Benefits and limitations of genome-wide association studies

Vivian Tam¹, Nikunj Patel¹, Michelle Turcotte¹, Yohan Bossé²,³, Guillaume Paré¹,⁴ and David Meyre¹,⁴,⁵ *

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.

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 decade⁶–⁸. Since the first GWAS for age-related macular degeneration (AMD) was published in 2005 (REF⁹), more than 50,000 associations of genome-wide significance ($P<5 \times 10^{-8}$) have been reported between genetic variants and common diseases and traits¹. 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 GWAS¹⁰, this study design has not been without controversy¹¹–¹³. Critics of GWAS have contended that single-nucleotide variants (SNVs) identified in GWAS explain only a small fraction of the heritability of complex traits¹⁴, may represent spurious associations¹⁵ and do not necessarily pinpoint causal variants and genes¹⁶, and that GWAS will yield too many loci (if SNVs in all genes are implicated, then this would be uninformative)¹⁷,¹⁸. 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)¹⁹–²¹. Unsurprisingly, these criticisms have led to scepticism among non-geneticians about the benefits of GWAS and hesitancy among national funding organizations to fund new GWAS²². 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 SNP–trait associations at a genome-wide significance threshold¹. GWAS have successfully identified risk loci for a vast number of diseases and traits, including anorexia nervosa¹³, major depressive disorder¹⁴, cancers and subtypes of cancers¹⁵–¹⁷, type 2...
diabetes mellitus (T2DM)\textsuperscript{17}, coronary artery disease\textsuperscript{18}, schizophrenia\textsuperscript{19}, inflammatory bowel disease\textsuperscript{20}, insomnia\textsuperscript{21}, body mass index (BMI)\textsuperscript{22} and educational attainment\textsuperscript{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\textsuperscript{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 variation.
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.

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, 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 [2]. Multi-locus analysis of AMD suscepti-
bility loci showed that 99% of the individuals with the highest-risk genotypes (including at CFH) had AMD; of these, 85% had advanced AMD [2]. 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 [2].

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 [4]. Rare loss-of-function mutations in HNF1A are known to cause a common form of MODY [4]. In 2008, two independent GWAS identified SNPs near the HNF1A gene associated with serum C-reactive protein (CRP) levels, a marker of inflammation [4]. 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 [4]. 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 [4].
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. 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.

GWAS have identified genetic variants that can be used to inform drug selection and dosage and prevent adverse drug reactions. 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 and genetic variants in SLCO1B1 that are associated with simvastatin-induced myopathy. 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 and for managing the T2DM risk allele.

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. 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, East Asian, Mexican, and Greenlandic 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 and Greenlandic populations.

### 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).

<table>
<thead>
<tr>
<th>Factor</th>
<th>SNP arrays</th>
<th>WGS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cost</td>
<td>Relatively inexpensive (~US$40 per sample)</td>
<td>Expensive (~US$1,000 per sample)</td>
</tr>
<tr>
<td>Reliability</td>
<td>Reliable, highly accurate technology</td>
<td>Less mature and less accurate technology</td>
</tr>
<tr>
<td>Genomic coverage</td>
<td>• Mainly restricted to common and low-frequency variants, although imputation of rare variants is increasingly accurate (ultra-rare variants, however, can never be identified)</td>
<td>From low-frequency, common variants to nearly all genetic variation in the genome, depending on the depth of sequencing</td>
</tr>
<tr>
<td>GWAS analysis</td>
<td>Well-established analytical pipeline and tools for data analysis</td>
<td>• Higher computational costs and greater analytical complexity</td>
</tr>
<tr>
<td>Other considerations</td>
<td>Custom genotyping arrays can be extremely cost-effective</td>
<td>• As all variation is ascertained, fine-mapping is easier</td>
</tr>
<tr>
<td>Suitable research objectives</td>
<td>• Analysing known or candidate associations in large cohorts</td>
<td>• Greater costs to store, process, analyse and interpret the resulting data</td>
</tr>
<tr>
<td></td>
<td>• Detecting low-frequency, common variant associations in extremely large sample sizes</td>
<td>• Detecting and fine-mapping rare variants</td>
</tr>
<tr>
<td></td>
<td>• Detecting ultra-rare risk variants when it becomes economically viable to perform WGS at a very large scale</td>
<td></td>
</tr>
</tbody>
</table>
and Europeans, but because the minor alleles are less common in Europeans, much larger sample sizes are needed to detect the association\(^6\). 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\(^6\). 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\(^6\). Ethnic differences in the effect size of a risk variant\(^6\) can also affect the likelihood of its discovery and its contribution to disease burden across populations\(^6\).

