Benefits and limitations of genome-wide association studies

*Vivian Tam[1], Nikunj Patel[1], Michelle Turcotte[1], Yohan Bossé[2,3], Guillaume Paré[1,4] and David Meyre[1,4,5]**

Abstract | Genome-wide association studies (GWAS) involve testing genetic variants across the genomes of many individuals to identify genotype–phenotype associations. GWAS have revolutionized the field of complex disease genetics over the past decade, providing numerous compelling associations for human complex traits and diseases. Despite clear successes in identifying novel disease susceptibility genes and biological pathways and in translating these findings into clinical care, GWAS have not been without controversy. Prominent criticisms include concerns that GWAS will eventually implicate the entire genome in disease predisposition and that most association signals reflect variants and genes with no direct biological relevance to disease. In this Review, we comprehensively assess the benefits and limitations of GWAS in human populations and discuss the relevance of performing more GWAS.

Single-nucleotide variants (SNVs). Single nucleotides in the genome that vary between individuals in a population. Single-nucleotide polymorphism refers to a SNV that occurs at an appreciable frequency in a population (for example, >1%).

[1]*Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Ontario, Canada.*

[2]*Institut Universitaire de Cardiologie et de Pneumologie de Québec-Université Laval, Québec City, Québec, Canada.*

[3]*Department of Molecular Medicine, Laval University, Québec City, Quebec, Canada.*

[4]*Department of Pathology and Molecular Medicine, McMaster University, Hamilton, Ontario, Canada.*

[5]*Inserm UMRS 954 N-GERE (Nutrition–Genetics–Environmental Risks), University of Lorraine, Faculty of Medicine, Nancy, France.*

**e-mail: meyred@mcmaster.ca*

https://doi.org/10.1038/s41576-019-0127-1

Genome-wide association studies (GWAS), in which hundreds of thousands to millions of genetic variants across the genomes of many individuals are tested to identify genotype–phenotype associations (FIG. 1), have revolutionized the field of complex disease genetics over the past decade[1,2]. Since the first GWAS for age-related macular degeneration (AMD) was published in 2005 (REF.[3]), more than 50,000 associations of genome-wide significance ($P < 5 \times 10^{-8}$) have been reported between genetic variants and common diseases and traits[4]. These associations have led to insights into the architecture of disease susceptibility (through the identification of novel disease-causing genes and mechanisms) and to advances in clinical care (for example, the identification of new drug targets and disease biomarkers) and personalized medicine (for example, risk prediction and optimization of therapies based on genotype).

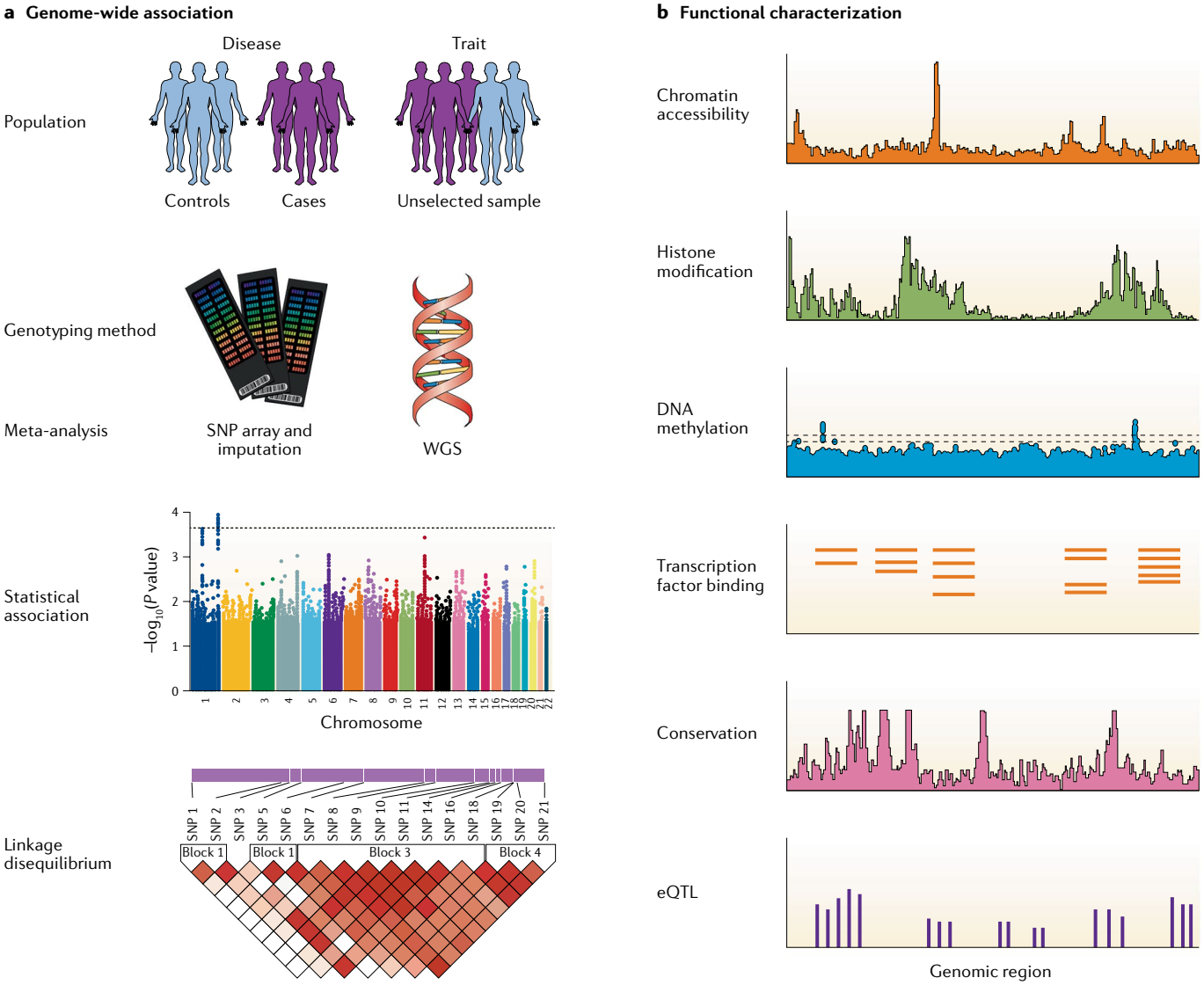
However, despite the clear successes of GWAS[5–7], this study design has not been without controversy[8–11]. Critics of GWAS have contended that single-nucleotide variants (SNVs) identified in GWAS explain only a small fraction of the heritability of complex traits[8], may represent spurious associations[9] and do not necessarily pinpoint causal variants and genes[10], and that GWAS will yield too many loci (if SNVs in all genes are implicated, then this would be uninformative)[10,11]. It has therefore been proposed to focus efforts on the analysis of ultra-rare variants and on post-GWAS experiments (for example, functional studies, gene network analysis and translational medicine)[9–11]. Unsurprisingly, these criticisms have led to scepticism among non-geneticists about the benefits of GWAS and hesitancy among national funding organizations to fund new GWAS[12]. In this context, it is timely for the scientific community, funding agencies and other stakeholders to consider the relevance of initiating more GWAS.

In this Review, we assess the benefits and limitations of performing GWAS in human populations. We draw primarily on lessons learned from the field of cardiometabolic diseases, given our expertise in this area, although we provide examples from other areas where applicable to highlight major advances and limitations of GWAS that might not have been observed prominently for cardiometabolic traits. Although a GWAS is a genome-wide analysis of genotypes that can be measured using numerous technologies — for example, whole-genome sequencing (WGS) or genome-wide single-nucleotide polymorphism (SNP) arrays plus imputation (BOX 1) — most GWAS are still performed using data from SNP arrays. Therefore, benefits and concerns that are relevant to GWAS as a genetic association study design in general and those that are unique to SNP array-based GWAS are discussed.

Benefits of GWAS

General benefits

GWAS have been very successful in identifying novel variant–trait associations. As of 11 January 2019, 3,730 GWAS have been published, with a total of 37,730 SNVs and 52,415 unique SNV–trait associations at a genome-wide significance threshold[4]. GWAS have successfully identified risk loci for a vast number of diseases and traits, including anorexia nervosa[13], major depressive disorder[14], cancers and subtypes of cancers[15,16], type 2



Heritability
The proportion of phenotypic variation between individuals in a population that is due to genetic factors.

diabetes mellitus (T2DM)[17], coronary artery disease[18], schizophrenia[19], inflammatory bowel disease[20], insomnia[21], body mass index (BMI)[22] and educational attainment[23], among others. This surge of replicable associations is in stark contrast to the pre-GWAS era, in which only a handful of robustly associated loci were identified[24].

Initial excitement was somewhat tempered by the realization that GWAS loci typically have small effect sizes and explain only a modest proportion of trait

◀ Fig. 1 | **GWAS study design. a** | The aim of a genome-wide association study (GWAS) is to detect associations between allele or genotype frequency and trait status. The first step of a GWAS involves identifying the disease or trait to be studied and selecting an appropriate study population (for example, cases and controls for a disease, or an unselected population sample for a trait). Genotyping can be performed using single-nucleotide polymorphism (SNP) arrays combined with imputation or whole-genome sequencing (WGS). Association tests are used to identify regions of the genome associated with the phenotype of interest at genome-wide significance, and meta-analysis is a common step to increase the statistical power to detect associations. Causal variants are usually not directly genotyped but are in linkage disequilibrium with the genotyped SNPs. **b** | Functional characterization of genetic variants is often required to move from statistical association to causal variants and genes, especially in the non-coding genome. Computational methods are used to predict the regulatory effect of non-coding variants on the basis of functional annotations. **c** | Target genes can be identified or confirmed using chromatin immunoprecipitation and chromosome conformation capture methods, and experimentally validated using cell-based systems and model organisms. **d** | Genetic variants exist along a spectrum of allele frequencies and effect sizes. Most risk variants identified by GWAS lie within the two diagonal lines. Rare variants with small effect sizes are difficult to identify using GWAS, and common variants with large effects are unusual for common complex diseases. eQTL, expression quantitative trait locus.

heritability[8]. However, the gap between the amount of heritability explained and the amount estimated by twin and family studies is now better understood. For many traits, SNPs are suggested to account for a majority of the 'missing' heritability, such that the 'missing' heritability is small, especially if heritability estimates are biased upwards[25–28]. Thus, even if GWAS cannot explain all the heritability of complex traits, they represent a practical means by which bona fide associations can be discovered, and increasing the sample size of GWAS should continue to yield new loci. In fact, empirical evidence demonstrates that for each complex trait there is a threshold sample size above which the rate of locus discovery accelerates in GWAS[1,29] (FIG. 2), and, to date, the identification of risk loci has yet to plateau for any trait[1,30].

GWAS can lead to the discovery of novel biological mechanisms. GWAS loci often implicate genes of unknown function or of previously unsuspected relevance[5], and experimental follow-up of such loci can lead to the discovery of novel biological mechanisms underlying disease[2,5]. For example, the role of autophagy in Crohn's disease was not known before the discovery of SNPs associated with disease risk in the genes *ATG16L1* (rs2241880)[31] and *IRGM* (rs1000113)[32] in GWAS. The rs2241880 SNP leads to a missense mutation (p.Thr300Ala) that results in enhanced caspase-3-mediated cleavage of ATG16L1 and diminished autophagy under cellular stress[33]; in turn, this reduction impairs intracellular bacterial clearance and increases inflammatory cytokine production, thereby establishing a chronic inflammatory state[33]. A similar effect on autophagy at the *IRGM* locus, mediated by a causal SNP in strong linkage disequilibrium with rs1000113, has also been reported[34]. Another example involves a risk haplotype spanning *SLC16A11* on chromosome 17p13 that was associated with T2DM in a GWAS of Mexican adults[35]. Genetic variants at this locus were shown to independently reduce SLC16A11 function in two ways: first, by decreasing *SLC16A11* expression in the liver in a gene dose-dependent manner; and second, by disrupting a key interaction with a chaperone protein, thereby reducing cell-surface localization of SLC16A11 (REF.[36]). Additional experiments demonstrated that SLC16A11 is a proton-coupled monocarboxylate transporter and that decreased SLC16A11 induces changes in cellular fatty acid and lipid metabolism that are associated with increased risk of T2DM[35,36]. Other well-known illustrations of quick translation from GWAS to biology include an association between the *CFH* gene and AMD, which implicates the complement system of innate immunity[3], and between the major histocompatibility complex locus and schizophrenia, which points to a role for complement component 4 activity[37,38].

Of note, the value of biological insights gained from GWAS is not proportional to the strength of association. For example, many genes that represent molecular targets of US Food and Drug Administration-approved drugs harbour common variants of modest effect that have been identified by GWAS[5]. This finding provides a strong argument for continuing to identify subtle associations using GWAS in even larger sample sizes[39].

GWAS findings have diverse clinical applications. A central objective of genetic research is to translate biological insights into medical advancements. Despite the considerable amount of time required to bring scientific discoveries from bench to bedside[40], a growing number of examples highlight the diverse areas in which GWAS findings can have clinical applications.

Genetic variants discovered by GWAS can be used to identify individuals at high risk of certain diseases, thereby improving patient outcomes through early detection, prevention or treatment. For example, a coding non-synonymous variant in the *CFH* gene (rs1061170) explains 50% of the population-attributable risk of AMD[3,41]. Multi-locus analysis of AMD susceptibility loci showed that 99% of the individuals with the highest-risk genotypes (including at *CFH*) had AMD; of these, 85% had advanced AMD[42]. Similarly, a GWAS for exfoliation glaucoma identified two non-synonymous SNPs in the gene *LOXL1* that explain 99% of the population-attributable risk of this disease[43].

GWAS findings can be applied to disease classification and subtyping. Maturity-onset diabetes of the young (MODY) accounts for 1–2% of all patients with diabetes but is commonly misdiagnosed as type 1 diabetes mellitus (T1DM) or T2DM[44]. Rare loss-of-function mutations in *HNF1A* are known to cause a common form of MODY[45]. In 2008, two independent GWAS identified SNPs near the *HNF1A* gene associated with serum C-reactive protein (CRP) levels, a marker of inflammation[46,47]. On the basis of these findings, it was hypothesized that serum levels of high-sensitivity CRP could represent a clinically useful biomarker to identify *HNF1A* mutations that cause MODY[48]. Patients with *HNF1A*-MODY were observed to have lower CRP levels than individuals without diabetes as well as individuals with T1DM, T2DM or non-*HNF1A* MODY, validating the use of high-sensitivity CRP as a clinical biomarker for diagnosing certain diabetes subtypes[48,49].

Rare variants
Variations in the genome for which the less prevalent form (minor allele) occurs at a frequency of 1% or less in the population.

Imputation
Statistical inference of unobserved genotypes from a reference panel of known haplotypes in a population.

Effect sizes
The magnitudes of the effect of alleles on phenotypic values.

Linkage disequilibrium
The nonrandom association of alleles at two or more loci due to limited recombination.

Haplotype
A set of genetic markers that are present on a single chromosome and in linkage disequilibrium.

Common variants
Variation in the genome for which the less prevalent form (minor allele) occurs at a frequency of 5% or greater in the population.

