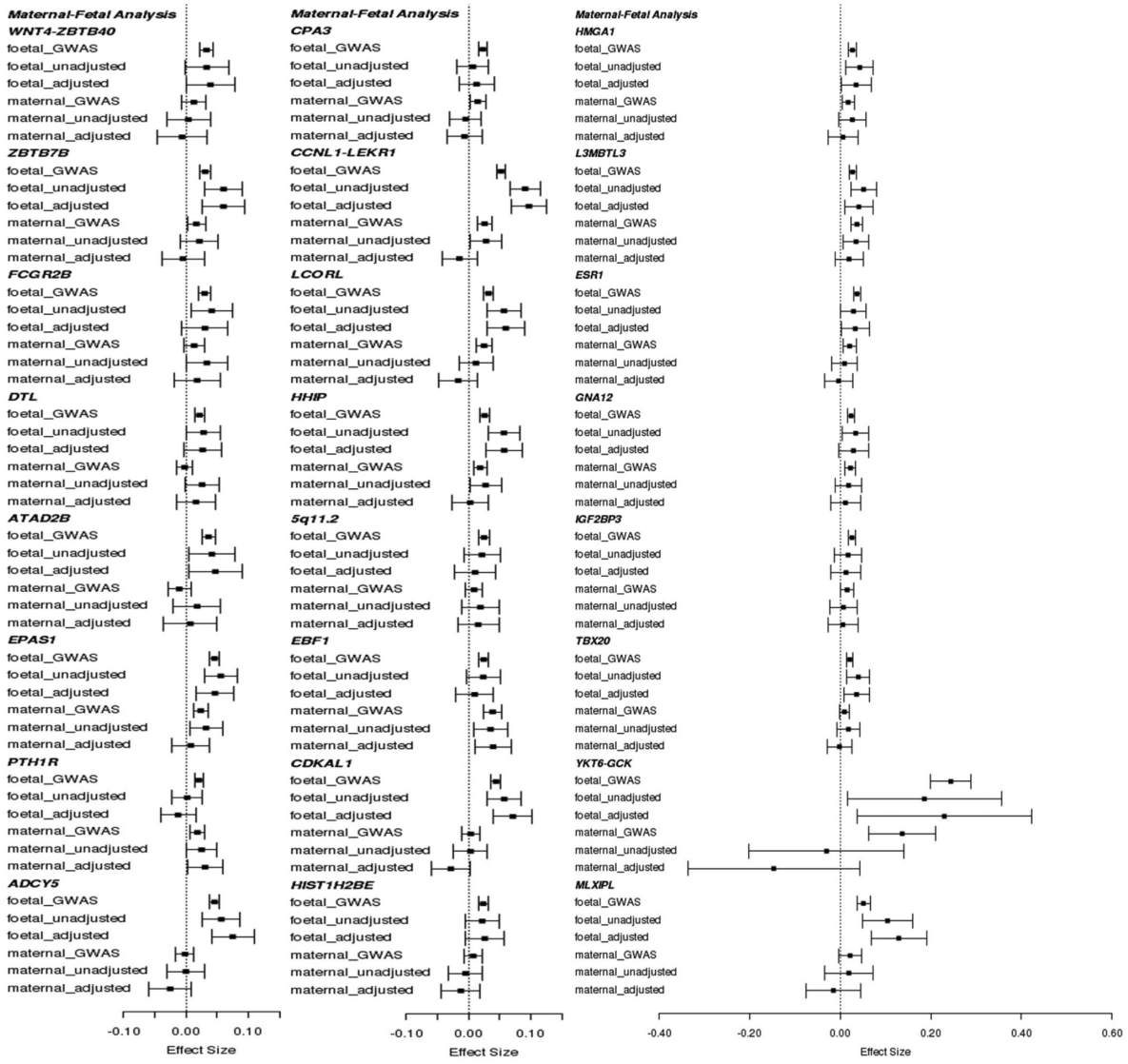
Extended Data Figure 1 | Flow chart of the study design.