**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\(^7\), 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\(^8\). 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\(^9\), haematological traits\(^10\),\(^11\),\(^12\), blood pressure level\(^10\), height\(^9\), BMI\(^9\) and T2DM\(^9\).

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\(^11\). 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)\(^11\) and the generation of population-specific and study-specific reference panels\(^12\),\(^13\),\(^14\),\(^15\),\(^16\),\(^17\) 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\(^14\). This reference panel claims accurate imputation down to a MAF of 0.1%\(^17\). 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\(^15\).

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\(^7\). 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\(^18\),\(^19\). This observation led scientists to hypothesize that genes identified by GWAS may be relevant candidates for discovering rare disease-causing mutations\(^20\),\(^21\). 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)\(^22\), T1DM (IFIH1)\(^23\), T2DM (MTN1B, SLC30A8 and PPARG)\(^24\),\(^25\),\(^26\) and inflammatory bowel disease (TNFRSF68, PRDM1, CARD9, IL23R and RNF186)\(^27\),\(^28\),\(^29\) 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\(^30\),\(^31\).
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 copy number variants (CNVs) 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.

Furthermore, analyses of SNP array data from GWAS have identified clonal mosaicism in autosomes and sex chromosomes, which is associated with age and increased risk of haematological and solid tumours, as well as diseases such as T2DM. 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 haematoepoiesis as a predictor of haematological cancer.

Beyond CNVs and mosaic events, other classes of genetic variation, including haplotypes, variable number tandem repeats (VNTRs), retrotransposon insertion polymorphisms, insertions or deletions (indels) and inversions, 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, determination of ancestry and population substructure, fine-scale estimation of location of birth, genome-wide assessment of linkage disequilibrium, estimation of SNP heritability for complex traits, estimation of genetic correlations between traits, Mendelian randomization studies, polygenic risk scores, forensic analyses, determination of cryptic relatedness, paternity testing, direct-to-consumer genetic testing, clinical diagnostic genetic testing, prenatal and pre-implantation genetic diagnosis, embryonic DNA fingerprinting, determination of perinatal loss, loss-of-heterozygosity and CNV analyses in tumours (for example, for disease subtyping and classification), validation of new analytic methods and quality control of next-generation sequencing data (by comparing sequence-based variant calls to array-based genotyping).

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, whereas modelling the underlying genetic architecture of disease has shown that diseases have diverse genetic architectures, with consequent effects on rates of loci discovery. Psychiatric diseases and mental health traits, for example, seem to be mostly polygenic, involving a continuum of variants with small effects. By contrast, most other diseases involve clusters of SNPs with distinct magnitudes of effects. 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. 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.

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. New algorithms have also been designed specifically for calling low-frequency and rare variants and for inferring haplotypes and structural variants. With GWAS increasingly relying on WGS data, a corresponding host of tools has been developed for variant discovery and SNP calling from sequencing data. Similarly, improvements have been made in statistical imputation of genotypes. For data management and

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

<table>
<thead>
<tr>
<th>Reference panel</th>
<th>Number of reference samples</th>
<th>Ancestry of reference samples</th>
<th>Number of variant sites</th>
<th>Indels available</th>
<th>Refs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Icelandic reference panel</td>
<td>15,220</td>
<td>European (Icelandic)</td>
<td>31.1 million</td>
<td>Yes</td>
<td>145</td>
</tr>
<tr>
<td>HapMap Project phase 3</td>
<td>1,011</td>
<td>Multi-ethnic</td>
<td>1.4 million</td>
<td>No</td>
<td>146</td>
</tr>
<tr>
<td>1000G phase 1</td>
<td>1,092</td>
<td>Multi-ethnic</td>
<td>28.9 million</td>
<td>Yes</td>
<td>147</td>
</tr>
<tr>
<td>1000G phase 3</td>
<td>2,504</td>
<td>Multi-ethnic</td>
<td>81.7 million</td>
<td>Yes</td>
<td>81</td>
</tr>
<tr>
<td>UK10K Project</td>
<td>3,781</td>
<td>European</td>
<td>42.0 million</td>
<td>Yes</td>
<td>148</td>
</tr>
<tr>
<td>HRC</td>
<td>32,470</td>
<td>Predominantly European (includes the 1000G reference panel samples)</td>
<td>40.4 million</td>
<td>No</td>
<td>71</td>
</tr>
<tr>
<td>TOPMed</td>
<td>62,784</td>
<td>Multi-ethnic</td>
<td>463.0 million</td>
<td>Yes</td>
<td>85</td>
</tr>
</tbody>
</table>