Box 1 | **GWAS using SNP arrays versus WGS**

The genome-wide association study (GWAS) is a study design used to detect associations between genetic variants and common diseases or traits in a population. Genetic variants can be genotyped using numerous technologies, including genome-wide single-nucleotide polymorphism (SNP) arrays (combined with statistical imputation of unobserved genotypes from population reference panels) and whole-genome sequencing (WGS). SNP arrays are the most widely used genotyping technology in GWAS, primarily owing to their lower costs, and performing WGS in very large sample sizes is currently cost-prohibitive. Although the switch to WGS is likely to be inevitable with declining sequencing costs, the choice to use SNP arrays or WGS in GWAS should be made taking into consideration other factors, such as the reliability of the technology, desired coverage of genetic variants, available resources and the study design and research objectives (see the table).

Factor	SNP arrays	WGS
Cost	Relatively inexpensive (~US$40 per sample)	Expensive (>US$1,000 per sample)
Reliability	Reliable, highly accurate technology	Less mature and less accurate technology
Genomic coverage	• Mainly restricted to common and low-frequency variants, although imputation of rare variants is increasingly accurate (ultra-rare variants, however, can never be identified) • Biased towards variants discovered in well-studied or sequenced populations	From low-frequency, common variants to nearly all genetic variation in the genome, depending on the depth of sequencing
GWAS analysis	Well-established analytical pipeline and tools for data analysis	• Higher computational costs and greater analytical complexity • Eventually, larger multiple testing burden when conducting single-variant tests
Other considerations	Custom genotyping arrays can be extremely cost-effective	• As all variation is ascertained, fine-mapping is easier • Greater costs to store, process, analyse and interpret the resulting data
Suitable research objectives	• Analysing known or candidate associations in large cohorts • Detecting low-frequency, common variant associations in extremely large sample sizes	• Detecting and fine-mapping rare variants • Detecting ultra-rare risk variants when it becomes economically viable to perform WGS at a very large scale

GWAS can inform drug development and repurposing. Although GWAS are not usually directly informative with respect to the causal genes or the disease mechanisms, post-GWAS functional experiments can illuminate new targets and pathways for therapeutic intervention. Because the percentage of drug mechanisms with direct support from human genetic studies increases across the drug development pipeline, selecting genetically supported drug targets could improve the success rate of drug development, reducing the time and costs of developing new drugs[50]. GWAS for several diseases (including T2DM, rheumatoid arthritis, ankylosing spondylitis, psoriasis, osteoporosis, schizophrenia and dyslipidaemia) have led to the identification of new and candidate drugs that are now being used in clinics or evaluated in clinical trials[2].

GWAS have identified genetic variants that can be used to inform drug selection and dosage and prevent adverse drug reactions[51]. Notable examples include SNPs near the *IL28B* gene (also known as *IFNL3*) that predict the likelihood of a positive response to pegylated interferon-α and ribavirin therapy for hepatitis C virus infection[52–54] and genetic variants in *SLCO1B1* that are associated with simvastatin-induced myopathy[55]. The Clinical Pharmacogenetics Implementation Consortium (CPIC) has established a scientifically rigorous approach to determining the clinical value and interpretation of genetic variants associated with drug response, aiding physicians in making prescription decisions. For example, CPIC guidelines have been published for pegylated interferon-α-based treatment regimens based on the *IL28B* genotype[56] and for managing the risk of simvastatin-induced myopathy in the context of *SLCO1B1* genotyping[57].

GWAS can provide insight into ethnic variation of complex traits. Although common variants are expected to be evolutionarily old and shared across ethnicities, some risk loci show considerable ethnic differences in frequency and/or effect size[2,58]. Performing GWAS in diverse ethnic groups can therefore reveal heterogeneity in genetic susceptibility to disease. For instance, GWAS have identified different genetic loci as having the strongest effect on T2DM risk in European (*TCF7L2*)[59], East Asian (*KCNQ1*)[60,61], Mexican (*SLC16A11*)[35] and Greenlandic (*TBC1D4*)[62] populations. A locus that is associated with disease in one ethnic group but not in another may indicate ethnic differences in risk allele frequency. For example, risk alleles in *KCNQ1* confer increased susceptibility to T2DM in both East Asians

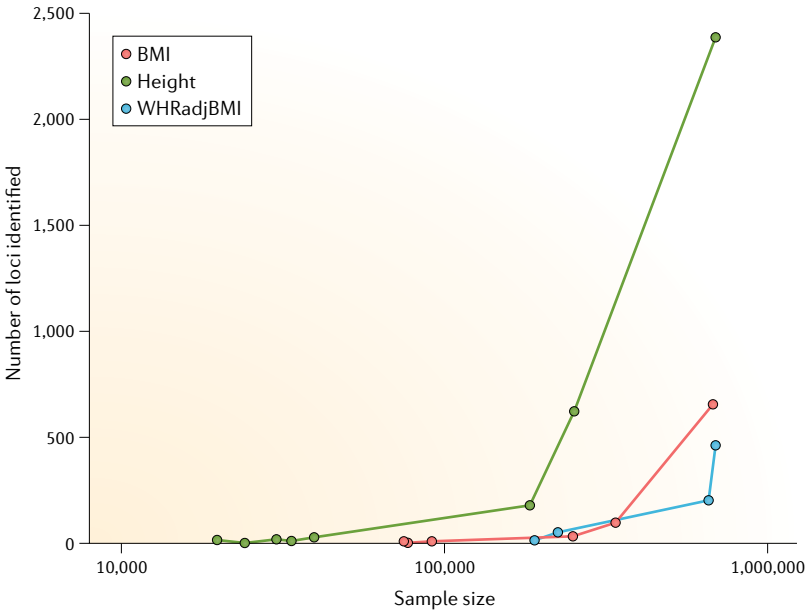


Fig. 2 | **Number of loci identified as a function of GWAS sample size.** A plot of the number of independent or near-independent genome-wide significant loci ($P < 5 \times 10^{-8}$) reported from genome-wide association studies (GWAS) in European or predominantly European populations for three anthropometric traits: body mass index (BMI)[22,103,104,182,333,334], height[22,335–341] and waist-to-hip ratio adjusted for BMI (WHRadjBMI)[167,342–344]. For each trait, there is a threshold sample size above which the rate of locus discovery accelerates. The identification of risk loci has yet to plateau for these traits.

and Europeans, but because the minor alleles are less common in Europeans, much larger sample sizes are needed to detect the association[61]. Similarly, the 15q25 locus has been unequivocally associated with lung cancer in European populations, but the association has not been replicated in Asian populations because of very low risk allele frequencies[63]. In more extreme cases, risk variants can be common in one population but virtually absent in others (founder mutations), as observed in the Greenlandic and Samoan isolated populations[62,64]. Ethnic differences in the effect size of a risk variant[65–67] can also affect the likelihood of its discovery and its contribution to disease burden across populations[68–70].

GWAS are relevant to the study of low-frequency and rare variants. Currently, most GWAS are performed using data obtained by SNP arrays. Genome-wide SNP arrays were originally designed to interrogate common genetic variation, but they have improved markedly over time to include a greater density of variants and a wider range of allele frequencies. Content on SNP arrays is now informed by data from large reference panels (TABLE 1), such as the Haplotype Reference Consortium panel[71], enabling many low-frequency and rare variants to be directly genotyped. Low-frequency and rare variants can also be genotyped using exome-centred custom arrays, which target ~240,000 rare, low-frequency and common coding variants observed in sequencing data from ~12,000 individuals[72]. These exonic variants can be added to larger arrays with a genome-wide scaffold of common and rare variants. Studies using exome-centred custom arrays have identified rare and low-frequency coding variants associated with various complex traits, including blood lipid levels[73], haematological traits[74–76], blood pressure level[77], height[78], BMI[70] and T2DM[79].

Genotype imputation is commonly done in array-based GWAS to increase the number of variants that can be tested for association, including low-frequency and rare variants that were not originally genotyped. It predicts the genotypes of untyped variants from a dense panel of reference haplotypes, derived ideally from sequencing a subset of the study population or a closely related one[80]. The size of a reference panel (that is, the number of haplotypes) is directly and inversely related to the imputable allele frequency. Therefore, with the availability of larger and more ethnically diverse reference panels (for example, from the 1000 Genomes Project)[81] and the generation of population-specific and study-specific reference panels[82–84], untyped variants can be accurately imputed at low minor allele frequencies (MAFs), provided that they are first observed in the reference population. Recently, the Haplotype Reference Consortium created a unified reference panel of 64,976 haplotypes by amalgamating WGS data from 20 studies of individuals of primarily European ancestry[71]. This reference panel claims accurate imputation down to a MAF of 0.1%[71]. Ongoing large-scale projects, such as the Trans-Omics for Precision Medicine (TOPMed) programme, are expected to produce reference panels of more than 100,000 individuals[85].

As WGS becomes cheaper and more widespread, the study of rare variants should become more tractable using the GWAS approach. GWAS based on high-depth WGS permit the full frequency spectrum of variants to be studied, including rare variants that are difficult to capture using SNP arrays and imputation.

GWAS can be used to identify novel monogenic and oligogenic disease genes. Up to one-fifth of loci identified by GWAS include a gene that is mutated in a corresponding single-gene disorder[5]. This phenomenon is usually due to the accumulation of independent rare and common causal variants in the same gene, although, in rare cases, partial linkage disequilibrium between rare pathogenic mutations and common non-causal SNPs can create a synthetic GWAS hit[86,87]. This observation led scientists to hypothesize that genes identified by GWAS may be relevant candidates for discovering rare disease-causing mutations[7,39]. Studies re-sequencing GWAS-implicated genes have since validated this hypothesis, identifying many novel monogenic or oligogenic genes for complex diseases, including for obesity (*SH2B1*, *NPC1* and *ADCY3*)[88–91], T1DM (*IFIH1*)[92], T2DM (*MTNR1B*, *SLC30A8* and *PPARG*)[93–95] and inflammatory bowel disease (*TNFRSF6B*, *PRDM1*, *CARD9*, *IL23R* and *RNF186*)[96–98], and confirming the value of targeting GWAS loci to efficiently identify genes for corresponding monogenic or familial forms of disease. The main advantages of targeting GWAS-identified genes for re-sequencing are lower costs and improved statistical power to detect associations compared with whole-exome sequencing (WES) or WGS[99,100].

Minor allele frequencies (MAFs). The frequencies of the less common allele of a genetic variant in a population.

Table 1 | **Commonly used population reference panels**

Reference panel	Number of reference samples	Ancestry of reference samples	Number of variant sites	Indels available	Refs
Icelandic reference panel	15,220	European (Icelandic)	31.1 million	Yes	345
HapMap Project phase 3	1,011	Multi-ethnic	1.4 million	No	346
1000G phase 1	1,092	Multi-ethnic	28.9 million	Yes	347
1000G phase 3	2,504	Multi-ethnic	81.7 million	Yes	81
UK10K Project	3,781	European	42.0 million	Yes	348
HRC	32,470	Predominantly European (includes the 1000G reference panel samples)	40.4 million	No	71
TOPMed[a]	62,784	Multi-ethnic	463.0 million	Yes	85

1000G, 1000 Genomes; HRC, Haplotype Reference Consortium; indels, insertions or deletions; TOPMed, Trans-Omics for Precision Medicine. [a]Figures are based on the latest status of the reference panel.

GWAS can study genetic variants other than SNVs. GWAS are primarily designed to test SNVs for association with complex diseases and traits. However, other types of genetic variants that contribute to disease susceptibility can also be detected by GWAS. For example, GWAS have associated rare[101,102] and common copy number variants (CNVs)[103–105] with BMI and obesity, among several other common traits and diseases. However, to date, there has been a paucity of GWAS investigating the effects of CNVs on disease risk. This area of untapped research has an important impact on conditions such as autism, bipolar disorder and schizophrenia, in which CNVs are known to play a particularly prominent role[106,107].

Furthermore, analyses of SNP array data from GWAS have identified clonal mosaicism in autosomes and sex chromosomes, which is associated with ageing and increased risk of haematological and solid tumours[108–110], as well as diseases such as T2DM[111]. These studies have revealed that somatic mosaicism is more common than initially thought, especially in the ageing genome, spurring extensive research into genomic stability and clonal haematopoiesis as a predictor of haematological cancer[112].

Beyond CNVs and mosaic events, other classes of genetic variation, including haplotypes[113], variable number tandem repeats[114], retrotransposon insertion polymorphisms[115], insertions or deletions (indels)[116] and inversions[117], have been associated with risk of disease in GWAS.

GWAS data are used for multiple applications beyond gene identification. The value of GWAS lies not only in their utility in identifying loci influencing disease predisposition but also in numerous applications for which GWAS data may be used. Beyond gene identification, GWAS data — in the form of either individual-level genotype data or summary-level association statistics — have enabled a wide range of applications, including reconstruction of population history[118–121], determination of ancestry and population substructure[122,123], fine-scale estimation of location of birth[124], genome-wide assessment of linkage disequilibrium[125], estimation of SNP heritability for complex traits[26], estimation of genetic correlations between traits[126], Mendelian randomization studies[127], polygenic risk scores[128], forensic analyses[129,130], determination of cryptic relatedness[131], paternity testing[132], direct-to-consumer genetic testing[133], clinical diagnostic genetic testing[133], prenatal and pre-implantation genetic diagnosis[134,135], embryonic DNA fingerprinting[136], determination of perinatal loss[137], loss-of-heterozygosity and CNV analyses in tumours (for example, for disease subtyping and classification)[138], validation of new analytic methods[139] and quality control of next-generation sequencing data (by comparing sequence-based variant calls to array-based genotyping)[140].

A particularly noteworthy area has been in estimating SNP heritability and modelling disease genetic architecture. Methods for estimating SNP heritability have helped to define the empirical bounds of GWAS[26,141], whereas modelling the underlying genetic architecture of disease has shown that diseases have diverse genetic architectures, with consequent effects on rates of loci discovery[142]. Psychiatric diseases and mental health traits, for example, seem to be mostly polygenic, involving a continuum of variants with small effects[142]. By contrast, most other diseases involve clusters of SNPs with distinct magnitudes of effects[142]. Sample sizes that are needed to explain most of the heritability of these traits range accordingly from a few hundred thousand to millions of individuals[142]. Discovery rates based on allele frequencies, effect sizes and accuracy of phenotyping have been instrumental in informing the design of future GWAS and in predicting the extent of their discovery[142,143].

GWAS data generation, management and analysis are straightforward. The success of GWAS can be attributed in part to technological and methodological advances that have facilitated their performance. For data generation, several algorithms for calling genotypes from SNP array data have been developed, with each generation heralding improvements in accuracy and call rate[32,144,145]. New algorithms have also been designed specifically for calling low-frequency and rare variants[146–149] and for inferring haplotypes and structural variants[114,150–152]. With GWAS increasingly relying on WGS data, a corresponding host of tools has been developed for variant discovery and SNP calling from sequencing data[153]. Similarly, improvements have been made in statistical imputation of genotypes[154]. For data management and

Copy number variants
(CNVs). A class of DNA sequence variants (including deletions and duplications) that lead to a departure from the expected diploid representation of DNA sequence.

Clonal mosaicism
The presence of clones of cells with different karyotypes within an individual derived from a single zygote.

analysis, software such as PLINK can handle and analyse whole-genome SNP array data in a computationally efficient manner[155,156]. Currently, BOLT-LMM[157] and SNPTEST[158] are popular for GWAS analysis, and Hail, a recently developed scalable framework for genomic data analysis, is gaining in popularity. In addition, several tools are available to visualize GWAS results (for example, qqman[159] and LocusZoom[160]) and to conduct GWAS meta-analyses (for example, METAL[161]).