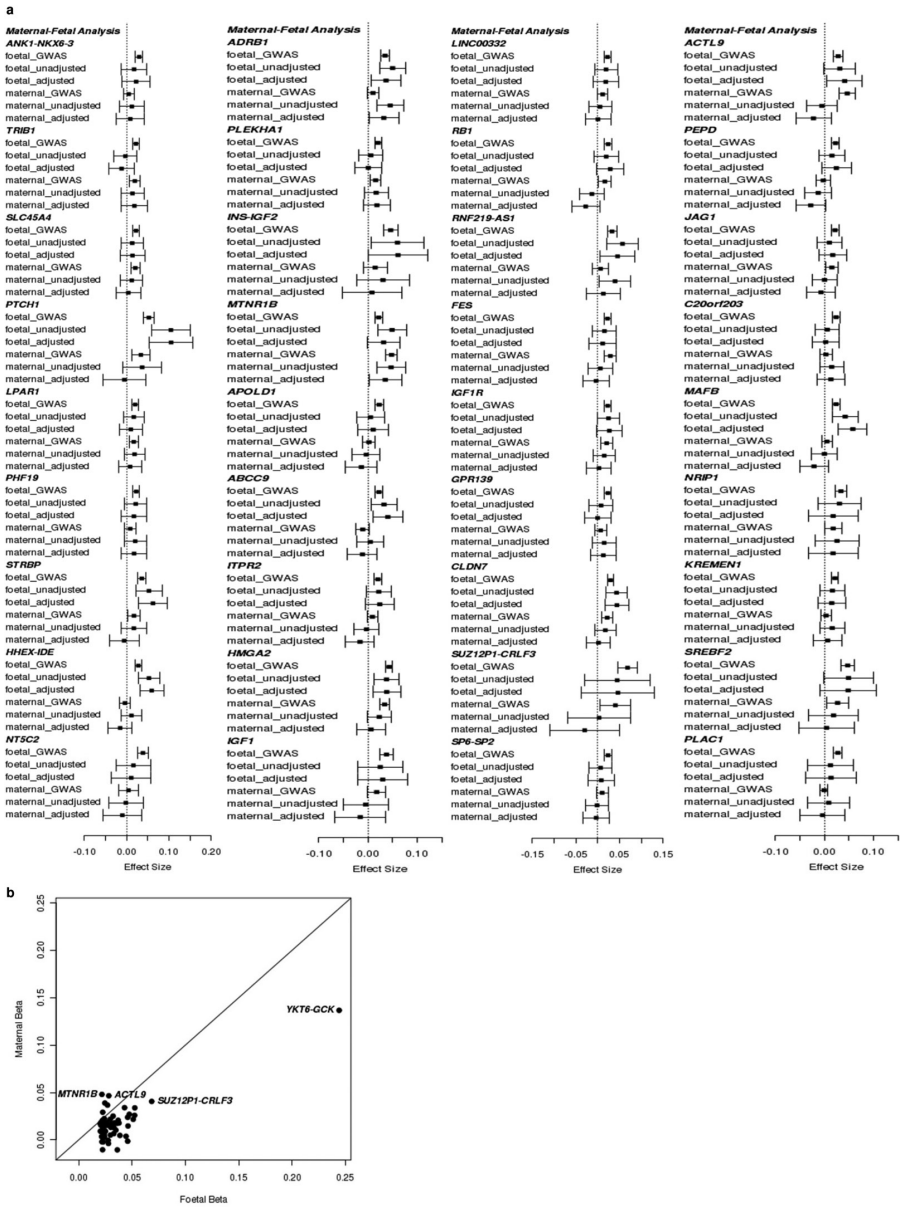
Extended Data Figure 2 | Manhattan and quantile–quantile (QQ) plots of the trans-ancestry meta-analysis for BW. a, Manhattan (main panel) and QQ (top right) plots of genome-wide association results for BW from trans-ancestry meta-analysis of up to 153,781 individuals. The association *P* value (on $-\log_{10}$ scale) for each of up to 22,434,434 SNPs (*y* axis) was plotted against the genomic position (NCBI Build 37; *x* axis). Association signals that reached genome-wide significance ($P < 5 \times 10^{-8}$) are shown in green if novel and pink if previously reported. In the QQ plot, the black dots represent observed *P* values and the grey line represents expected *P* values under the null distribution. The red dots represent observed *P* values after excluding the previously identified signals[5]. **b**, Manhattan (main panel) and QQ (top right) plots of trans-ethnic GWAS meta-analysis for BW highlighting the reported imprinted regions described in Supplementary Table 14. Novel association signals that reached genome-wide significance ($P < 5 \times 10^{-8}$) and mapped to imprinted regions are shown in green. Genomic regions outside imprinted regions are shaded in grey. SNPs in the imprinted regions are shown in light blue or dark blue, depending on chromosome number (odd or even). In the QQ plot, the black dots represent observed *P* values and the grey lines represent expected *P* values and their 95% confidence intervals under the null distribution for the SNPs within the imprinted regions.