1000G, 1000 Genomes; HRC, Haplotype Reference Consortium; indels, insertions or deletions; TOPMed, Trans-Omics for Precision Medicine. Figures are based on the latest status of the reference panel.
The process of localizing novel trait associations. For example, UK Biobank is still collecting genotype and phenotype information from 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, impulsivity and neuroticism.

Consensus among the genetics community to adopt a standardized significance threshold ($P < 5 \times 10^{-8}$) 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.

GWAS data are easily shared and publicly available data facilitates novel discoveries. GWAS data will continue to be useful for identifying novel traits 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. This wealth of data can be analysed in various ways to gain insight into the genetic basis of complex traits, for example, publicly available summary statistics have enabled discoveries of novel risk loci, estimates of SNP-based heritability, cross-trait analyses, polygenic risk prediction, and fine-mapping, 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.

A commitment by researchers to make summary associations publicly available summary statistics with linkage disequilibrium information, given the importance of these data in providing a foundation for a wide range of important analyses. 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.

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). 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.

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. Existing GWAS findings therefore represent the low-hanging fruit of GWAS discoveries, 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, BMI changes over time (that is, longitudinal data), BMI changes in response to obesity risk, or obesity-protective exposures. BMI variance, the extreme tails of BMI distribution, deep obesity phenotypes, obesity intermediate traits, and obesity composite traits. 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...
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. 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) 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. 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 under-represented in genetic association studies to date; these groups include Latin Americans, Native Americans, Indigenous Australians, Arabs, South Asians, Roma and Pacific Islanders. The study of isolated, that is, founder and highly consanguineous populations, as well as multi-ethnic and admixed groups will also be valuable.

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\textsuperscript{210}, dominant\textsuperscript{192}, overdominant\textsuperscript{211}, multiplicative\textsuperscript{12}, parent-of-origin-specific\textsuperscript{217} and X-linked inheritance\textsuperscript{11} models in GWAS can improve the statistical power to detect additional variants\textsuperscript{215}. In addition, GWAS accounting for gene–gene and gene–environment interactions\textsuperscript{216,217}, GWAS using genomic region-based or gene-based association tests\textsuperscript{218,219}, GWAS using Bayesian analyses\textsuperscript{220}, GWAS using machine learning approaches\textsuperscript{221}, GWAS in different study designs (for example, family-based, case–control–case, case-only, intervention and hypothesis-driven study designs)\textsuperscript{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\textsuperscript{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\textsuperscript{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\textsuperscript{225} and meta-analyses of GWAS\textsuperscript{226}. 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\textsuperscript{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\textsuperscript{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\textsuperscript{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\textsuperscript{230}.

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\textsuperscript{231}. 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\textsuperscript{232}. A similar challenge arises for GWAS in small isolated populations\textsuperscript{233}.

A second strategy is to reduce the number of tests performed. This can be achieved by using gene-based\textsuperscript{234} or pathway-based association tests\textsuperscript{235,236}, or by restricting analyses to candidate genomic regions, such as linkage regions\textsuperscript{237}, genes specifically expressed in an important tissue\textsuperscript{238}, genes showing differential expression patterns related to the disease\textsuperscript{239}, prioritized candidate genes\textsuperscript{240}, potentially damaging SNPs\textsuperscript{241} or SNPs harbouring evolutionary signatures\textsuperscript{242}. Combining supporting biological evidence with statistical significance also increases the probability that the result is a true positive\textsuperscript{243}. 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\textsuperscript{244}. Unlike frequentist methods, such as the Bonferroni correction, Bayesian methods consider the universe of possible hypotheses in the human genome\textsuperscript{245}. 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\textsuperscript{246,247}. Several reasons for the missing heritability have been proposed\textsuperscript{239,248,249}. One probable explanation is that SNPs of modest effect are missed because they do not reach the...
Reviews