Consensus among the genetics community to adopt a standardized significance threshold ($P < 5 \times 10^{-8}$)[162] has allowed the field to enjoy highly reproducible findings, and this threshold is likely to become more stringent with the increasing number of low MAF SNPs being analysed in higher-density arrays and/or more comprehensive imputation reference panels[163].

GWAS data are easily shared and publicly available data facilitates novel discoveries. GWAS data will continue to be useful for identifying novel trait associations for some time, as the value of the vast amount of publicly available GWAS data that has accumulated has been realized only partially. Several initiatives are expected to deliver genome-wide genotype and rich phenotypic information on a record number of individuals. These data offer the opportunity to aggregate very large sample sizes, from which novel associations can be discovered.

The availability of GWAS summary statistics has increased dramatically in recent years, and hundreds of such data sets are now publicly available[2]. This wealth of data can be analysed in various ways to gain insight into the genetic basis of complex traits[164]. For example, publicly available summary statistics have enabled discoveries of novel risk loci[22,165–167], estimates of SNP-based heritability[168,169], cross-trait analyses[126,170], polygenic risk prediction[171] and fine-mapping[172], among other applications. Despite considerable progress, there is still a need for improved sharing of summary statistics, including summary statistics with linkage disequilibrium information, given the importance of these data in providing a foundation for a wide range of important analyses[164]. A commitment by researchers to make summary association statistics publicly available will have a broad impact on genomics research. Funding agencies and journals should have a stronger role in enforcing data-sharing requirements.

Growth in resources that link electronic health record data to genotype data has also been observed in recent years (for example, UK Biobank; Kaiser Permanente's Research Program on Genes, Environment, and Health; and the Electronic Medical Records and Genomics Network)[173]. The density and phenotypic diversity of longitudinal clinical data afforded by electronic health records permit a deeper understanding of genotype–phenotype associations[173,174]. Furthermore, large-scale initiatives (for example, national biobanks and participant-centric initiatives) in both the private and public sectors have collected, and in some cases are still collecting, genotype and phenotype information on a large number of participants, and such resources have been particularly transformative for the discovery of novel trait associations. For example, UK Biobank recently released genome-wide genotypes and rich phenotypic data on ~500,000 individuals[175]. In the private sector, similar successes can be found. The company 23andMe, which offers direct-to-consumer genetic services, has amassed genotype and phenotype data on millions of individuals. These data have been combined with other sources of GWAS data in consortium-led studies to identify multiple risk loci for many complex diseases and traits, including most recently educational attainment[23], impulsivity[176] and neuroticism[177].

Combining different sets of genetic data and more open data sharing can provide a new paradigm for discovering novel associations. Integrated knowledge databases that are accessible through interactive public portals, such as one envisioned for T2DM[178], might allow new types of data (for example, clinical trial data on patient drug responses) to be integrated with GWAS for the first time. Such databases would enhance the value of GWAS by creating a synergistic link between data contributors and researchers[178].

GWAS findings published to date represent only the tip of the iceberg. Complex diseases result from the interplay between biological and environmental factors. For example, obesity arises from complex interactions between genetic predisposition, demographic factors (for example, age), medical conditions (for example, depression), lifestyle factors (for example, sedentary lifestyle, unhealthy dietary patterns, smoking cessation and medication or drug use) and environmental exposures (for example, pollution and built environment)[179,180]. Such gene–environment interactions have support at the molecular level. For example, studies have shown that environmental factors can alter methylation patterns of obesity genes, and the plasticity of the methylome supports the notion that different subsets of genes predispose individuals to obesity in response to specific environmental exposures[181].

Despite unequivocal evidence of the interplay between environment and genetics in mediating disease risk, most GWAS performed to date have focused only on easy-to-measure phenotypes, such as BMI, for example, without accounting for relevant biological and environmental exposures[182]. Existing GWAS findings therefore represent the low-hanging fruit of GWAS discoveries[183], and exploring a wider range of phenotypes in GWAS is likely to lead to additional discoveries. For obesity, these include BMI in different age groups[184,185], BMI changes over time (that is, longitudinal data)[186], BMI changes in response to obesity risk[187,188] or obesity-protective exposures[189,190], BMI variance[191], the extreme tails of BMI distribution[192–194], deep obesity phenotypes[183,195,196], obesity intermediate traits[197–199] and biomarkers[200–202], and obesity composite traits[203]. A similar extension of the phenotypes studied in GWAS of other diseases and traits is likely to result in the identification of corresponding novel loci.

To this end, large prospective cohort studies with longitudinally measured clinical, demographic, lifestyle and environmental exposure data are needed, as are electronic health records and other sources of real-world evidence that provide a treasure trove of

Fine-mapping
The process of localizing association signals to causal variants using statistical, bioinformatic or functional methods.

additional information (for example, laboratory test results, physician notes and health administrative data). In particular, integrating behavioural health-tracking data with genetic data could contribute enormously to our understanding of neuropsychiatric disorders, which currently lack large cohorts phenotyped for quantitative behavioural traits[204]. Longitudinal data might also enable the discovery of prognostic loci, which are valuable for understanding the progression of diseases such as cancer and Parkinson disease.

In addition to extending the phenotypes studied in GWAS, a similar expansion in scale at multiple levels (sample size, populations studied, methods and study design used)[205] can help to reveal more of the 'GWAS discoveries' iceberg (FIG. 3). First, as sample size is the primary limiting factor in risk variant discovery, larger sample sizes will necessarily result in the identification of additional loci. Sample sizes of over 1 million individuals are now becoming a reality for some traits, especially with the increasing public availability of summary statistics from large-consortia GWAS and the establishment of large-scale initiatives collecting genotype or sequencing data and clinical information[21,23]. For rare diseases and conditions, for example, T1DM or suicidal behaviours, the use of electronic health records to identify affected individuals at a national level might be promising.

Second, performing large GWAS in understudied ethnic groups will be informative, especially to uncover ethnic-specific risk variants. Several ethnic groups have been neglected or disproportionately underrepresented in genetic association studies to date; these groups include Latin Americans, Native Americans, Indigenous Australians, Arabs, South Asians, Roma and Pacific Islanders[205]. The study of isolated, that is, founder[64] and highly consanguineous populations[206], as well as multi-ethnic and admixed groups will also be valuable[207–209].

The rest of the 'iceberg' may be uncovered by using innovative GWAS methods and study designs. GWAS analyses are usually performed under an autosomal additive model, which is quite restrictive. The routine

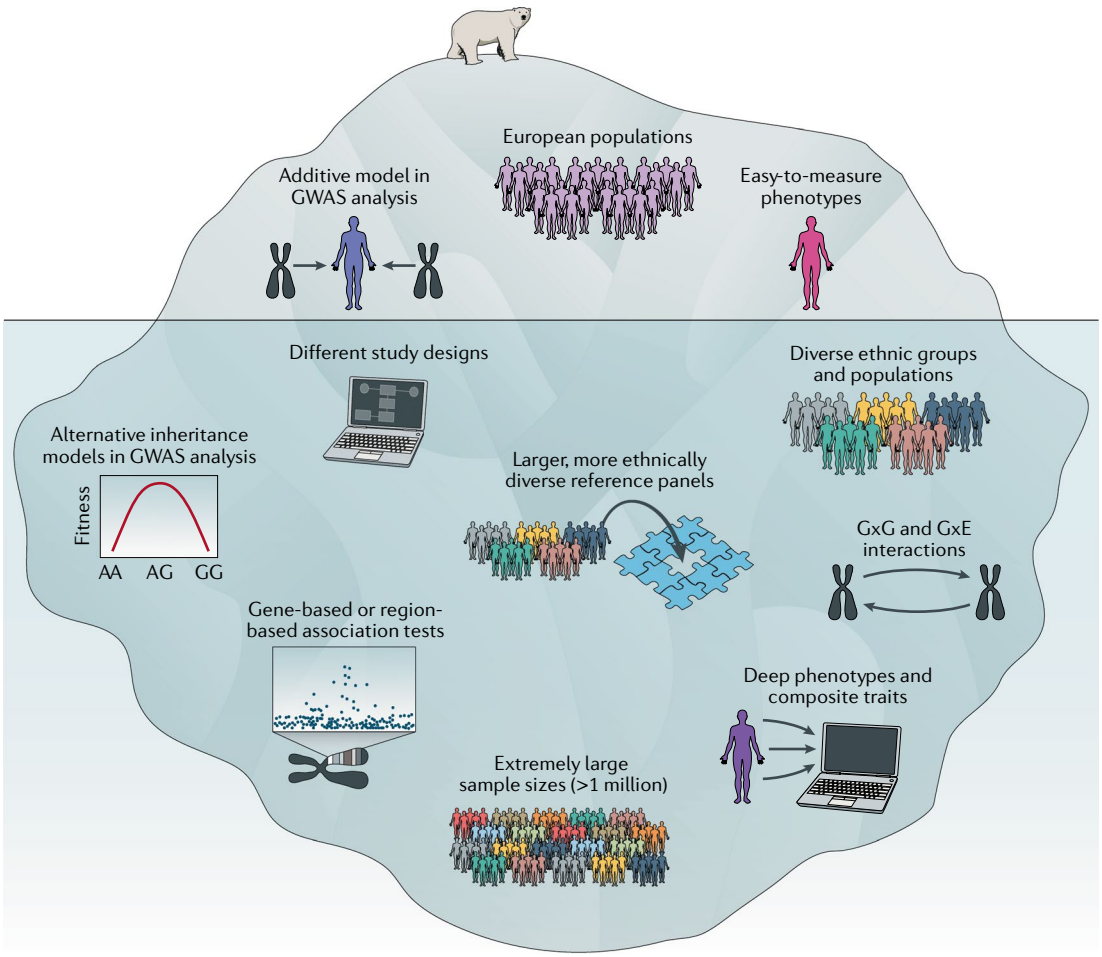


Fig. 3 | **GWAS performed to date represent the tip of the iceberg.** The discoveries that can be made using genome-wide association studies (GWAS) are represented by an iceberg. The portion of the iceberg above water represents the discoveries that have been made by GWAS to date, using easy-to-measure phenotypes, predominantly European populations, and an additive genetic model. Most of the iceberg is submerged under water. The submerged portion represents the vast number of discoveries that can potentially be made by expanding the current paradigm of GWAS to include a wider range of phenotypes, substantially larger sample sizes, more diverse populations and ethnic groups, and different study designs and analyses. GxG, gene–gene; GxE, gene–environment.

use of recessive[210], dominant[192], overdominant[211], multiplicative[212], parent-of-origin-specific[213] and X-linked inheritance[214] models in GWAS can improve the statistical power to detect additional variants[215]. In addition, GWAS accounting for gene–gene and gene–environment interactions[216,217], GWAS using genomic region-based or gene-based association tests[218,219], GWAS using Bayesian analyses[220], GWAS using machine learning approaches[221], GWAS in different study designs (for example, family-based, case–control–case, case-only, intervention and hypothesis-driven study designs)[222] and GWAS using methods to improve power to analyse heterogeneous traits (and investigate overlap between traits, such as cancer and cancer subtypes), such as ASSET[223], for example, can also lead to the discovery of previously undetected associations.

Benefits specific to SNP array-based GWAS

GWAS based on SNP arrays use reliable genotyping technology. Accurate genotyping is crucial to the success of any large-scale genetic association study, as systemic biases induced by even small sources of error may inflate the number of false-positive and false-negative associations[224]. Genotyping using SNP arrays has become extremely efficient and reliable with the extensive number of GWAS performed over the past decade. According to company specifications, contemporary genome-wide SNP arrays achieve call rates, HapMap concordance, Mendelian consistency and reproducibility of >99.7%. In addition, best practices for SNP quality control have been developed for GWAS[224] and meta-analyses of GWAS[225]. SNPs that fail to meet acceptable quality control thresholds, which are usually set independently for each study (for example, a SNP call rate of >97%, a duplicate concordance rate of >99%, Mendelian consistency of >99% and Hardy–Weinberg equilibrium test $P > 1 \times 10^{-6}$), are removed before data analysis, thus ensuring confidence in GWAS results.

GWAS based on SNP arrays are cost-effective for identifying risk loci. GWAS are a cost-effective approach for identifying risk loci, given that SNP array analysis followed by imputation of variants down to a MAF of 0.1% is now very affordable, allowing for much of the genetic variation in the genome to be explored at a reasonable cost even in very large sample sizes. At the time of writing, genome-wide SNP arrays, such as the Illumina Infinium Global Screening Array and the Thermo Fisher Axiom Precision Medicine Research Array, cost approximately US$40 per sample. As the cost of WGS continues to decline, GWAS using SNP arrays will eventually be replaced by GWAS using WGS[2]. Currently, the price differential between the two technologies is at least 30-fold, and ancillary costs for performing GWAS using WGS, such as those required for data storage and processing, computational infrastructure and research staff, remain prohibitively high[226–228]. Until then, however, we are optimistic that the majority of the common variants and a substantial fraction of the low-frequency and rare variants that contribute to disease risk can be identified using affordable SNP arrays combined with imputation to increasingly large WGS reference panels.

Limitations of GWAS

General limitations

GWAS are penalized by an important multiple testing burden. A major limitation of genome-wide approaches is the need to adopt a high level of significance to account for the multiple tests. This is commonly done in GWAS by using a Bonferroni correction to maintain the genome-wide false-positive rate at 5%, based on the assumption of 1 million independent tests for common genetic variation. As a result, conventional GWAS are underpowered to detect all the heritability explained by SNPs, because association signals must reach a threshold of $P < 5 \times 10^{-8}$ to be considered significant[8,229]. The limitation of multiple testing is likely to be exacerbated in the future, as genomic coverage of GWAS increases in parallel with the use of WGS data and the number of independent tests becomes much larger[163].

One strategy to overcome the limitation of multiple testing in GWAS is to increase the sample size. This approach has been successfully used by large international consortia to study traits that are available in multiple cohorts and are relatively inexpensive to measure[7]. However, assembling large sample sizes is not always possible. For example, certain 'deep' phenotypes (for example, weight/body-composition change phenotypes) are costly and difficult to measure, and only a limited number of suitably characterized individuals may be gathered for such traits[7]. A similar challenge arises for GWAS in small isolated populations[230].

A second strategy is to reduce the number of tests performed. This can be achieved by using gene-based[231] or pathway-based association tests[232], or by restricting analyses to candidate genomic regions, such as linkage regions[233], genes specifically expressed in an important tissue[234], genes showing differential expression patterns related to the disease[235], prioritized candidate genes[192], potentially damaging SNPs[70] or SNPs harbouring evolutionary signatures[236]. Combining supporting biological evidence with statistical significance also increases the probability that the result is a true positive[237]. However, caution should be exercised when interpreting findings from underpowered studies that do not account for all explicit and implicit hypothesis testing. A more rigorous approach to dealing with multiple hypothesis testing is to use Bayesian methods[220]. Unlike frequentist methods, such as the Bonferroni correction, Bayesian methods consider the universe of possible hypotheses in the human genome[220]. A statistical advantage is offered when the number of possible tests is reduced, for example, by assaying biological units such as genes instead of single variants.

GWAS explain only a modest fraction of the missing heritability. GWAS have identified an unprecedented number of genetic variants associated with common diseases and traits, but, apart from a few notable exceptions (for example, AMD, exfoliation glaucoma and T1DM), these variants account for only a modest proportion of the estimated heritability of most complex traits[8,238]. Several reasons for the missing heritability have been proposed[8,239,240]. One probable explanation is that SNPs of modest effect are missed because they do not reach the

stringent significance threshold[8,25,239]. In line with this hypothesis, recent genome-wide complex trait analyses suggest that SNPs may explain one-third to two-thirds of the heritability of most complex traits[25–28]. With increasingly large sample sizes[2,22], as well as the adoption of new methods and study designs[220–222,241], GWAS findings may soon account for a substantial fraction of the heritability for many complex diseases.