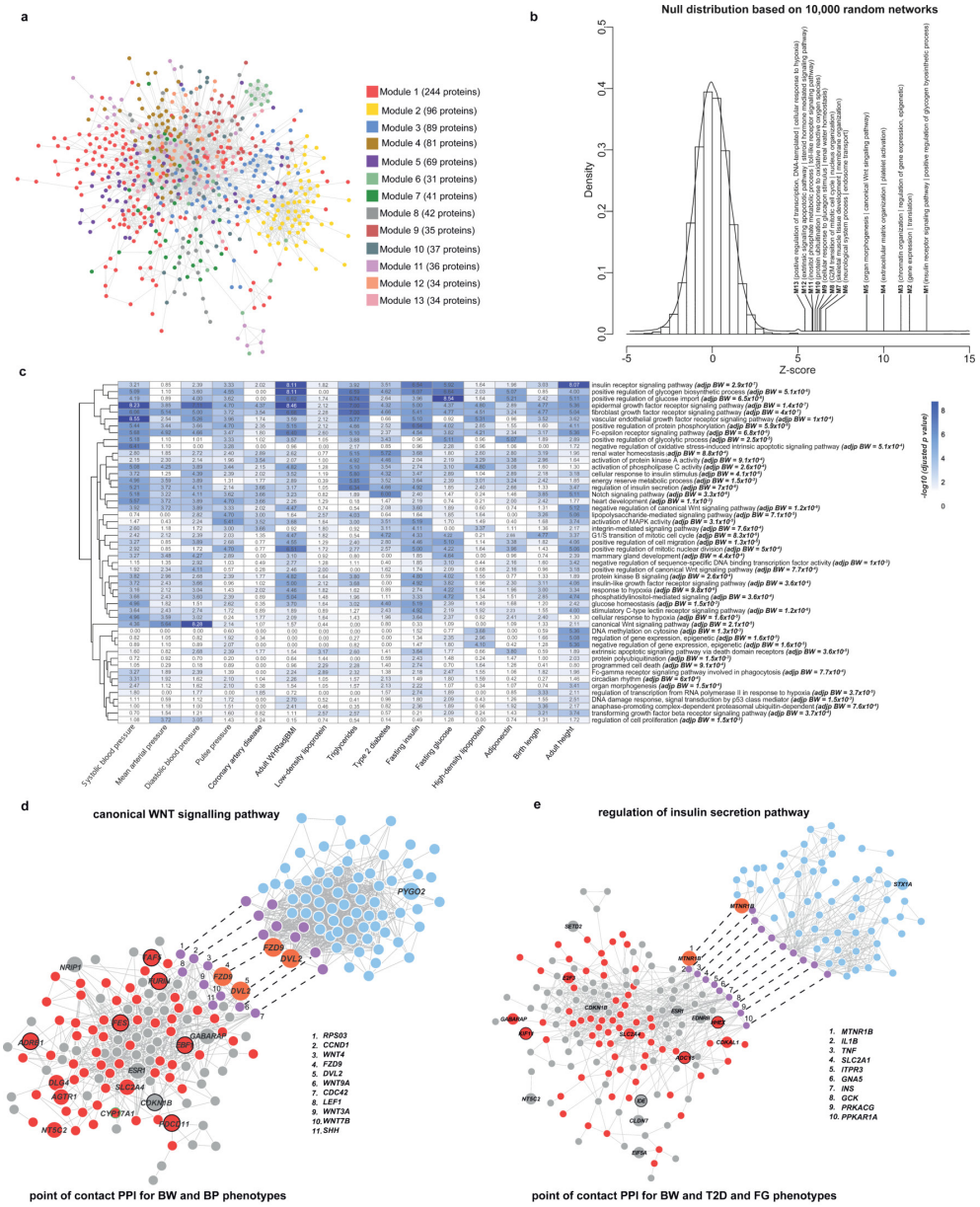
Extended Data Figure 3 | Regional plots for multiple distinct signals at three BW loci. Regional plots for each locus, *ZBTB7B* (**a**), *HMGA1* (**b**) and *PTCH1* (**c**), are displayed from: the unconditional European-specific meta-analysis of up to 143,677 individuals (left); the approximate conditional meta-analysis for the primary signal after adjustment for the index variant for the secondary signal (middle); and the approximate conditional meta-analysis for the secondary signal after adjustment for the index variant for the primary signal (right). Directly genotyped or imputed SNPs were plotted with their association *P* values (on a $-\log_{10}$ scale) as a function of genomic position (NCBI Build 37). Estimated recombination rates (blue lines) were plotted to reflect the local linkage-disequilibrium structure around the index SNPs and their correlated proxies. SNPs were coloured in reference to linkage-disequilibrium with the particular index SNP according to a blue to red scale from $R^2 = 0$ to 1, based on pairwise R^2 values estimated from a reference of 5,000 individuals of white British origin, randomly selected from the UK Biobank.



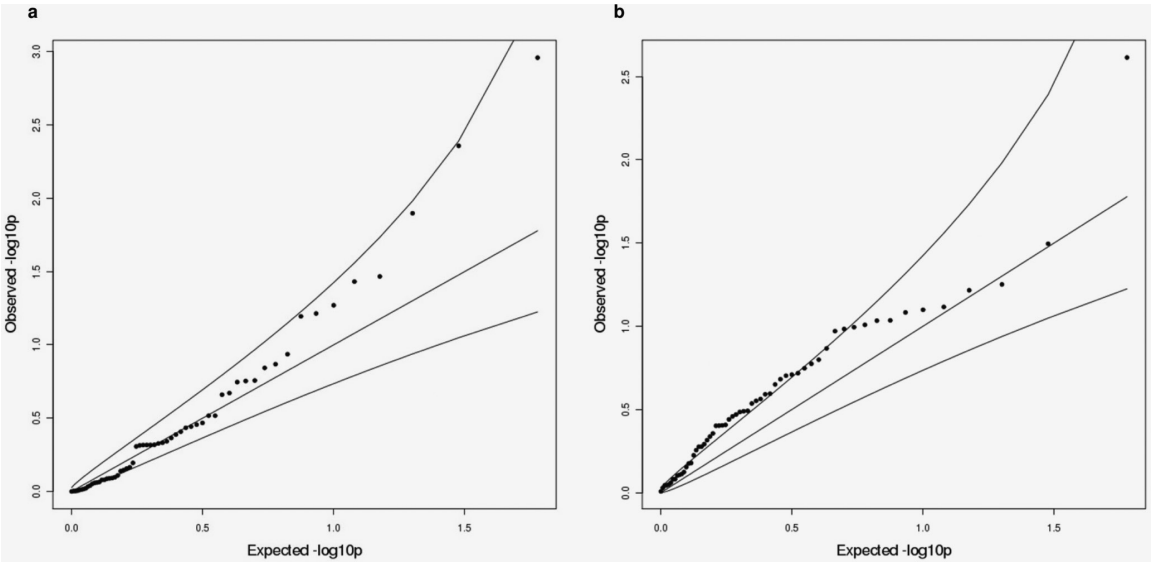
Extended Data Figure 4 | Comparison of fetal effect sizes and maternal effect sizes at 60 known and novel birth weight loci, for the first 24 loci. The remaining loci are shown in Extended Data Fig. 5a. For each BW locus, the following six effect sizes (with 95% CI) are shown, all aligned to the same BW-raising allele: fetal_GWAS, fetal allelic effect on BW (from European ancestry meta-analysis of up to $n = 143,677$ individuals); fetal_unadjusted, fetal allelic effect on BW (unconditioned in $n = 12,909$ mother–child pairs); fetal_adjusted, fetal effect (conditioned on maternal genotype, $n = 12,909$); maternal_GWAS, maternal allelic effect on offspring BW (from meta-analysis of up to $n = 68,254$ European mothers)[7]; maternal_unadjusted, maternal allelic effect on offspring BW (unconditioned, $n = 12,909$); maternal_adjusted, maternal effect (conditioned on fetal genotype, $n = 12,909$). The 60 BW loci were ordered by chromosome and position (Supplementary Tables 10, 11). These plots illustrate that, in large GWAS of BW, fetal effect size estimates are larger than those of maternal at 55 out of 60 identified loci (binomial $P = 1 \times 10^{-11}$), suggesting that most of the associations are driven by the fetal genotype. In conditional analyses that modelled the effects of both maternal and fetal genotypes ($n = 12,909$ mother–child pairs), confidence intervals around the estimates were wide, precluding inference about the likely contribution of maternal versus fetal genotype at individual loci.