Some have used the argument of missing heritability to suggest a failure of GWAS to explain the genetic underpinnings of disease. However, it must be noted that genetic susceptibility to disease should be studied in the context of environmental risk factors, with which it is inextricably linked. Gene–gene and gene–environment interactions are expected to explain some of the missing heritability[242,243]. This might be especially true of diseases for which susceptibility is highly influenced by the environment. For example, different subsets of genes are expected to play a role in obesity depending on the risk environment (for example, nicotine withdrawal, pregnancy and antidepressant medication)[181,244]. In addition, heritability estimates may be inflated as a result of shared environmental effects, especially in classical twin studies[245–247].

GWAS do not necessarily pinpoint causal variants and genes. Genetic mapping is a double-edged sword: local correlation of multiple genetic variants due to linkage disequilibrium facilitates the initial identification of a locus but makes it difficult to discern the causal variant or variants[39]. Most association signals map to non-coding regions of the genome, for which biological interpretation is inherently challenging[248,249]. Consequently, once a GWAS has been performed, additional steps are often required to identify the causal variants and their target genes, for example, multi-ethnic or admixed population re-sequencing and fine-mapping, methodological developments, functional analyses or evolutionary genetic analyses[250–256]. Although identifying causal variants might be easier for GWAS using WGS than for GWAS using SNP arrays, in that all genetic variation has been ascertained, functional characterization is challenging regardless of the technology used — a key reason being that hypotheses about the underlying mechanisms are typically required.

Another challenge in assigning causality in GWAS is that too many hits may be involved. Not long ago, a lack of replication plagued genetic association studies[257]. However, akin to a transition from drought to flood, the number of discoveries in human genetics has rapidly accelerated over the past 13 years[4]. For example, recent simulations have shown that 90,000–100,000 SNPs may be needed to explain 80% of the heritability of height[10,11]. This means that a substantial fraction of all genes may contribute to variation of complex traits[10]. In pointing at 'everything', the danger is that GWAS could point at 'nothing'[11]. The recently proposed 'omnigenic' model suggests that gene regulatory networks are sufficiently interconnected such that all genes expressed in disease-relevant cells are liable to affect the functions of core disease-related genes and that most of the heritability can be explained by the effects of genes outside core pathways[10]. Genes associated with the continuum from monogenic to polygenic forms of disease may be more likely to contribute to core biological pathways and can be prioritized for functional investigation[244].

Despite the difficulties in interpreting GWAS associations, much progress has been made in moving from association to function and causation. Several advances have aided in fine-mapping and prioritizing variants for functional follow-up, especially in the non-coding genome. First, improvements in the density of SNP arrays and imputation reference panels have allowed the mapping resolution of common variant associations in GWAS to approach that of a fine-mapping study. A caveat is that the association of a common variant may be the result of partial linkage disequilibrium with one or more rare variants of large effect that happen to segregate on common haplotypes, a phenomenon known as synthetic association[86]. Although synthetic associations have been reported in the literature (for example, *NOD2* (REF.[258]), *HBB*[86], *MYH6* (REF.[259]) or *SERPINA1* (REF.[260])), multiple lines of evidence suggest that they are rare[87,261], and studies have increasingly demonstrated support for independent contributions of rare and common variants at a single locus[262,263]. Second, custom genotyping arrays that provide dense SNP coverage in candidate disease-associated regions, such as the MetaboChip, iCOGS array and ImmunoChip, provide a cost-effective strategy for fine-mapping certain diseases and traits[264–269]. Third, trans-ethnic and admixed population fine-mapping can be used to refine regions of association by exploiting population differences in linkage disequilibrium and can be incorporated into the initial GWAS step[128,215,253]. Assuming the causal variant is associated with the disease across ethnicities and linkage disequilibrium varies with ethnicity at the associated locus, meta-analysis of genetic data from different ethnic backgrounds can magnify the associations of the causal variants and tone down the associations of proxies. Fourth, advances in methods for statistical fine-mapping, such as Bayesian approaches, have made great inroads into narrowing down the possible causal variants, for example, with credible sets of SNVs[270]. Last, the rapid development of publicly available databases of regulatory elements across a range of tissues and cells types (for example, ENCODE[271], Epigenome RoadMap[272], FANTOM5 (REF.[273]) and GTEx[274]), as well as tools for querying such databases (for example, RegulomeDB[275] and HaploReg[276,277]), has allowed GWAS findings to be integrated with functional genomics data at multiple levels, prioritizing candidate variants for functional follow-up — for example, by testing for colocalization with expression or methylation quantitative trait loci, or for overlap with accessible chromatin, transcription factor binding or regulatory histone marks[278].

In parallel, progress has also been observed in the functional characterization of causal variants and the identification of target genes. Several experimental approaches are available to test the functions of candidate variants and to determine the molecular mechanisms[278]. Chromosome conformation capture and its derivatives can be used to visualize the 3D organization of chromatin and are important methods for

determining target genes[279]; the relationship between the regulatory variant and the target gene is complex, with functional studies suggesting that only about one-third of causal genes are the nearest gene to the GWAS hit[280,281]. An example of this complexity is the *FTO* locus in obesity. Intronic variants in *FTO* associated with obesity in GWAS were initially assumed by many to regulate *FTO*[282,283]. However, functional studies found that the obesity-associated *FTO* region interacts with the *IRX3* promoter, located several hundred kilobases away, and that these intronic variants are associated with the expression of *IRX3* but not *FTO* in human brains[284]. Additional work using mouse models confirmed *IRX3* as a likely causal gene[284], and an additional target, *IRX5*, has also been identified[285].

GWAS cannot identify all genetic determinants of complex traits. It is unlikely that GWAS will ever explain 100% of the heritability of complex traits. This limitation is not exclusive to GWAS, as no method or technology to date can identify all the genetic components of complex traits. The difficulty in detecting common variants with very small effects, rare variants with small effects, genes harbouring ultra-rare variants and complex interactions (gene–gene and gene–environment) makes explaining all the heritability of complex traits an impossible task[39]. Another challenge lies in accurately estimating the heritability of complex traits[245,286,287].

GWAS have been largely unsuccessful in detecting epistasis in humans. Although evidence of non-additive heritability is difficult to assess in humans, model organisms (for example, yeast, worm, fly or mouse) have established epistasis as a pivotal component of the genetic architecture of complex traits[288–290]. However, the identification of significant gene–gene interactions has been challenging in GWAS and post-GWAS experiments in humans, owing primarily to a lack of statistical power and to methodological challenges[128,290–292]. As there is still limited evidence that epistasis contributes to a large fraction of the total genetic variation of complex traits in humans, very large sample sizes may be needed to detect significant gene–gene interactions[2]. Furthermore, the loss of information caused by imperfect linkage disequilibrium between genotyped and causal variants is larger for interactions than for main effects[2]. Recent methodological developments (for example, data filtering, Bayesian methods and artificial intelligence algorithms) may boost the identification of epistatic interactions in humans[290]. However, epistasis remains challenging regardless of whether one uses a GWAS or WGS, as it depends on the power of much larger samples and wider computing throughput to accommodate the exponential increase in statistical tests.

GWAS signals may be due to cryptic population stratification. An important concern in genetic association studies is population stratification, which can result in spurious associations if not properly accounted for. Population stratification is especially a challenge in large GWAS, for which perfect matching of cases and controls is virtually impossible[9], but is also a concern when studying recently admixed populations and variants with very small effect sizes[1].

Most GWAS signals (OR < 1.5) have been suggested to be attributable to cryptic population stratification[9]. However, this view is probably too extreme owing to several reasons. First, efficient methods are used in GWAS to control for population stratification[293,294]. Second, many GWAS loci are enriched for biologically relevant variants[10,240] or localize in genes or pathways known or postulated to play a role in disease[6]. Third, GWAS hits have been confirmed by family-based association tests, which are robust to population stratification[293,295–297]. Fourth, although population stratification is a greater concern in case–control studies than in population-based studies, analysing the continuous and binary versions of a trait in a single data set leads to an almost entirely overlapping list of GWAS signals[193]. Last, identification of the same disease-associated GWAS SNPs in diverse ethnic groups and association of GWAS SNPs with future risk of disease in prediction models suggest that the majority of GWAS hits are indeed true signals[1].

GWAS have limited clinical predictive value. The modest proportion of heritability explained and the small effect sizes of GWAS-identified SNVs limit their clinical predictive value. For most complex traits, identified SNVs in aggregate perform poorly at discriminating between individuals with and without the disease[298,299]. Between private, highly penetrant mutations of large effect (OR > 10) and common variants of modest effect (OR 1.05–1.30) exists an intermediate category of variants with low allele frequencies and modest-to-strong effects on disease. These variants are extremely relevant to disease prediction[300] and may account for a significant fraction of complex trait heritability[78]. They are also likely to be causal variants, owing to low linkage disequilibrium with other variants. For example, the obesity-associated low-frequency SNV rs6232 (p.Asn221Asp) in *PCSK1* is not in strong linkage disequilibrium ($r^2 > 0.8$) with any other variant[68]. Such variants will constitute the main target of the new generation of GWAS. For obesity, these include rs2229616 (p.Val103Ile) and rs52820871 (p.Ile251Leu) in *MC4R* (MAF 0.5–1%, OR 0.52–0.80)[301], rs6232 (p.Asn221Asp) in *PCSK1* (MAF 3%, OR 1.34)[68], rs116454156 (p.Arg270His) in *FFAR4* (also known as *GRP120*) (MAF 1%, OR 1.62)[302] and rs28932472 (p.Arg236Gly) in *POMC* (MAF 0.5%, OR 4)[303]. Recently, a large-scale exome-wide meta-analysis identified 14 rare and low-frequency coding variants that increase weight by 0.315–7.05 kg (REF.[70]). As the number of identified disease-associated variants increases, their cumulative predictive value will undoubtedly gain in importance[240].

The relevance of using GWAS findings to predict, prevent and treat disease remains a subject of intense debate. Even for a disease such as T1DM, for which most of the heritability can be explained by GWAS loci, genetic screening at the population level is not feasible, as the number of false positives would greatly exceed the number of true positives[238]. In addition, identifying individuals at risk of disease may not be meaningful if no personalized treatment is available. Instead, screening for rare and low-frequency monogenic and

Epistasis
Statistical interaction between loci in their effect on a trait such that the effect of a genotype at one locus is dependent on the genotypes at the other locus (or loci).

Population stratification
Differences in allele frequencies between cases and controls resulting from systematic differences in ancestry rather than association of genes with disease.

oligogenic variants is more likely to lead to actionable treatments[304–308]. Clinical prediction might also prove to be especially useful in small isolated populations where deleterious variants with strong effect have risen to high frequency (for example, Inuit in Greenland)[309].

For a subset of diseases, however, polygenic risk scores (PRSs) — quantitative measures of risk summed across multiple risk alleles — have begun to show promise in their ability to separate a population into categories with sufficiently distinct risks to affect clinical and personal decision-making[310]. For example, the value of a PRS in cancer as a tool for stratification in public health has been exemplified in several studies[311], including a study of breast cancer that combined a PRS with conventional risk factors to identify 16% of the population who could benefit from earlier screening (and 32% who could delay screening)[312]. For coronary artery disease, PRSs have identified individuals with risk equivalent to rare monogenic mutations[313,314]; these individuals constitute a substantially larger fraction of the population than do individuals with rare monogenic mutations[313] and might derive greater benefit from early lifestyle interventions and initiation of statin therapy than individuals at lower genetic risk[315,316]. Although knowledge of individual genetic risk can improve readiness to adopt a healthier lifestyle, human behaviour is complex and genetic testing may not necessarily translate into improved long-term clinical outcomes[317,318].

Limitations specific to SNP array-based GWAS

GWAS based on SNP arrays rely on pre-existing genetic variant reference panels. A limitation of SNP array-based GWAS is that they depend on the completeness of the sequencing studies and resulting reference panels that are used to inform genotyping array design and to impute untyped variants in GWAS[319,320]. For example, early genome-wide SNP arrays were designed by selecting tag SNPs from reference panels of predominantly European populations[321]. Because linkage disequilibrium patterns vary across ethnic groups, these arrays often provided poor coverage in non-European populations[39,321]. This problem has since improved with the development of a new generation of high-density arrays whose contents are based on sequencing data from more diverse populations, as well as the development of ethnic-specific and trans-ethnic arrays designed specifically to optimize genomic coverage in non-European ethnic groups[322]. However, although these new arrays should collectively enable a large harvest of ethnic-specific disease signals, many ethnic groups (for example, Indigenous Australians, Native Americans, Pacific Islanders, Middle Eastern Arabs and African Pygmies) still have not been sequenced. Hence, optimal GWAS and genotype imputation cannot yet be performed in these populations[323,324]. Even for populations that have been sequenced, larger reference panels are still needed to improve genomic coverage, genotype imputation and, subsequently, the detection of novel SNV–trait associations[84,325]. Moreover, much like the design of early genome-wide SNP arrays, content on the ExomeChip and other custom genotyping arrays is based primarily on sequencing data from European individuals[72].

GWAS based on SNP arrays cannot detect ultra-rare mutations contributing to disease. Whether the remaining heritability is explained by common variants of modest effect or rare variants of large effect is still under debate[5,8,9,11,39]. Although empirical evidence suggests that much of the heritability of complex traits can be explained by common variants[26–28], rare and ultra-rare variants are also expected to contribute[326,327]. In this context, it is important to note that SNP array-based GWAS are unable to detect ultra-rare variants associated with disease.

Genotype imputation using reference panels derived from WES and WGS projects enables GWAS using common SNP arrays to identify associations with rare variants, by recovering some of the information lost to imperfect linkage disequilibrium. WES and WGS efforts are underway in many populations worldwide to uncover rare variants that are specific to an ethnic group or population[83,328]. However, these initiatives are far from complete, and it will take substantial time and resources before imputation of even a subset of rare variants representative of the worldwide ethnic and geographical diversity can be achieved[252]. At the present time, statistical imputation using large reference population panels is reasonably accurate for variants as infrequent as 1 in 1,000 (REF.[71]). However, imputation of ultra-rare variants (for example, a frequency of 1 in 100,000) and private mutations (for example, those only found in one pedigree or individual worldwide) is nearly impossible, as these variants are unlikely to be identified in such sequencing efforts[2].

Although large-scale WES and WGS initiatives can identify at least a subset of ultra-rare variants, demonstrating a correlation between these variants and a trait or disease will be extremely challenging. For instance, GWAS using WGS in more than 1 million individuals may be required to detect an association between an ultra-rare variant with a large effect (for example, one phenotypic standard deviation unit) and a quantitative trait[2]. Power may be increased by studying extreme cases, using family-based study designs (for example, studying families with multiple cases of a rare disease) and using rare-variant burden tests across genes and targeted candidate gene strategies[2,99,329].

Finally, for GWAS to successfully identify a gene for a given pathology, disease-associated genetic variation at the locus must exist. In theory, there could be ultra-conserved regions in the genome containing genes with important biological roles in disease that do not exhibit deleterious variation. Crucially, GWAS will not be able to identify these genes.