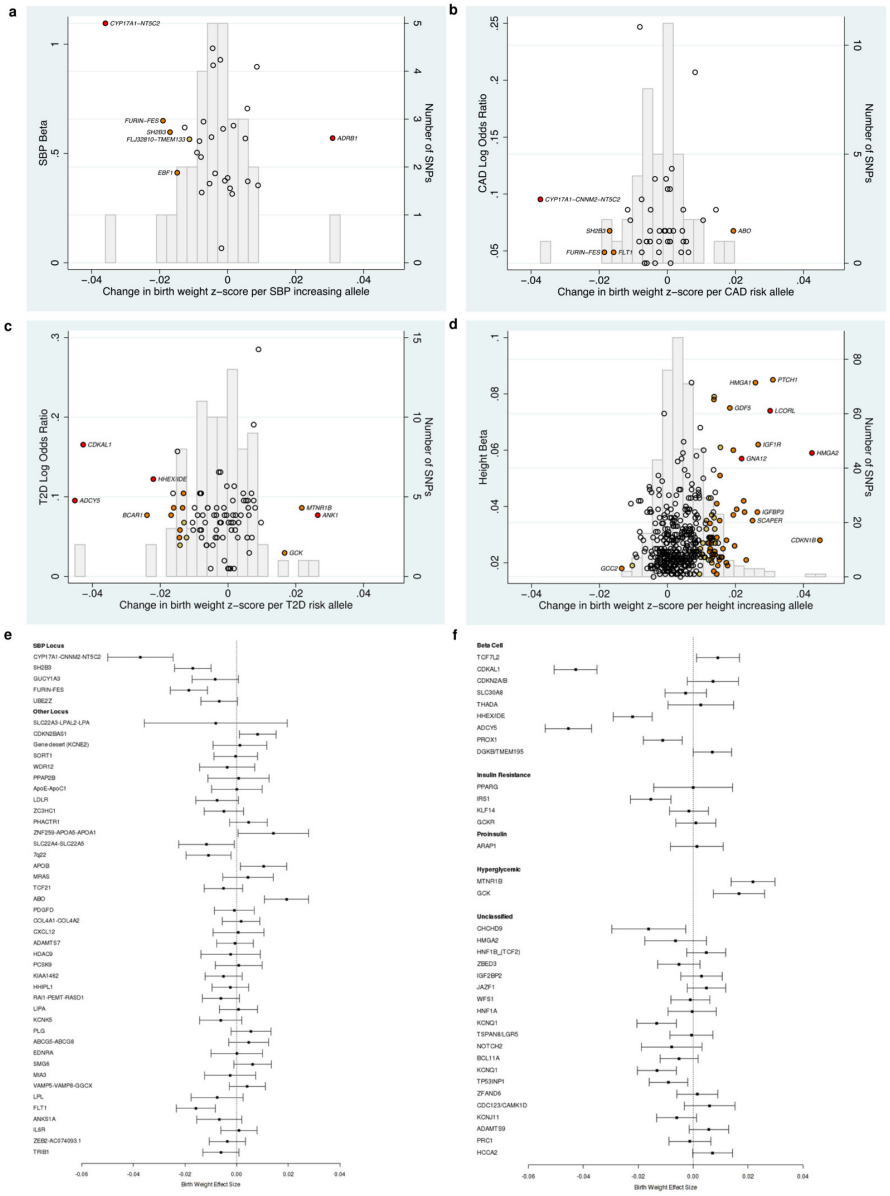
Extended Data Figure 5 | Comparison of fetal effect sizes and maternal effect sizes at 60 known and novel birth weight loci, for the remaining 36 loci. a, Continued from Extended Data Fig. 4. **b**, The scatter plot illustrates the difference between the fetal (*x* axis) and maternal (*y* axis) effect sizes in the overall maternal versus fetal GWAS results.



Extended Data Figure 6 | Protein–protein Interaction (PPI) Network analysis. a, The largest global component of BW PPI network containing 13 modules is shown. **b**, The histogram shows the null distribution of *Z* scores of BW PPI networks based on 10,000 random networks, and where the *Z* scores for the 13 BW modules (M1–13) lie. For each module, the two most significant GO terms are shown. **c**, A heat map is shown, which takes the top 50 biological processes over-represented in the global BW PPI network (listed at the right of the plot), and displays the extent of enrichment for the various trait-specific "point of contact" (PoC) PPI networks. **d**, **e**, Trait-specific PoC PPI networks composed of proteins that are shared in both the global BW PPI network and networks generated using the same pipeline for each of the adult traits: **d**, canonical Wnt signalling pathway enriched for PoC PPI between BW and blood pressure (BP)-related phenotypes; and **e**, regulation of insulin secretion pathway enriched for PoC between BW and T2D/fasting glucose (FG). Red nodes indicate those present in PoC for BW and traits of interest; blue nodes correspond to the pathway nodes; purple nodes are those present in both the pathway and PoC; orange nodes are genes in BW loci that overlap with both the pathway and PoC. Large nodes correspond to genes in BW loci (within 300 kb from the lead SNP), and have a black border if they, amongst all BW loci, have a stronger (top 5) association with at least one of the pairing adult traits.



Extended Data Figure 7 | Quantile–Quantile (QQ) plots of variance comparison between heterozygotes and homozygotes analysis in 57,715 UK Biobank samples and parent-of-origin specific analysis in 4,908 ALSPAC mother–child pairs at 59 autosomal BW loci plus *DLK1*. a, QQ plot from the Quicktest analysis (ref. 77) comparing the BW variance of heterozygotes with homozygotes in 57,715 UK Biobank samples. **b**, QQ plot from the parent-of-origin specific analysis testing the association between BW and maternally transmitted versus paternally transmitted alleles in 4,908 mother–child pairs from the ALSPAC study (Methods, Supplementary Tables 15, 16). In both panels, the black dots represent lead SNPs at 59 identified autosomal BW loci and a further sub-genome-wide significant signal for BW near *DLK1* (rs6575803; $P = 5.6 \times 10^{-8}$). The grey lines represent expected *P* values and their 95% confidence intervals under the null distribution for the 60 SNPs. Both results show trends in favour of imprinting effects at BW loci; however, despite the large sample size, these analyses were underpowered (see Methods) and much larger sample sizes are required for definitive analysis.



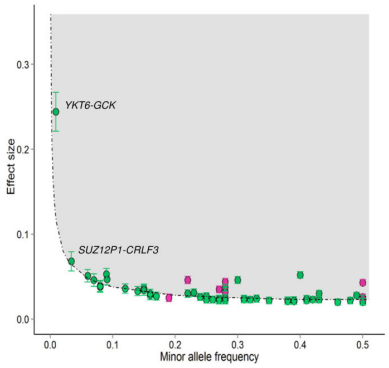
Extended Data Figure 8 | Summary of previously reported loci for SBP, CAD, T2D and adult height and their effect on birth weight. a–d, Effect sizes (left *y* axis) of previously reported 30 SBP loci[13,14], 45 CAD loci[23], 84 T2D loci[24] and 422 adult height loci[25] were plotted against effects on BW (*x* axis). Effect sizes were aligned to the adult trait (or risk)-raising allele. The colour of each dot indicates BW association *P* value: red, $P < 5 \times 10^{-8}$; orange, $5 \times 10^{-8} \leq P < 0.001$; yellow, $0.001 \leq P < 0.01$; white, $P \geq 0.01$. The superimposed grey frequency histogram shows the number of SNPs (right *y* axis) in each category of BW effect size. **e**, Effect sizes (with 95% CI) on BW of 45 known CAD loci were plotted arranged in the order of CAD effect size from highest to lowest, separating out the known SBP loci. CAD loci with a larger effect on BW concentrated amongst loci with primary blood pressure association. **f**, Effect sizes (with 95% CI) on BW of 32 known T2D loci were plotted, subdivided by previously reported categories derived from detailed adult physiological data[27]. Heterogeneity in BW effect sizes between five T2D loci groups with different mechanistic categories was substantial (Cochran's *Q* statistic $P_{het} = 1.2 \times 10^{-9}$). In pairwise comparisons, the 'beta cell' group of variants differed from the other four groups: fasting hyperglycaemia ($P_{het} = 3 \times 10^{-11}$), insulin resistance ($P_{het} = 0.002$), proinsulin ($P_{het} = 0.78$) and unclassified ($P_{het} = 0.02$) groups. All of the BW effect sizes plotted in the forest plots were aligned to the trait (or risk)-raising allele.