Conclusions

The emergence of GWAS well over a decade ago has caused a remarkable shift in our capacity to understand the genetic basis of human disease. The ever-expanding list of replicable associations has extended beyond common variation to rare variation[4], and the value of these associations in realizing fundamental goals of human genetics is now clear. GWAS loci have generated new insights into disease biology that have supported clinical translation[2] and are beginning to show promise in

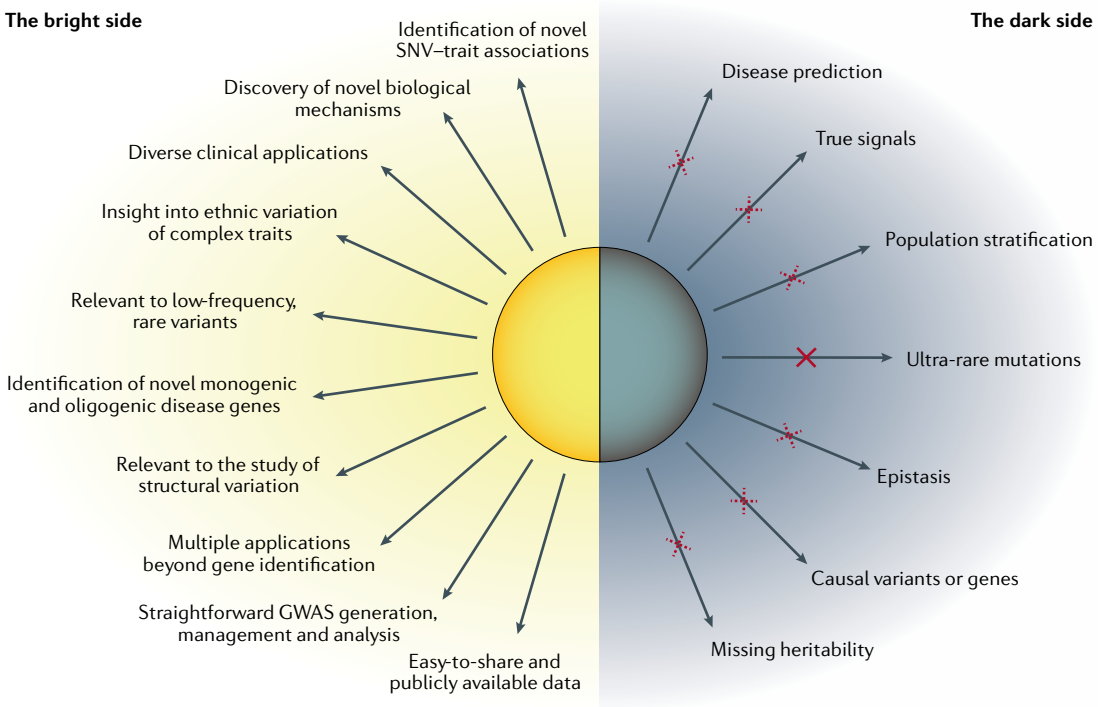


Fig. 4 | **Benefits and limitations of GWAS using SNP arrays.** A visual depiction of the current benefits (the bright side) and limitations (the dark side) of genome-wide association studies (GWAS). The solid X indicates a permanent limitation. The dotted Xs represent limitations that have the potential to be overcome, at least to some extent, in the future (for example, with larger sample sizes, technological and methodological advancements, and a shift from the use of single-nucleotide polymorphism (SNP) arrays to whole-genome sequencing). SNV, single-nucleotide variant.

population risk stratification[310]. And still, the full potential of GWAS has not been unlocked. Research institutions and funding agencies should feel encouraged by the numerous arguments that continue to support the GWAS design (FIG. 4).

Some researchers have expressed important concerns about the value of GWAS, for example, that multiple rare variants may induce synthetic associations with common SNPs[86]. Although the literature suggests that synthetic associations are likely to be rare[100,272], we note that the new generation of dense genome-wide SNP arrays and the availability of large WGS reference panels provide an opportunity to test the synthetic association hypothesis more accurately than could be achieved in the past[330].

Others have suggested that most GWAS signals are the result of cryptic population stratification[9]. Although population stratification probably does not account for most GWAS signals, we propose several strategies to overcome this possibility: first, correct for population substructure; second, perform family-based GWAS or at least use family-based replication designs to validate GWAS associations from case–control studies; third, compare GWAS associations across ethnic backgrounds; and fourth, strengthen the results of case–control studies (for example, on obesity) with those of corresponding quantitative trait studies (for example, on BMI).

Concerns have also been raised that GWAS will implicate too many loci: most disease-associated variants will have infinitesimally small effect sizes and will not pinpoint core genes with direct effects on disease[10,11]. However, the fact that complex diseases are extremely polygenic, involving many variants of small effect sizes, does not preclude clinical utility of identified variants. For example, statins are effective in lowering cardiovascular disease risk, yet GWAS loci at the drug-target locus explain only about 1% of the phenotypic variation[5]. The real question therefore is not whether there are too many loci, but rather how do they cumulatively explain disease, and can they be used for individual prediction or population stratification. Although identifying more loci might complicate the identification of causal genes, this is certainly an advantage when it comes to risk prediction — an equally important goal to that of clarifying the causal genes and their complex interactions. Already, PRSs comprising hundreds of thousands to millions of SNPs provide clinical utility for certain diseases and are being introduced into therapeutic decision-making[313]. As more powerful GWAS are performed, future PRSs, consisting of common, low-frequency and rare variants, as well as incorporating or complementing familial and environmental risk factors, will provide even better risk stratification[331], with important implications for future research, screening and primary prevention, personal and therapeutic decision-making, public health and clinical trial enrichment, and for making differential diagnoses.

Many of the current limitations of GWAS are not insurmountable or can be overcome at least to some extent (FIG. 4). For example, larger sample sizes, advances in technology, methodology and computing, as well as

a shift from the use of SNP arrays to WGS have the potential to resolve many of the limitations (for example, improve risk prediction, identify missed signals, account for population stratification, identify ultra-rare mutations, identify gene–gene and gene–environment interactions, identify causal variants and genes, and explain more of the missing heritability).

WGS is the gold standard in GWAS[79,328,330]. As the price of sequencing falls, GWAS using WGS in large samples will become increasingly realistic. In the meantime, GWAS based on dense SNP arrays combined with imputation to large WGS reference panels will be complementary to the study of rare variants and will continue to provide major advances in the field of complex disease genetics. This view is supported by the success of recent GWAS that imputed SNVs using large WGS reference panels[330,332] and by a recent large-scale WGS study of T2DM, which identified variants that were overwhelmingly common and located in regions already discovered using SNP arrays[79]. For a study design that is already more than a decade old, the still growing number of published GWAS is a testament to the continued success of this approach in elucidating the genetic basis of complex human traits[2].

Published online 8 May 2019

1. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *Am. J. Hum. Genet.* **90**, 7–24 (2012).
This paper reviews the progress and discoveries made in the first 5 years of GWAS.
2. Visscher, P. M. et al. 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
This paper provides an overview of the lessons learned from the past decade of GWAS.
3. Klein, R. J. et al. Complement factor H polymorphism in age-related macular degeneration. *Science* **308**, 385–389 (2005).
This study may be considered the first GWAS to be published.
4. MacArthur, J. et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2017).
This paper provides a description of the GWAS Catalog, a continuously updated list of GWAS and their results.
5. Hirschhorn, J. N. Genomewide association studies — illuminating biologic pathways. *N. Engl. J. Med.* **360**, 1699–1701 (2009).
6. Klein, R. J., Xu, X., Mukherjee, S., Willis, J. & Hayes, J. Successes of genome-wide association studies. *Cell* **142**, 350–351 (2010).
7. Speakman, J., Loos, R., O'Rahilly, S., Hirschhorn, J. & Allison, D. GWAS for BMI: a treasure trove of fundamental insights into the genetic basis of obesity. *Int. J. Obes.* **42**, 1524–1531 (2018).
8. Manolio, T. A. et al. Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
9. McClellan, J. & King, M. C. Genetic heterogeneity in human disease. *Cell* **141**, 210–217 (2010).
The authors of this paper suggest that most GWAS findings may be due to cryptic population stratification.
10. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).
This reference provides a theoretical groundwork for the omnigenic model of inheritance.
11. Goldstein, D. B. Common genetic variation and human traits. *N. Engl. J. Med.* **360**, 1696–1698 (2009).
In this paper, the author is among the first to suggest that GWAS may eventually implicate most of the genome.
12. Meyre, D. Give GWAS a chance. *Diabetes* **66**, 2741–2742 (2017).
13. Duncan, L. et al. Significant locus and metabolic genetic correlations revealed in genome-wide association study of anorexia nervosa. *Am. J. Psychiatry* **174**, 850–858 (2017).
14. Hyde, C. L. et al. Identification of 15 genetic loci associated with risk of major depression in individuals of European descent. *Nat. Genet.* **48**, 1031–1036 (2016).
15. Milne, R. L. et al. Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. *Nat. Genet.* **49**, 1767–1778 (2017).
16. Sud, A., Kinnersley, B. & Houlston, R. S. Genome-wide association studies of cancer: current insights and future perspectives. *Nat. Rev. Cancer* **17**, 692–704 (2017).
17. Zhao, W. et al. Identification of new susceptibility loci for type 2 diabetes and shared etiological pathways with coronary heart disease. *Nat. Genet.* **49**, 1450–1457 (2017).
18. Nikpay, M. et al. A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat. Genet.* **47**, 1121–1130 (2015).
19. Li, Z. et al. Genome-wide association analysis identifies 30 new susceptibility loci for schizophrenia. *Nat. Genet.* **49**, 1576–1583 (2017).
20. de Lange, K. M. et al. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat. Genet.* **49**, 256–261 (2017).
21. Jansen, P. R. et al. Genome-wide analysis of insomnia in 1,331,010 individuals identifies new risk loci and functional pathways. *Nat. Genet.* **51**, 394–403 (2019).
This is the largest GWAS published to date.
22. Yengo, L. et al. Meta-analysis of genome-wide association studies for height and body mass index in approximately 700000 individuals of European ancestry. *Hum. Mol. Genet.* **27**, 3641–3649 (2018).
23. Lee, J. J. et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat. Genet.* **50**, 1112–1121 (2018).
24. Lohmueller, K. E., Pearce, C. L., Pike, M., Lander, E. S. & Hirschhorn, J. N. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat. Genet.* **33**, 177–182 (2003).
25. Yang, J. et al. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
This study helped to largely resolve the 'missing' heritability problem, by demonstrating that a large portion of the heritability can be explained by common SNPs.
26. Yang, J. et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat. Genet.* **47**, 1114–1120 (2015).
27. Loh, P.-R. et al. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat. Genet.* **47**, 1385–1392 (2015).
28. Yang, J. et al. Ubiquitous polygenicity of human complex traits: genome-wide analysis of 49 traits in Koreans. *PLOS Genet.* **9**, e1003355 (2013).
29. Ahlqvist, E., van Zuydam, N. R., Groop, L. C. & McCarthy, M. I. The genetics of diabetic complications. *Nat. Rev. Nephrol.* **11**, 277–287 (2015).
30. Wray, N. R., Wijmenga, C., Sullivan, P. F., Yang, J. & Visscher, P. M. Common disease is more complex than implied by the core gene omnigenic model. *Cell* **173**, 1573–1580 (2018).
31. Hampe, J. et al. A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1. *Nat. Genet.* **39**, 207–211 (2007).
32. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
33. Murthy, A. et al. A Crohn's disease variant in Atg16l1 enhances its degradation by caspase 3. *Nature* **506**, 456–462 (2014).
34. Brest, P. et al. A synonymous variant in IRGM alters a binding site for miR-196 and causes deregulation of IRGM-dependent xenophagy in Crohn's disease. *Nat. Genet.* **43**, 242–245 (2011).
35. Williams, A. L. et al. Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico. *Nature* **506**, 97–101 (2014).
36. Rusu, V. et al. Type 2 diabetes variants disrupt function of SLC16A11 through two distinct mechanisms. *Cell* **170**, 199–212 (2017).
37. Sekar, A. et al. Schizophrenia risk from complex variation of complement component 4. *Nature* **530**, 177–183 (2016).
38. Stefansson, H. et al. Common variants conferring risk of schizophrenia. *Nature* **460**, 744–747 (2009).
39. Altshuler, D., Daly, M. J. & Lander, E. S. Genetic mapping in human disease. *Science* **322**, 881–888 (2008).
40. Morris, Z. S., Wooding, S. & Grant, J. The answer is 17 years, what is the question: understanding time lags in translational research. *J. R. Soc. Med.* **104**, 510–520 (2011).
41. Edwards, A. O. et al. Complement factor H polymorphism and age-related macular degeneration. *Science* **308**, 421–424 (2005).
42. Chen, W. et al. Genetic variants near TIMP3 and high-density lipoprotein-associated loci influence susceptibility to age-related macular degeneration. *Proc. Natl Acad. Sci. USA* **107**, 7401–7406 (2010).
43. Thorleifsson, G. et al. Common sequence variants in the LOXL1 gene confer susceptibility to exfoliation glaucoma. *Science* **317**, 1397–1400 (2007).
44. Shields, B. M. et al. Maturity-onset diabetes of the young (MODY): how many cases are we missing? *Diabetologia* **53**, 2504–2508 (2010).
45. Yamagata, K. et al. Mutations in the hepatocyte nuclear factor-1α gene in maturity-onset diabetes of the young (MODY3). *Nature* **384**, 455–458 (1996).
46. Ridker, P. M. et al. Loci related to metabolic-syndrome pathways including LEPR,HNF1A, IL6R, and GCKR associate with plasma C-reactive protein: the Women's Genome Health Study. *Am. J. Hum. Genet.* **82**, 1185–1192 (2008).
47. Reiner, A. P. et al. Polymorphisms of the HNF1A gene encoding hepatocyte nuclear factor-1α are associated with C-reactive protein. *Am. J. Hum. Genet.* **82**, 1193–1201 (2008).
48. Owen, K. R. et al. Assessment of high-sensitivity C-reactive protein levels as diagnostic discriminator of maturity-onset diabetes of the young due to HNF1A mutations. *Diabetes Care* **33**, 1919–1924 (2010).
49. Thanabalasingham, G. et al. A large multi-centre European study validates high-sensitivity C-reactive protein (hsCRP) as a clinical biomarker for the diagnosis of diabetes subtypes. *Diabetologia* **54**, 2801–2810 (2011).
50. Nelson, M. R. et al. The support of human genetic evidence for approved drug indications. *Nat. Genet.* **47**, 856–860 (2015).
51. Giacomini, K. M. et al. Genome-wide association studies of drug response and toxicity: an opportunity for genome medicine. *Nat. Rev. Drug Discov.* **16**, 1 (2017).
52. Ge, D. et al. Genetic variation in IL28B predicts hepatitis C treatment-induced viral clearance. *Nature* **461**, 399–401 (2009).
53. Suppiah, V. et al. IL28B is associated with response to chronic hepatitis C interferon-α and ribavirin therapy. *Nat. Genet.* **41**, 1100–1104 (2009).
54. Tanaka, Y. et al. Genome-wide association of IL28B with response to pegylated interferon-α and ribavirin therapy for chronic hepatitis C. *Nat. Genet.* **41**, 1105–1109 (2009).