Extended Data Table 1 | Sixty loci associated with BW ($P < 5 \times 10^{-8}$) in European ancestry meta-analysis of up to 143,677 individuals and/or trans-ancestry meta-analysis of up to 153,781 individuals

a

Locus	Lead SNP	Chr.	Position (bp, b37)	Alleles Effect/Other	EAF	European ancestry β (SE)	European ancestry *P*-value	Trans-ancestry β (SE)	Trans-ancestry *P*-value
Previously reported loci									
CCNL1-LEKR1	rs13322435	3	156,795,468	A/G	0.59	0.053 (0.004)	3.7×10^{-41}	0.052 (0.004)	1.3×10^{-42}
HMGA2	rs1351394	12	66,351,826	T/C	0.48	0.044 (0.004)	1.9×10^{-32}	0.043 (0.004)	2.0×10^{-33}
CDKAL1	rs35261542	6	20,675,792	C/A	0.73	0.044 (0.004)	4.4×10^{-27}	0.044 (0.004)	9.7×10^{-29}
ADCY5	rs11719201	3	123,068,744	T/C	0.23	0.046 (0.004)	2.4×10^{-26}	0.046 (0.004)	6.4×10^{-27}
ADRB1	rs7076938	10	115,789,375	T/C	0.73	0.036 (0.004)	4.7×10^{-18}	0.035 (0.004)	4.7×10^{-18}
LCORL	rs925098	4	17,919,811	G/A	0.28	0.034 (0.004)	5.4×10^{-16}	0.032 (0.004)	1.3×10^{-15}
5q11.2	rs854037	5	57,091,783	A/G	0.80	0.027 (0.005)	2.2×10^{-8}	0.025 (0.005)	3.5×10^{-8}
Novel loci									
EPAS1	rs1374204	2	46,484,205	T/C	0.70	0.047 (0.004)	6.2×10^{-29}	0.046 (0.004)	1.5×10^{-29}
YKT6-GCK	rs138715366	7	44,246,271	C/T	0.99	0.241 (0.023)	7.2×10^{-26}	0.244 (0.023)	1.4×10^{-26}
ESR1	rs1101081	6	152,032,917	C/T	0.73	0.038 (0.004)	1.6×10^{-19}	0.037 (0.004)	6.1×10^{-20}
PTCH1	rs28510415	9	98,245,026	G/A	0.09	0.056 (0.007)	1.5×10^{-17}	0.053 (0.006)	4.0×10^{-18}
CLDN7	rs113086489	17	7,171,356	T/C	0.55	0.031 (0.004)	9.1×10^{-16}	0.030 (0.004)	1.3×10^{-15}
HHEX-IDE	rs61862780	10	94,468,643	T/C	0.52	0.028 (0.004)	3.0×10^{-14}	0.028 (0.004)	9.5×10^{-15}
STRBP	rs700059	9	125,824,055	G/A	0.16	0.033 (0.005)	4.7×10^{-10}	0.036 (0.005)	1.2×10^{-12}
HHIP	rs6537307	4	145,601,863	G/A	0.48	0.025 (0.004)	9.5×10^{-12}	0.026 (0.004)	1.3×10^{-12}
ZBTB7B	rs3753639	1	154,986,091	C/T	0.23	0.031 (0.004)	7.3×10^{-12}	0.031 (0.004)	1.3×10^{-12}
SREBF2	rs62240962	22	42,259,524	C/T	0.92	0.047 (0.007)	9.7×10^{-12}	0.047 (0.007)	3.7×10^{-12}
MLXIPL	rs62466330	7	73,056,805	C/T	0.07	0.049 (0.008)	1.2×10^{-10}	0.051 (0.007)	5.9×10^{-12}
ANK1-NKX6-3	rs13266210	8	41,533,514	A/G	0.79	0.031 (0.005)	1.3×10^{-11}	0.030 (0.004)	1.6×10^{-11}
L3MBTL3	rs1415701	6	130,345,835	G/A	0.73	0.025 (0.004)	2.6×10^{-9}	0.027 (0.004)	4.0×10^{-11}
ATAD2B	rs7575873	2	23,962,647	A/G	0.88	0.038 (0.006)	1.3×10^{-11}	0.036 (0.006)	6.2×10^{-11}
C20orf203	rs28530618	20	31,275,581	A/G	0.50	0.026 (0.004)	7.7×10^{-12}	0.024 (0.004)	8.4×10^{-11}
MAFB	rs6016377	20	39,172,728	T/C	0.45	0.024 (0.004)	9.5×10^{-10}	0.024 (0.004)	3.7×10^{-10}
CPA3	rs10935733	3	148,622,968	T/C	0.42	0.022 (0.004)	9.2×10^{-9}	0.023 (0.004)	6.2×10^{-10}
INS-IGF2	rs72851023	11	2,130,620	T/C	0.07	0.048 (0.008)	2.9×10^{-10}	0.046 (0.007)	6.8×10^{-10}
IGF2BP3	rs11765649	7	23,479,013	T/C	0.76	0.027 (0.004)	5.8×10^{-10}	0.026 (0.004)	1.0×10^{-9}
WNT4-ZBTB40	rs2473248	1	22,536,643	C/T	0.87	0.033 (0.006)	1.1×10^{-8}	0.033 (0.005)	1.1×10^{-9}
IGF1R	rs7402982	15	99,193,269	A/G	0.42	0.023 (0.004)	2.3×10^{-9}	0.023 (0.004)	1.1×10^{-9}
PLAC1	rs11096402	X	133,827,868	G/A	0.25	0.028 (0.005)	1.3×10^{-8}	N/A	N/A
EBF1	rs7729301	5	157,886,953	A/G	0.72	0.024 (0.004)	1.6×10^{-8}	0.025 (0.004)	1.3×10^{-9}
SUZ12P1-CRLF3	rs144843919	17	29,037,339	G/A	0.96	0.066 (0.012)	1.4×10^{-8}	0.068 (0.011)	1.5×10^{-9}
FCGR2B	rs72480273	1	161,644,871	C/A	0.17	0.031 (0.005)	8.0×10^{-10}	0.030 (0.005)	1.5×10^{-9}
RNF219-AS1	rs1819436	13	78,580,283	C/T	0.87	0.033 (0.006)	6.3×10^{-9}	0.033 (0.005)	1.8×10^{-9}
NT5C2	rs74233809	10	104,913,940	C/T	0.08	0.037 (0.007)	5.2×10^{-8}	0.039 (0.006)	1.8×10^{-9}
SLC45A4	rs12543725	8	142,247,979	G/A	0.60	0.023 (0.004)	1.2×10^{-9}	0.022 (0.004)	1.9×10^{-9}
GPR139	rs1011939	16	19,992,996	G/A	0.31	0.022 (0.004)	1.3×10^{-7}	0.024 (0.004)	2.7×10^{-9}
SP6-SP2	rs12942207	17	45,968,294	C/T	0.30	0.022 (0.004)	5.1×10^{-8}	0.024 (0.004)	3.0×10^{-9}
GNA12	rs798489	7	2,801,803	C/T	0.74	0.023 (0.004)	2.0×10^{-8}	0.024 (0.004)	5.0×10^{-9}
PHF19	rs7847628	9	123,631,225	G/A	0.67	0.023 (0.004)	1.0×10^{-8}	0.023 (0.004)	5.4×10^{-9}
PLEKHA1	rs2421016	10	124,167,512	T/C	0.48	0.021 (0.004)	1.8×10^{-8}	0.021 (0.004)	6.1×10^{-9}
JAG1	rs6040076	20	10,658,882	C/G	0.51	0.023 (0.004)	2.0×10^{-9}	0.022 (0.004)	7.2×10^{-9}
LINC00332	rs2324499	13	40,662,001	G/C	0.67	0.022 (0.004)	7.3×10^{-8}	0.023 (0.004)	8.3×10^{-9}
IGF1	rs7964361	12	102,994,878	A/G	0.08	0.039 (0.007)	4.7×10^{-9}	0.038 (0.007)	9.7×10^{-9}
FES	rs12906125	15	91,427,612	G/A	0.69	0.023 (0.004)	1.7×10^{-8}	0.023 (0.004)	1.0×10^{-8}
TBX20	rs6959887	7	35,295,365	A/G	0.61	0.023 (0.004)	1.5×10^{-9}	0.021 (0.004)	1.0×10^{-8}
HMGA1	rs7742369	6	34,165,721	G/A	0.19	0.028 (0.005)	1.0×10^{-8}	0.027 (0.005)	1.1×10^{-8}
HIST1H2BE	rs9379832	6	26,186,200	A/G	0.71	0.023 (0.004)	6.6×10^{-9}	0.024 (0.004)	1.2×10^{-8}
PTH1R	rs2242116	3	46,941,116	A/G	0.39	0.022 (0.004)	1.4×10^{-8}	0.021 (0.004)	1.2×10^{-8}
NRIP1	rs2229742	21	16,339,172	G/C	0.87	0.036 (0.006)	2.2×10^{-9}	0.034 (0.006)	1.5×10^{-8}
RB1	rs2854355	13	48,882,363	G/A	0.26	0.023 (0.004)	9.8×10^{-9}	0.024 (0.004)	2.2×10^{-8}
KREMEN1	rs134594	22	29,468,456	C/T	0.35	0.023 (0.004)	1.0×10^{-8}	0.022 (0.004)	2.2×10^{-8}
APOLD1	rs11055034	12	12,890,626	C/A	0.73	0.022 (0.004)	1.8×10^{-7}	0.023 (0.004)	2.3×10^{-8}
PEPD	rs10402712	19	33,926,013	A/G	0.27	0.022 (0.004)	4.4×10^{-7}	0.023 (0.004)	2.3×10^{-8}
ACTL9	rs61154119	19	8,787,750	T/G	0.84	0.028 (0.005)	1.1×10^{-7}	0.028 (0.005)	2.3×10^{-8}
LPAR1	rs2150052	9	113,945,067	T/A	0.50	0.021 (0.004)	2.2×10^{-8}	0.020 (0.004)	2.8×10^{-8}
ITPR2	rs12823128	12	26,872,730	T/C	0.56	0.021 (0.004)	1.9×10^{-8}	0.020 (0.004)	3.2×10^{-8}
DTL	rs61830764	1	212,289,976	A/G	0.36	0.022 (0.004)	5.6×10^{-8}	0.022 (0.004)	4.5×10^{-8}
TRIB1	rs6989280	8	126,508,746	G/A	0.70	0.022 (0.004)	2.2×10^{-7}	0.022 (0.004)	5.0×10^{-8}
MTNR1B	rs10830963	11	92,708,710	G/C	0.27	0.023 (0.004)	2.9×10^{-8}	0.022 (0.004)	1.0×10^{-7}
ABCC9	rs139975827	12	22,068,161	G/A	0.63	0.025 (0.004)	1.1×10^{-8}	0.022 (0.004)	1.0×10^{-7}