55. Link, E. et al. SLCO1B1 variants and statin-induced myopathy — a genomewide study. *N. Engl. J. Med.* **359**, 789–799 (2008).
56. Muir, A. J. et al. Clinical Pharmacogenetics Implementation Consortium (CPIC) guidelines for IFNL3 (IL28B) genotype and PEG interferon-α-based regimens. *Clin. Pharmacol. Ther.* **95**, 141–146 (2014).
57. Ramsey, L. B. et al. The clinical pharmacogenetics implementation consortium guideline for SLCO1B1 and simvastatin-induced myopathy: 2014 update. *Clin. Pharmacol. Ther.* **96**, 423–428 (2014).
58. Liu, J. Z. et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47**, 979–986 (2015).
59. Sladek, R. et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**, 881–885 (2007).
This study is the first true GWAS for a complex disease that used SNP arrays with exhaustive coverage of the genome.
60. Yasuda, K. et al. Variants in KCNQ1 are associated with susceptibility to type 2 diabetes mellitus. *Nat. Genet.* **40**, 1092–1097 (2008).
61. Unoki, H. et al. SNPs in KCNQ1 are associated with susceptibility to type 2 diabetes in East Asian and European populations. *Nat. Genet.* **40**, 1098–1102 (2008).
62. Moltke, I. et al. A common Greenlandic TBC1D4 variant confers muscle insulin resistance and type 2 diabetes. *Nature* **512**, 190–193 (2014).
63. Bosse, Y. & Amos, C. I. A. Decade of GWAS results in lung cancer. *Cancer Epidemiol. Biomarkers Prev.* **27**, 363–379 (2018).
64. Minster, R. L. et al. A thrifty variant in CREBRF strongly influences body mass index in Samoans. *Nat. Genet.* **48**, 1049–1054 (2016).
65. Nead, K. T. et al. Contribution of common non-synonymous variants in PCSK1 to body mass index variation and risk of obesity: a systematic review and meta-analysis with evidence from up to 331 175 individuals. *Hum. Mol. Genet.* **24**, 3582–3594 (2015).
66. Choquet, H., Kasberger, J., Hamidovic, A. & Jorgenson, E. Contribution of common PCSK1 genetic variants to obesity in 8,359 subjects from multi-ethnic American population. *PLOS ONE* **8**, e57857 (2013).
67. Kurokawa, N. et al. The ADRB3 Trp64Arg variant and BMI: a meta-analysis of 44 833 individuals. *Int. J. Obes.* **32**, 1240–1249 (2008).
68. Benzinou, M. et al. Common nonsynonymous variants in PCSK1 confer risk of obesity. *Nat. Genet.* **40**, 943–945 (2008).
69. Wen, W. et al. Meta-analysis identifies common variants associated with body mass index in East Asians. *Nat. Genet.* **44**, 307–311 (2012).
70. Turcot, V. et al. Protein-altering variants associated with body mass index implicate pathways that control energy intake and expenditure in obesity. *Nat. Genet.* **50**, 26–41 (2018).
71. McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
This study describes one of the largest reference panels to date, which comprises 64,976 haplotypes and provides accurate genotype imputation at MAFs as low as 0.1%.
72. Grove, M. L. et al. Best practices and joint calling of the HumanExome BeadChip: the CHARGE Consortium. *PLOS ONE* **8**, e68095 (2013).
73. Peloso, G. M. et al. Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. *Am. J. Hum. Genet.* **94**, 223–232 (2014).
74. Auer, P. L. et al. Rare and low-frequency coding variants in CXCR2 and other genes are associated with hematological traits. *Nat. Genet.* **46**, 629–634 (2014).
75. CHARGE Consortium Hematology Working Group. Meta-analysis of rare and common exome chip variants identifies S1PR4 and other loci influencing blood cell traits. *Nat. Genet.* **48**, 867–876 (2016).
76. Eicher, J. D. et al. Platelet-related variants identified by exomechip meta-analysis in 157,293 individuals. *Am. J. Hum. Genet.* **99**, 40–55 (2016).
77. Surendran, P. et al. Trans-ancestry meta-analyses identify rare and common variants associated with blood pressure and hypertension. *Nat. Genet.* **48**, 1151–1161 (2016).
78. Marouli, E. et al. Rare and low-frequency coding variants alter human adult height. *Nature* **542**, 186–190 (2017).
79. Fuchsberger, C. et al. The genetic architecture of type 2 diabetes. *Nature* **536**, 41–47 (2016).
This is a large GWAS for T2DM that finds little evidence for low-frequency and rare variants despite being sufficiently powered to detect such associations.
80. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11**, 499–511 (2010).
81. Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68 (2015).
82. Deelen, P. et al. Improved imputation quality of low-frequency and rare variants in European samples using the 'Genome of The Netherlands'. *Eur. J. Hum. Genet.* **22**, 1321–1326 (2014).
83. Gudbjartsson, D. F. et al. Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* **47**, 435–444 (2015).
84. Huang, J. et al. Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat. Commun.* **6**, 8111 (2015).
85. Taliun, D. et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Preprint at *bioRxiv* https://www.biorxiv.org/content/10.1101/563866v1 (2019).
86. Dickson, S. P., Wang, K., Krantz, I., Hakonarson, H. & Goldstein, D. B. Rare variants create synthetic genome-wide associations. *PLOS Biol.* **8**, e1000294 (2010).
The authors of this paper lay out the theoretical basis for synthetic associations in GWAS.
87. Anderson, C. A., Soranzo, N., Zeggini, E. & Barrett, J. C. Synthetic associations are unlikely to account for many common disease genome-wide association signals. *PLOS Biol.* **9**, e1000580 (2011).
88. Doche, M. E. et al. Human SH2B1 mutations are associated with maladaptive behaviors and obesity. *J. Clin. Invest.* **122**, 4732–4736 (2012).
89. Liu, R. et al. Rare loss-of-function variants in NPC1 predispose to human obesity. *Diabetes* **66**, 935–947 (2017).
90. Saeed, S. et al. Loss-of-function mutations in ADCY3 cause monogenic severe obesity. *Nat. Genet.* **50**, 175–179 (2018).
91. Grarup, N. et al. Loss-of-function variants in ADCY3 increase risk of obesity and type 2 diabetes. *Nat. Genet.* **50**, 172–174 (2018).
92. Nejentsev, S., Walker, N., Riches, D., Egholm, M. & Todd, J. A. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science* **324**, 387–389 (2009).
93. Bonnefond, A. et al. Rare MTNR1B variants impairing melatonin receptor 1B function contribute to type 2 diabetes. *Nat. Genet.* **44**, 297–301 (2012).
94. Flannick, J. et al. Loss-of-function mutations in SLC30A8 protect against type 2 diabetes. *Nat. Genet.* **46**, 357–363 (2014).
95. Majithia, A. R. et al. Rare variants in PPARG with decreased activity in adipocyte differentiation are associated with increased risk of type 2 diabetes. *Proc. Natl Acad. Sci. USA* **111**, 13127–13132 (2014).
96. Cardinale, C. J. et al. Targeted resequencing identifies defective variants of decoy receptor 3 in pediatric-onset inflammatory bowel disease. *Genes Immun.* **14**, 447–452 (2013).
97. Ellinghaus, D. et al. Association between variants of PRDM1 and NDP52 and Crohn's disease, based on exome sequencing and functional studies. *Gastroenterology* **145**, 339–347 (2013).
98. Beaudoin, M. et al. Deep resequencing of GWAS loci identifies rare variants in CARD9, IL23R and RNF186 that are associated with ulcerative colitis. *PLOS Genet.* **9**, e1003723 (2013).
99. Philippe, J. et al. A nonsense loss-of-function mutation in PCSK1 contributes to dominantly inherited human obesity. *Int. J. Obes.* **39**, 295–302 (2015).
100. Lessard, S. et al. Testing the role of predicted gene knockouts in human anthropometric trait variation. *Hum. Mol. Genet.* **25**, 2082–2092 (2016).
101. Walters, R. G. et al. A new highly penetrant form of obesity due to deletions on chromosome 16p11.2. *Nature* **463**, 671–675 (2010).
102. Bochukova, E. G. et al. Large, rare chromosomal deletions associated with severe early-onset obesity. *Nature* **463**, 666–670 (2010).
103. Willer, C. J. et al. Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat. Genet.* **41**, 25–34 (2009).
104. Speliotes, E. K. et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet.* **42**, 937–948 (2010).
105. Wheeler, E. et al. Genome-wide SNP and CNV analysis identifies common and low-frequency variants associated with severe early-onset obesity. *Nat. Genet.* **45**, 513–517 (2013).
106. Malhotra, D. et al. High frequencies of de novo CNVs in bipolar disorder and schizophrenia. *Neuron* **72**, 951–963 (2011).
107. Pinto, D. et al. Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466**, 368–372 (2010).
108. Jacobs, K. B. et al. Detectable clonal mosaicism and its relationship to aging and cancer. *Nat. Genet.* **44**, 651–658 (2012).
109. Laurie, C. C. et al. Detectable clonal mosaicism from birth to old age and its relationship to cancer. *Nat. Genet.* **44**, 642–650 (2012).
110. Forsberg, L. A. et al. Mosaic loss of chromosome Y in peripheral blood is associated with shorter survival and higher risk of cancer. *Nat. Genet.* **46**, 624–628 (2014).
111. Bonnefond, A. et al. Association between large detectable clonal mosaicism and type 2 diabetes with vascular complications. *Nat. Genet.* **45**, 1040–1043 (2013).
112. Genovese, G. et al. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N. Engl. J. Med.* **371**, 2477–2487 (2014).
113. Tregouet, D. A. et al. Genome-wide haplotype association study identifies the SLC22A3-LPAL2-LPA gene cluster as a risk locus for coronary artery disease. *Nat. Genet.* **41**, 283–285 (2009).
114. El-Sayed Moustafa, J. S. et al. Novel association approach for variable number tandem repeats (VNTRs) identifies DOCK5 as a susceptibility gene for severe obesity. *Hum. Mol. Genet.* **21**, 3727–3738 (2012).
115. Stacey, S. N. et al. Insertion of an SVA-E retrotransposon into the CASP8 gene is associated with protection against prostate cancer. *Hum. Mol. Genet.* **25**, 1008–1018 (2016).
116. de Vries, P. S. et al. A meta-analysis of 120 246 individuals identifies 18 new loci for fibrinogen concentration. *Hum. Mol. Genet.* **25**, 358–370 (2016).
117. Ma, J., Xiong, M., You, M., Lozano, G. & Amos, C. I. Genome-wide association tests of inversions with application to psoriasis. *Hum. Genet.* **133**, 967–974 (2014).
118. Reich, D. et al. Reconstructing Native American population history. *Nature* **488**, 370–374 (2012).
119. Reich, D., Thangaraj, K., Patterson, N., Price, A. L. & Singh, L. Reconstructing Indian population history. *Nature* **461**, 489–494 (2009).
120. Fu, Q. et al. The genetic history of Ice Age Europe. *Nature* **534**, 200–205 (2016).
121. Lazaridis, I. et al. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**, 409–413 (2014).
122. Price, A. L. et al. Discerning the ancestry of European Americans in genetic association studies. *PLOS Genet.* **4**, e236 (2008).
123. Jakkula, E. et al. The genome-wide patterns of variation expose significant substructure in a founder population. *Am. J. Hum. Genet.* **83**, 787–794 (2008).
124. Hoggart, C. J. et al. Fine-scale estimation of location of birth from genome-wide single-nucleotide polymorphism data. *Genetics* **190**, 669–677 (2012).
125. Goode, E. L. & Jarvik, G. P. Assessment and implications of linkage disequilibrium in genome-wide single-nucleotide polymorphism and microsatellite panels. *Genet. Epidemiol.* **29**, S72–S76 (2005).
126. Bulik-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
127. Ross, S. et al. Mendelian randomization analysis supports the causal role of dysglycaemia and diabetes in the risk of coronary artery disease. *Eur. Heart J.* **36**, 1454–1462 (2015).
128. Liu, H. Y. et al. Fine-mapping of 98 obesity loci in Mexican children. *Int. J. Obes.* **43**, 23–32 (2018).
129. Homer, N. et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLOS Genet.* **4**, e1000167 (2008).
130. Kling, D. et al. DNA microarray as a tool in establishing genetic relatedness — current status and future prospects. *Forens. Sci. Int. Genet.* **6**, 322–329 (2012).
131. Ramstetter, M. D. et al. Benchmarking relatedness inference methods with genome-wide data from thousands of relatives. *Genetics* **207**, 75–82 (2017).

132. Kerr, S. M. et al. Pedigree and genotyping quality analyses of over 10,000 DNA samples from the Generation Scotland: Scottish Family Health Study. *BMC Med. Genet.* **14**, 38 (2013).
133. Katsanis, S. H. & Katsanis, N. Molecular genetic testing and the future of clinical genomics. *Nat. Rev. Genet.* **14**, 415–426 (2013).
134. Srebniak, M. I. et al. Prenatal SNP array testing in 1000 fetuses with ultrasound anomalies: causative, unexpected and susceptibility CNVs. *Eur. J. Hum. Genet.* **24**, 645–651 (2016).
135. Treff, N. R. et al. Single nucleotide polymorphism microarray-based concurrent screening of 24-chromosome aneuploidy and unbalanced translocations in preimplantation human embryos. *Fertil. Steril.* **95**, 1606–1612 (2011).
136. Treff, N. R. et al. A novel single-cell DNA fingerprinting method successfully distinguishes sibling human embryos. *Fertil. Steril.* **94**, 477–484 (2010).
137. Rosenfeld, J. A. et al. Diagnostic utility of microarray testing in pregnancy loss. *Ultrasound Obstet. Gynecol.* **46**, 478–486 (2015).
138. Monzon, F. A. et al. Whole genome SNP arrays as a potential diagnostic tool for the detection of characteristic chromosomal aberrations in renal epithelial tumors. *Mod. Pathol.* **21**, 599–608 (2008).
139. Deng, W. Q. & Pare, G. A fast algorithm to optimize SNP prioritization for gene–gene and gene–environment interactions. *Genet. Epidemiol.* **35**, 729–738 (2011).
140. Bonnefond, A. et al. Molecular diagnosis of neonatal diabetes mellitus using next-generation sequencing of the whole exome. *PLOS ONE* **5**, e13630 (2010).
141. Speed, D. et al. Reevaluation of SNP heritability in complex human traits. *Nat. Genet.* **49**, 986–992 (2017).
142. Zhang, Y., Qi, G., Park, J. H. & Chatterjee, N. Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. *Nat. Genet.* **50**, 1318–1326 (2018).
143. Park, J. H. et al. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat. Genet.* **42**, 570–575 (2010).
144. Xiao, Y., Segal, M. R., Yang, Y. H. & Yeh, R. F. A multi-array multi-SNP genotyping algorithm for Affymetrix SNP microarrays. *Bioinformatics* **23**, 1459–1467 (2007).
145. Korn, J. M. et al. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat. Genet.* **40**, 1253–1260 (2008).
146. Li, G., Gelernter, J., Kranzler, H. R. & Zhao, H. M(3): an improved SNP calling algorithm for Illumina BeadArray data. *Bioinformatics* **28**, 358–365 (2012).
147. Goldstein, J. I. et al. zCall: a rare variant caller for array-based genotyping: genetics and population analysis. *Bioinformatics* **28**, 2543–2545 (2012).
148. Shah, T. S. et al. optiCall: a robust genotype-calling algorithm for rare, low-frequency and common variants. *Bioinformatics* **28**, 1598–1603 (2012).
149. Liu, R., Dai, Z., Yeager, M., Irizarry, R. A. & Ritchie, M. E. KRLMM: an adaptive genotype calling method for common and low frequency variants. *BMC Bioinformatics* **15**, 158 (2014).
150. Winchester, L., Yau, C. & Ragoussis, J. Comparing CNV detection methods for SNP arrays. *Brief. Funct. Genomic Proteomic* **8**, 353–366 (2009).
151. Coin, L. J. et al. cnvHap: an integrative population and haplotype-based multiplatform model of SNPs and CNVs. *Nat. Methods* **7**, 541–546 (2010).
152. Hauser, E., Cremer, N., Hein, R. & Deshmukh, H. Haplotype-based analysis: a summary of GAW16 Group 4 analysis. *Genet. Epidemiol.* **33**, S24–S28 (2009).
153. Nielsen, R., Paul, J. S., Albrechtsen, A. & Song, Y. S. Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.* **12**, 443–451 (2011).
154. Das, S., Abecasis, G. R. & Browning, B. L. Genotype imputation from large reference panels. *Annu. Rev. Genomics Hum. Genet.* **19**, 73–96 (2018).
155. 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).
156. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
157. Loh, P. R. et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).
158. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39**, 906–913 (2007).
159. Turner, S. D. qqman: an R package for visualizing GWAS results using QQ and manhattan plots. Preprint at *bioRxiv* https://www.biorxiv.org/content/10.1101/005165v1 (2014).
160. Pruim, R. J. et al. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26**, 2336–2337 (2010).
161. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
162. Pe'er, I., Yelensky, R., Altshuler, D. & Daly, M. J. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet. Epidemiol.* **32**, 381–385 (2008).
163. Pulit, S. L., de With, S. A. & de Bakker, P. I. Resetting the bar: statistical significance in whole-genome sequencing-based association studies of global populations. *Genet. Epidemiol.* **41**, 145–151 (2017).
164. Pasaniuc, B. & Price, A. L. Dissecting the genetics of complex traits using summary association statistics. *Nat. Rev. Genet.* **18**, 117–127 (2017).
165. 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 (2012).
166. Xue, A. et al. Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. *Nat. Commun.* **9**, 2941 (2018).
167. Pulit, S. L. et al. Meta-analysis of genome-wide association studies for body fat distribution in 694,649 individuals of European ancestry. *Hum. Mol. Genet.* **28**, 166–174 (2019).
168. Shi, H., Kichaev, G. & Pasaniuc, B. Contrasting the genetic architecture of 30 complex traits from summary association data. *Am. J. Hum. Genet.* **99**, 139–153 (2016).
169. Bulik-Sullivan, B. K. et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
170. Pickrell, J. K. et al. Detection and interpretation of shared genetic influences on 42 human traits. *Nat. Genet.* **48**, 709–717 (2016).
171. Vilhjalmsson, B. J. et al. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.* **97**, 576–592 (2015).
172. Kichaev, G. & Pasaniuc, B. Leveraging functional-annotation data in trans-ethnic fine-mapping studies. *Am. J. Hum. Genet.* **97**, 260–271 (2015).
173. Jensen, P. B., Jensen, L. J. & Brunak, S. Mining electronic health records: towards better research applications and clinical care. *Nat. Rev. Genet.* **13**, 395–405 (2012).
174. Bush, W. S., Oetjens, M. T. & Crawford, D. C. Unravelling the human genome-phenome relationship using phenome-wide association studies. *Nat. Rev. Genet.* **17**, 129–145 (2016).
175. Sudlow, C. et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Med.* **12**, e1001779 (2015).
176. Sanchez-Roige, S. et al. Genome-wide association study of delay discounting in 23,217 adult research participants of European ancestry. *Nat. Neurosci.* **21**, 16–18 (2018).
177. Nagel, M. et al. Meta-analysis of genome-wide association studies for neuroticism in 449,484 individuals identifies novel genetic loci and pathways. *Nat. Genet.* **50**, 920–927 (2018).
178. Flannick, J. & Florez, J. C. Type 2 diabetes: genetic data sharing to advance complex disease research. *Nat. Rev. Genet.* **17**, 535–549 (2016).
179. McAllister, E. J. et al. Ten putative contributors to the obesity epidemic. *Crit. Rev. Food Sci. Nutr.* **49**, 868–913 (2009).
180. Pigeyre, M., Yazdi, F. T., Kaur, Y. & Meyre, D. Recent progress in genetics, epigenetics and metagenomics unveils the pathophysiology of human obesity. *Clin. Sci.* **130**, 943–986 (2016).
181. Reddon, H., Gueant, J. L. & Meyre, D. The importance of gene–environment interactions in human obesity. *Clin. Sci.* **130**, 1571–1597 (2016).
182. Locke, A. E. et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).
183. Müller, M. J. et al. The case of GWAS of obesity: does body weight control play by the rules? *Int. J. Obes.* **42**, 1395–1405 (2018).
184. Felix, J. F. et al. Genome-wide association analysis identifies three new susceptibility loci for childhood body mass index. *Hum. Mol. Genet.* **25**, 389–403 (2016).
185. Winkler, T. W. et al. The influence of age and sex on genetic associations with adult body size and shape: a large-scale genome-wide interaction study. *PLOS Genet.* **11**, e1005378 (2015).
186. Warrington, N. M. et al. A genome-wide association study of body mass index across early life and childhood. *Int. J. Epidemiol.* **44**, 700–712 (2015).
187. Yu, H. et al. Genome-wide association study suggested the PTPRD polymorphisms were associated with weight gain effects of atypical antipsychotic medications. *Schizophr. Bull.* **42**, 814–823 (2016).
188. Taylor, A. E. et al. Stratification by smoking status reveals an association of CHRNA5-A3-B4 genotype with body mass index in never smokers. *PLOS Genet.* **10**, e1004799 (2014).
189. Hatoum, I. J. et al. Weight loss after gastric bypass is associated with a variant at 15q26.1. *Am. J. Hum. Genet.* **92**, 827–834 (2013).
190. McCaffery, J. M. et al. Human cardiovascular disease IBC chip-wide association with weight loss and weight regain in the look AHEAD trial. *Hum. Hered.* **75**, 160–174 (2013).
191. Yang, J. et al. FTO genotype is associated with phenotypic variability of body mass index. *Nature* **490**, 267–272 (2012).
192. Meyre, D. et al. Genome-wide association study for early-onset and morbid adult obesity identifies three new risk loci in European populations. *Nat. Genet.* **41**, 157–159 (2009).
193. Berndt, S. I. et al. Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. *Nat. Genet.* **45**, 501–512 (2013).
194. Blakemore, A. I. et al. A rare variant in the visfatin gene (NAMPT/PBEF1) is associated with protection from obesity. *Obesity* **17**, 1549–1553 (2009).
195. Liu, Y. J. et al. Genome-wide association scans identified CTNNBL1 as a novel gene for obesity. *Hum. Mol. Genet.* **17**, 1803–1813 (2008).
196. Lu, Y. et al. New loci for body fat percentage reveal link between adiposity and cardiometabolic disease risk. *Nat. Commun.* **7**, 10495 (2016).
197. Tanaka, T. et al. Genome-wide meta-analysis of observational studies shows common genetic variants associated with macronutrient intake. *Am. J. Clin. Nutr.* **97**, 1395–1402 (2013).
198. De Moor, M. H. et al. Genome-wide association study of exercise behavior in Dutch and American adults. *Med. Sci. Sports Exerc.* **41**, 1887–1895 (2009).
199. Lane, J. M. et al. Genome-wide association analyses of sleep disturbance traits identify new loci and highlight shared genetics with neuropsychiatric and metabolic traits. *Nat. Genet.* **49**, 274–281 (2017).
200. Kilpelainen, T. O. et al. Genome-wide meta-analysis uncovers novel loci influencing circulating leptin levels. *Nat. Commun.* **7**, 10494 (2016).
201. Sun, Q. et al. Genome-wide association study identifies polymorphisms in LEPR as determinants of plasma soluble leptin receptor levels. *Hum. Mol. Genet.* **19**, 1846–1855 (2010).
202. Dastani, Z. et al. Novel loci for adiponectin levels and their influence on type 2 diabetes and metabolic traits: a multi-ethnic meta-analysis of 45,891 individuals. *PLOS Genet.* **8**, e1002607 (2012).
203. Ried, J. S. et al. A principal component meta-analysis on multiple anthropometric traits identifies novel loci for body shape. *Nat. Commun.* **7**, 13357 (2016).
204. Freimer, N. B. & Mohr, D. C. Integrating behavioural health tracking in human genetics research. *Nat. Rev. Genet.* **20**, 129–130 (2019).
205. Tam, V., Turcotte, M. & Meyre, D. Established and emerging strategies to crack the genetic code of obesity. *Obes. Rev.* **20**, 212–240 (2019).
206. Yengo, L. et al. Detection and quantification of inbreeding depression for complex traits from SNP data. *Proc. Natl Acad. Sci. USA* **114**, 8602–8607 (2017).
207. Medina-Gomez, C. et al. Challenges in conducting genome-wide association studies in highly admixed multi-ethnic populations: the Generation R Study. *Eur. J. Epidemiol.* **30**, 317–330 (2015).
208. Li, Y. R. & Keating, B. J. Trans-ethnic genome-wide association studies: advantages and challenges of mapping in diverse populations. *Genome Med.* **6**, 91 (2014).
209. Morris, A. P. Transethnic meta-analysis of genomewide association studies. *Genet. Epidemiol.* **35**, 809–822 (2011).

210. Wood, A. R. et al. Variants in the FTO and CDKAL1 loci have recessive effects on risk of obesity and type 2 diabetes, respectively. *Diabetologia* **59**, 1214–1221 (2016).
211. Wermter, A. K. et al. Preferential reciprocal transfer of paternal/maternal DLK1 alleles to obese children: first evidence of polar overdominance in humans. *Eur. J. Hum. Genet.* **16**, 1126–1134 (2008).
212. Joo, J., Kwak, M., Ahn, K. & Zheng, G. A robust genome-wide scan statistic of the Wellcome Trust Case-Control Consortium. *Biometrics* **65**, 1115–1122 (2009).
213. 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).
214. Tukiainen, T. et al. Chromosome X-wide association study identifies loci for fasting insulin and height and evidence for incomplete dosage compensation. *PLOS Genet.* **10**, e1004127 (2014).
215. Bush, W. S. & Moore, J. H. Chapter 11: genome-wide association studies. *PLOS Comput. Biol.* **8**, e1002822 (2012).
216. Manning, A. K. et al. Meta-analysis of gene-environment interaction: joint estimation of SNP and SNP × environment regression coefficients. *Genet. Epidemiol.* **35**, 11–18 (2011).
217. Aschard, H., Hancock, D. B., London, S. J. & Kraft, P. Genome-wide meta-analysis of joint tests for genetic and gene–environment interaction effects. *Hum. Hered.* **70**, 292–300 (2011).
218. Liu, J. Z. et al. A versatile gene-based test for genome-wide association studies. *Am. J. Hum. Genet.* **87**, 139–145 (2010).
219. Mishra, A. & Macgregor, S. VEGAS2: software for more flexible gene-based testing. *Twin Res. Hum. Genet.* **18**, 86–91 (2015).
220. Stephens, M. & Balding, D. J. Bayesian statistical methods for genetic association studies. *Nat. Rev. Genet.* **10**, 681–690 (2009).
221. Szymczak, S. et al. Machine learning in genome-wide association studies. *Genet. Epidemiol.* **33**, S51–S57 (2009).
222. Li, A. & Meyre, D. Challenges in reproducibility of genetic association studies: lessons learned from the obesity field. *Int. J. Obes.* **37**, 559–567 (2013).
223. Bhattacharjee, S. et al. A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. *Am. J. Hum. Genet.* **90**, 821–835 (2012).
224. Anderson, C. A. et al. Data quality control in genetic case–control association studies. *Nat. Protoc.* **5**, 1564–1573 (2010).
225. Winkler, T. W. et al. Quality control and conduct of genome-wide association meta-analyses. *Nat. Protoc.* **9**, 1192–1212 (2014).
226. Muir, P. et al. The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biol.* **17**, 53 (2016).
227. Richter, B. G. & Sexton, D. P. Managing and analyzing next-generation sequence data. *PLOS Comput. Biol.* **5**, e1000369 (2009).
228. Mardis, E. R. The $1,000 genome, the $100,000 analysis? *Genome Med.* **2**, 84 (2010).
229. Dudbridge, F. & Gusnanto, A. Estimation of significance thresholds for genomewide association scans. *Genet. Epidemiol.* **32**, 227–234 (2008).
230. Hatzikotoulas, K., Gilly, A. & Zeggini, E. Using population isolates in genetic association studies. *Brief. Funct. Genomics* **13**, 371–377 (2014).
231. Hagg, S. et al. Gene-based meta-analysis of genome-wide association studies implicates new loci involved in obesity. *Hum. Mol. Genet.* **24**, 6849–6860 (2015).
232. Liu, Y. J. et al. Biological pathway-based genome-wide association analysis identified the vasoactive intestinal peptide (VIP) pathway important for obesity. *Obesity* **18**, 2339–2346 (2010).
233. Johansson, A. et al. Linkage and genome-wide association analysis of obesity-related phenotypes: association of weight with the MGAT1 gene. *Obesity* **18**, 803–808 (2010).
234. Du, H. et al. Genetic polymorphisms in the hypothalamic pathway in relation to subsequent weight change — the DiOGenes study. *PLOS ONE* **6**, e17436 (2011).
235. Greenawalt, D. M. et al. A survey of the genetics of stomach, liver, and adipose gene expression from a morbidly obese cohort. *Genome Res.* **21**, 1008–1016 (2011).
236. Grossman, S. R. et al. A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science* **327**, 883–886 (2010).
237. Ioannidis, J. P. Why most published research findings are false. *PLOS Med.* **2**, e124 (2005).
238. Manolio, T. A. Bringing genome-wide association findings into clinical use. *Nat. Rev. Genet.* **14**, 549–558 (2013).
239. Eichler, E. E. et al. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.* **11**, 446–450 (2010).
240. Manolio, T. A. Genomewide association studies and assessment of the risk of disease. *N. Engl. J. Med.* **363**, 166–176 (2010).
241. Pare, G., Asma, S. & Deng, W. Q. Contribution of large region joint associations to complex traits genetics. *PLOS Genet.* **11**, e1005103 (2015).
This paper demonstrates the contribution of joint association of multiple weakly associated variants over large chromosomal regions to complex traits.
242. Frazer, K. A., Murray, S. S., Schork, N. J. & Topol, E. J. Human genetic variation and its contribution to complex traits. *Nat. Rev. Genet.* **10**, 241–251 (2009).
243. Aschard, H. et al. Inclusion of gene–gene and gene–environment interactions unlikely to dramatically improve risk prediction for complex diseases. *Am. J. Hum. Genet.* **90**, 962–972 (2012).
244. Choquet, H. & Meyre, D. Genetics of obesity: what have we learned? *Curr. Genomics* **12**, 169–179 (2011).
245. Zuk, O., Hechter, E., Sunyaev, S. R. & Lander, E. S. The mystery of missing heritability: genetic interactions create phantom heritability. *Proc. Natl Acad. Sci. USA* **109**, 1193–1198 (2012).
246. Zaitlen, N. et al. Leveraging population admixture to characterize the heritability of complex traits. *Nat. Genet.* **46**, 1356–1362 (2014).
247. Mayhew, A. J. & Meyre, D. Assessing the heritability of complex traits in humans: methodological challenges and opportunities. *Curr. Genomics* **18**, 332–340 (2017).
248. Hindorff, L. A. et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA* **106**, 9362–9367 (2009).
249. Mahajan, A. et al. Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes. *Nat. Genet.* **50**, 559–571 (2018).
250. McCarthy, M. I. et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* **9**, 356–369 (2008).
251. Edwards, S. L., Beesley, J., French, J. D. & Dunning, A. M. Beyond GWASs: illuminating the dark road from association to function. *Am. J. Hum. Genet.* **93**, 779–797 (2013).
252. Stryjecki, C., Alyass, A. & Meyre, D. Ethnic and population differences in the genetic predisposition to human obesity. *Obes. Rev.* **19**, 62–80 (2018).
253. Magi, R. et al. Trans-ethnic meta-regression of genome-wide association studies accounting for ancestry increases power for discovery and improves fine-mapping resolution. *Hum. Mol. Genet.* **26**, 3639–3650 (2017).
254. Gaulton, K. J. et al. Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. *Nat. Genet.* **47**, 1415–1425 (2015).
255. Thurner, M. et al. Integration of human pancreatic islet genomic data refines regulatory mechanisms at type 2 diabetes susceptibility loci. *eLife* **7**, e31977 (2018).
256. Ng, M. C. Y. et al. Discovery and fine-mapping of adiposity loci using high density imputation of genome-wide association studies in individuals of African ancestry: African Ancestry Anthropometry Genetics Consortium. *PLOS Genet.* **13**, e1006719 (2017).
257. [No authors listed.] Freely associating. *Nat. Genet.* **22**, 1–2 (1999).
258. Wang, K. et al. Interpretation of association signals and identification of causal variants from genome-wide association studies. *Am. J. Hum. Genet.* **86**, 730–742 (2010).
259. Holm, H. et al. A rare variant in MYH6 is associated with high risk of sick sinus syndrome. *Nat. Genet.* **43**, 316–320 (2011).
260. Thun, G. A. et al. Causal and synthetic associations of variants in the SERPINA gene cluster with α1-antitrypsin serum levels. *PLOS Genet.* **9**, e1003585 (2013).
261. Wray, N. R., Purcell, S. M. & Visscher, P. M. Synthetic associations created by rare variants do not explain most GWAS results. *PLOS Biol.* **9**, e1000579 (2011).
262. Scherag, A. et al. Investigation of a genome wide association signal for obesity: synthetic association and haplotype analyses at the melanocortin 4 receptor gene locus. *PLOS ONE* **5**, e13967 (2010).
263. Creemers, J. W. et al. Heterozygous mutations causing partial prohormone convertase 1 deficiency contribute to human obesity. *Diabetes* **61**, 383–390 (2012).
264. Voight, B. F. et al. The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLOS Genet.* **8**, e1002793 (2012).
265. Cortes, A. & Brown, M. A. Promise and pitfalls of the Immunochip. *Arthritis Res. Ther.* **13**, 101 (2011).
266. Gong, J. et al. Fine mapping and identification of BMI loci in African Americans. *Am. J. Hum. Genet.* **93**, 661–671 (2013).
267. Bahcall, O. G. iCOGS collection provides a collaborative model. Foreword. *Nat. Genet.* **45**, 343 (2013).
268. Onengut-Gumuscu, S. et al. Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. *Nat. Genet.* **47**, 381–386 (2015).
269. Ghoussaini, M. et al. Evidence that breast cancer risk at the 2q35 locus is mediated through IGFBP5 regulation. *Nat. Commun.* **4**, 4999 (2014).
270. Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* **19**, 491–504 (2018).
271. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
272. Kundaje, A. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
273. Andersson, R. et al. An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
274. GTEx Consortium. The genotype-tissue expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
275. Boyle, A. P. et al. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* **22**, 1790–1797 (2012).
276. Ward, L. D. & Kellis, M. 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–D881 (2016).
277. 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).
278. Gallagher, M. D. & Chen-Plotkin, A. S. The post-GWAS era: from association to function. *Am. J. Hum. Genet.* **102**, 717–730 (2018).
279. Denker, A. & de Laat, W. The second decade of 3C technologies: detailed insights into nuclear organization. *Genes Dev.* **30**, 1357–1382 (2016).
280. Gusev, A. et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).
281. Zhu, Z. et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).
282. Church, C. et al. Overexpression of Fto leads to increased food intake and results in obesity. *Nat. Genet.* **42**, 1086–1092 (2010).
283. Fischer, J. et al. Inactivation of the Fto gene protects from obesity. *Nature* **458**, 894–898 (2009).
284. Smemo, S. et al. Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature* **507**, 371–375 (2014).
285. Claussnitzer, M. et al. FTO obesity variant circuitry and adipocyte browning in humans. *N. Engl. J. Med.* **373**, 895–907 (2015).
286. Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics era — concepts and misconceptions. *Nat. Rev. Genet.* **9**, 255–266 (2008).
287. Witte, J. S., Visscher, P. M. & Wray, N. R. The contribution of genetic variants to disease depends on the ruler. *Nat. Rev. Genet.* **15**, 765–776 (2014).
288. Buchner, D. A. & Nadeau, J. H. Contrasting genetic architectures in different mouse reference populations used for studying complex traits. *Genome Res.* **25**, 775–791 (2015).
289. Mackay, T. F. Epistasis and quantitative traits: using model organisms to study gene–gene interactions. *Nat. Rev. Genet.* **15**, 22–33 (2014).
290. Wei, W. H., Hemani, G. & Haley, C. S. Detecting epistasis in human complex traits. *Nat. Rev. Genet.* **15**, 722–733 (2014).