b



a, Effects (β values) were aligned to the BW-raising allele. Effect allele frequency (EAF) was obtained from the trans-ancestry meta-analysis, except for *PLAC1*, for which the EAF was obtained from the European ancestry meta-analysis due to lack of X chromosome data from the non-European studies. Chr, chromosome; bp, base pair; b37, build 37; EAF, effect allele frequency; SE, standard error. **b**, The effect of the lead SNP (absolute value of β, *y* axis) is given as a function of minor allele frequency (*x* axis) for 60 known (pink) and novel (green) BW loci from the trans-ancestry meta-analysis. Error bars are proportional to the standard error of the effect size. The dashed line indicates 80% power to detect association at genome-wide significance level for the sample size in trans-ancestry meta-analysis.

Extended Data Table 2 | Gene set enrichment analysis and protein–protein interaction (PPI) analysis

a. Gene set enrichment analysis

			95th percentile enrichment cutoff			75th percentile enrichment cutoff		
Database	**Gene set**	**Number of genes (mapped to MAGENTA)**	***P***	**FDR**	**Expected (observed) number of genes**	***P***	**FDR**	**Expected (observed) number of genes**
GOTERM	Positive regulation of glycogen biosynthetic process	10 (10)	5.6×10^{-5}	0.005	1 (5)	3.6×10^{-3}	0.18	3 (7)
GOTERM	Insulin-like growth factor receptor binding	13 (13)	2.4×10^{-5}	0.006	1 (6)	0.02	0.35	3 (7)
GOTERM	Positive regulation of glucose import	22 (22)	1.0×10^{-4}	0.019	1 (7)	0.02	0.36	6 (10)
GOTERM	Insulin receptor signalling pathway	35 (34)	2.8×10^{-5}	0.022	2 (9)	4.3×10^{-3}	0.27	9 (16)
GOTERM	Chromatin remodelling complex	11 (9)	9.0×10^{-4}	0.036	0 (4)	0.16	0.55	2 (4)
KEGG	Glycosphyngolipid biosynthesis globo-series	14 (13)	2.6×10^{-3}	0.037	1 (4)	0.21	0.48	3 (5)
KEGG	Melanoma	71 (67)	1.6×10^{-3}	0.037	3 (10)	0.05	0.35	17 (23)
KEGG	Terpenoid backbone biosynthesis	15 (15)	5.9×10^{-3}	0.039	1 (1)	0.15	0.44	4 (6)
KEGG	Type 2 Diabetes Mellitus	47 (45)	2.2×10^{-3}	0.040	2 (8)	0.14	0.46	11 (15)
Panther	Cholesterol biosynthesis	11 (11)	1.8×10^{-3}	0.040	1 (4)	0.29	0.64	3 (4)
BIOCARTA	Growth hormone pathway	28 (27)	3.0×10^{-4}	0.044	1 (7)	0.11	0.25	7 (10)
KEGG	Oocyte meiosis	114 (108)	1.0×10^{-3}	0.048	5 (14)	0.07	0.45	27 (34)
Custom gene set of imprinted genes								
GTEX	Imprinted genes (All)	77 (72)	1.9×10^{-4}	-	4 (12)	0.11	-	18 (23)
GTEX	Imprinted genes (Primary)	38 (35)	6.9×10^{-3}	-	2 (6)	0.14	-	9 (12)
GTEX	Imprinted genes (Primary + Suggestive)	55 (50)	0.010	-	3 (7)	0.25	-	13 (15)

b. Protein-protein interaction analysis

Database	**Pathway**	**Number of genes (overlapped with PPI network)**	***Z* score**	***P***	**adjusted *P*[a]**
GOTERM	Epidermal growth factor receptor signalling pathway	198 (31)	7.97	3.3×10^{-10}	1.4×10^{-7}
GOTERM	Insulin receptor signalling pathway	151 (26)	7.90	1.1×10^{-9}	2.9×10^{-7}
GOTERM	Stimulatory C-type lectin receptor signalling pathway	121 (22)	7.59	7.5×10^{-9}	1.2×10^{-6}
GOTERM	Negative regulation of canonical Wnt signalling pathway	152 (25)	7.46	6.2×10^{-9}	1.2×10^{-6}
GOTERM	Notch signalling pathway	129 (22)	7.21	2.6×10^{-8}	3.3×10^{-6}
GOTERM	Cellular response to insulin stimulus	71 (16)	7.62	3.7×10^{-8}	4.1×10^{-6}
GOTERM	Positive regulation of glycogen biosynthetic process	15 (8)	9.39	5.3×10^{-8}	5.1×10^{-6}
GOTERM	Positive regulation of protein phosphorylation	114 (20)	7.03	6.8×10^{-8}	5.9×10^{-6}
GOTERM	Positive regulation of glucose import	27 (10)	8.42	8.3×10^{-8}	6.5×10^{-6}
GOTERM	Fc-epsilon receptor signalling pathway	186 (26)	6.58	9.6×10^{-8}	6.8×10^{-6}

Two complementary analyses of the overall GWAS summary data identified enrichment of BW associations in biological pathways related to metabolism, growth and development. **a**, The top results (FDR < 0.05 at the 95th percentile enrichment threshold) from a total of 3,216 biological pathways tested for enrichment of multiple modest associations with BW. Additionally, results are shown for custom sets of imprinted genes: Primary, genes identified as highly likely to be imprinted in the GTEx database (tested $n = 38$); Primary + suggestive, genes identified as highly likely and suggestively imprinted in GTEx ($n = 55$); All, the above plus genes selected from the literature where imprinting status is consistent in GTEx ($n = 77$). **b**, The results of a complementary analysis of empirical PPI data, displaying the top 10 most significant pathways enriched for BW-association scores.

[a]*P* value is adjusted for multiple correction using the Benjamini–Hochberg method.