291. Okada, Y. et al. Common variants at CDKAL1 and KLF9 are associated with body mass index in East Asian populations. *Nat. Genet.* **44**, 302–306 (2012).
292. Cortes, A. et al. Major histocompatibility complex associations of ankylosing spondylitis are complex and involve further epistasis with ERAP1. *Nat. Commun.* **6**, 7146 (2015).
293. Wang, K., Bucan, M., Grant, S. F., Schellenberg, G. & Hakonarson, H. Strategies for genetic studies of complex diseases. *Cell* **142**, 351–353 (2010).
294. Liu, L., Zhang, D., Liu, H. & Arendt, C. Robust methods for population stratification in genome wide association studies. *BMC Bioinformatics* **14**, 132 (2013).
295. Hinney, A. et al. Genome wide association (GWA) study for early onset extreme obesity supports the role of fat mass and obesity associated gene (FTO) variants. *PLOS ONE* **2**, e1361 (2007).
296. Wang, K. et al. Common genetic variants on 5p14.1 associate with autism spectrum disorders. *Nature* **459**, 528–533 (2009).
297. Ma, D. et al. A genome-wide association study of autism reveals a common novel risk locus at 5p14.1. *Ann. Hum. Genet.* **73**, 263–273 (2009).
298. Loos, R. J. F. & Janssens, A. Predicting polygenic obesity using genetic information. *Cell Metab.* **25**, 535–543 (2017).
299. Janssens, A. C. & van Duijn, C. M. Genome-based prediction of common diseases: advances and prospects. *Hum. Mol. Genet.* **17**, R166–R173 (2008).
300. Janssens, A. C. et al. The impact of genotype frequencies on the clinical validity of genomic profiling for predicting common chronic diseases. *Genet. Med.* **9**, 528–535 (2007).
301. Stutzmann, F. et al. Non-synonymous polymorphisms in melanocortin-4 receptor protect against obesity: the two facets of a Janus obesity gene. *Hum. Mol. Genet.* **16**, 1837–1844 (2007).
302. Ichimura, A. et al. Dysfunction of lipid sensor GPR120 leads to obesity in both mouse and human. *Nature* **483**, 350–354 (2012).
303. Challis, B. G. et al. A missense mutation disrupting a dibasic prohormone processing site in pro-opiomelanocortin (POMC) increases susceptibility to early-onset obesity through a novel molecular mechanism. *Hum. Mol. Genet.* **11**, 1997–2004 (2002).
304. Bonnefond, A. et al. Eating behavior, low-frequency functional mutations in the melanocortin-4 receptor (MC4R) gene, and outcomes of bariatric operations: a 6-year prospective study. *Diabetes Care* **39**, 1384–1392 (2016).
305. Kuhnen, P. et al. Proopiomelanocortin deficiency treated with a melanocortin-4 receptor agonist. *N. Engl. J. Med.* **375**, 240–246 (2016).
306. Farooqi, I. S. et al. Effects of recombinant leptin therapy in a child with congenital leptin deficiency. *N. Engl. J. Med.* **341**, 879–884 (1999).
307. Collet, T. H. et al. Evaluation of a melanocortin-4 receptor (MC4R) agonist (setmelanotide) in MC4R deficiency. *Mol. Metab.* **6**, 1321–1329 (2017).
308. Clement, K. et al. MC4R agonism promotes durable weight loss in patients with leptin receptor deficiency. *Nat. Med.* **24**, 551–555 (2018).
309. Manning, A. et al. A low-frequency inactivating AKT2 variant enriched in the Finnish population is associated with fasting insulin levels and type 2 diabetes risk. *Diabetes* **66**, 2019–2032 (2017).
310. Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* **19**, 581–590 (2018).
This is a recent in-depth review on the emerging personal and clinical utility of PRSs.
311. Gronberg, H. et al. Prostate cancer screening in men aged 50–69 years (STHLM3): a prospective population-based diagnostic study. *Lancet Oncol.* **16**, 1667–1676 (2015).
312. Maas, P. et al. Breast cancer risk from modifiable and nonmodifiable risk factors among white women in the United States. *JAMA Oncol.* **2**, 1295–1302 (2016).
313. Khera, A. V. et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224 (2018).
This paper demonstrates the utility of using PRSs to identify individuals with a level of risk equivalent to that of rare monogenic mutations.
314. Theriault, S. et al. Polygenic contribution in individuals with early-onset coronary artery disease. *Circ. Genom. Precis. Med.* **11**, e001849 (2018).
315. Natarajan, P. et al. Polygenic risk score identifies subgroup with higher burden of atherosclerosis and greater relative benefit from statin therapy in the primary prevention setting. *Circulation* **135**, 2091–2101 (2017).
316. Mega, J. L. et al. Genetic risk, coronary heart disease events, and the clinical benefit of statin therapy: an analysis of primary and secondary prevention trials. *Lancet* **385**, 2264–2271 (2015).
317. Meisel, S. F., Beeken, R. J., van Jaarsveld, C. H. & Wardle, J. Genetic susceptibility testing and readiness to control weight: results from a randomized controlled trial. *Obesity* **23**, 305–312 (2015).
318. Meisel, S. F., Walker, C. & Wardle, J. Psychological responses to genetic testing for weight gain: a vignette study. *Obesity* **20**, 540–546 (2012).
319. Xing, C. et al. Evaluation of power of the Illumina HumanOmni5M-4v1 BeadChip to detect risk variants for human complex diseases. *Eur. J. Hum. Genet.* **24**, 1029–1034 (2016).
320. Abecasis, G. R. et al. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
321. Rosenberg, N. A. et al. Genome-wide association studies in diverse populations. *Nat. Rev. Genet.* **11**, 356–366 (2010).
322. Hoffmann, T. J. et al. Design and coverage of high throughput genotyping arrays optimized for individuals of East Asian, African American, and Latino race/ethnicity using imputation and a novel hybrid SNP selection algorithm. *Genomics* **98**, 422–430 (2011).
323. Southam, L. et al. Whole genome sequencing and imputation in isolated populations identify genetic associations with medically-relevant complex traits. *Nat. Commun.* **8**, 15606 (2017).
324. Mitt, M. et al. Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. *Eur. J. Hum. Genet.* **25**, 869–876 (2017).
325. Tachmazidou, I. et al. Whole-genome sequencing coupled to imputation discovers genetic signals for anthropometric traits. *Am. J. Hum. Genet.* **100**, 865–884 (2017).
326. Pigeyre, M. & Meyre, D. in *Pediatric Obesity: Etiology, Pathogenesis and Treatment* 2nd edn (ed. Freemark, M.) 135–152 (Humana Press, 2018).
327. Bonnefond, A. & Froguel, P. Rare and common genetic events in type 2 diabetes: what should biologists know? *Cell Metab.* **21**, 357–368 (2015).
328. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
329. Hendricks, A. E. et al. Rare variant analysis of human and rodent obesity genes in individuals with severe childhood obesity. *Sci. Rep.* **7**, 4394 (2017).
330. Steinthorsdottir, V. et al. Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. *Nat. Genet.* **46**, 294–298 (2014).
331. Chatterjee, N., Shi, J. & Garcia-Closas, M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat. Rev. Genet.* **17**, 392–406 (2016).
332. Scott, R. A. et al. An expanded genome-wide association study of type 2 diabetes in Europeans. *Diabetes* **66**, 2888–2902 (2017).
333. Loos, R. J. et al. Common variants near MC4R are associated with fat mass, weight and risk of obesity. *Nat. Genet.* **40**, 768–775 (2008).
334. Thorleifsson, G. et al. Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. *Nat. Genet.* **41**, 18–24 (2009).
335. Weedon, M. N. et al. A common variant of HMGA2 is associated with adult and childhood height in the general population. *Nat. Genet.* **39**, 1245–1250 (2007).
336. Gudbjartsson, D. F. et al. Many sequence variants affecting diversity of adult human height. *Nat. Genet.* **40**, 609–615 (2008).
337. Weedon, M. N. et al. Genome-wide association analysis identifies 20 loci that influence adult height. *Nat. Genet.* **40**, 575–583 (2008).
338. Lettre, G. et al. Identification of ten loci associated with height highlights new biological pathways in human growth. *Nat. Genet.* **40**, 584–591 (2008).
339. Soranzo, N. et al. Meta-analysis of genome-wide scans for human adult stature identifies novel loci and associations with measures of skeletal frame size. *PLoS Genet.* **5**, e1000445 (2009).
340. Lango Allen, H. et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–838 (2010).
341. 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).
342. Heid, I. M. et al. Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. *Nat. Genet.* **42**, 949–960 (2010).
343. Shungin, D. New genetic loci link adipose and insulin biology to body fat distribution. *Nature* **518**, 187–196 (2015).
344. Lotta, L. A. et al. Association of genetic variants related to gluteofemoral vs abdominal fat distribution with type 2 diabetes, coronary disease, and cardiovascular risk factors. *JAMA* **320**, 2553–2563 (2018).
345. Jónsson, H. et al. Whole genome characterization of sequence diversity of 15,220 Icelanders. *Sci. Data* **4**, 170115 (2017).
346. International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).
347. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
348. UK10K Consortium. The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015).
This is one of the first large-scale attempts to use WGS to identify low-frequency and rare variants associated with common diseases and traits in the general population.

Acknowledgements
D.M. holds a Canada Research Chair in Genetics of Obesity. Y.B. holds a Canada Research Chair in Genomics of Heart and Lung Diseases. G.P. holds the Canada Research Chair in Genetic and Molecular Epidemiology.

Author contributions
V.T., M.T. and D.M. researched the literature. V.T., M.T., Y.B., G.P. and D.M. provided substantial contributions to discussion of the content. V.T., N.P., Y.B. and D.M. wrote the article. M.T., Y.B., G.P. and D.M. reviewed and/or edited the manuscript before submission.

Competing interests
The authors declare no competing interests.

Publisher's note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Reviewer information
Nature Reviews Genetics thanks S. Chanock, J. Florez and M. Nelson for their contribution to the peer review of this work.