or variants. Most association signals map to non-coding loci but makes it difficult to discern the causal variant. Disequilibrium facilitates the initial identification of a correlation of multiple genetic variants due to linkage studies. Shared environmental effects, especially in classical twin studies, for example, different subsets of genes are expected to play a role in obesity depending on the environment. For example, different subsets of genes might be involved in the pathogenesis of diseases for which susceptibility is highly influenced by genes.

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. Most association signals map to non-coding regions of the genome, for which biological interpretation is inherently challenging. 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. 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. However, akin to a transition from drought to flood, the number of discoveries in human genetics has rapidly accelerated over the past 13 years. For example, recent simulations have shown that 90,000–100,000 SNPs may be needed to explain 80% of the heritability of height. This means that a substantial fraction of all genes may contribute to variation of complex traits. In pointing at ‘everything’, the danger is that GWAS could point at ‘nothing’. 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. 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.

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. Although synthetic associations have been reported in the literature (for example, NOD2 (REF. 250), HBB (REF. 251), MYH6 (REF. 252) or SERPINA1 (REF. 253)), multiple lines of evidence suggest that they are rare, and studies have increasingly demonstrated support for independent contributions of rare and common variants at a single locus. 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. 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. Assumming 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. Last, the rapid development of publicly available databases of regulatory elements across a range of tissues and cells types (for example, ENCODE, Epigenome RoadMap, FANTOM5 (REF. 254) and GTEx (REF. 255), as well as tools for querying such databases (for example, RegulomeDB and HaploReg)), 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.

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. Chromosome conformation capture and its derivatives can be used to visualize the 3D organization of chromatin and are important methods for
determining target genes\textsuperscript{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\textsuperscript{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\textsuperscript{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\textsuperscript{284}. Additional work using mouse models confirmed IRX3 as a likely causal gene\textsuperscript{285}, and an additional target, IRX5, has also been identified\textsuperscript{283}.

**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\textsuperscript{28}. Another challenge lies in accurately estimating the heritability of complex traits\textsuperscript{285,286,287}.

**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\textsuperscript{290–292}. 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\textsuperscript{293,294–296}. 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\textsuperscript{297}. Furthermore, the loss of information caused by imperfect linkage disequilibrium between genotyped and causal variants is larger for interactions than for main effects\textsuperscript{298}. Recent methodological developments (for example, data filtering, Bayesian methods and artificial intelligence algorithms) may boost the identification of epistatic interactions in humans\textsuperscript{299}. 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\textsuperscript{2}, but is also a concern when studying recently admixed populations and variants with very small effect sizes\textsuperscript{3}.

Most GWAS signals (OR < 1.5) have been suggested to be attributable to cryptic population stratification\textsuperscript{3}. However, this view is probably too extreme owing to several reasons. First, efficient methods are used in GWAS to control for population stratification\textsuperscript{295,296}. Second, many GWAS loci are enriched for biologically relevant variants\textsuperscript{297,298} or localize in genes or pathways known or postulated to play a role in disease\textsuperscript{3}. Third, GWAS hits have been confirmed by family-based association tests, which are robust to population stratification\textsuperscript{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\textsuperscript{293}. 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\textsuperscript{3}.

**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\textsuperscript{290,292}. 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\textsuperscript{300} and may account for a significant fraction of complex trait heritability\textsuperscript{290}. They are also likely to be causal variants, owing to low linkage disequilibrium with other variants. For example, the obesity-associated low-frequency SNP rs26232 (p.Asn221Asp) in PCSK1 is not in strong linkage disequilibrium (\(r^2 > 0.8\)) with any other variant\textsuperscript{301}. 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)\textsuperscript{302}, rs6232 (p.Asn221Asp) in PCSK1 (MAF 3%, OR 1.34)\textsuperscript{303}, rs116454156 (p.Arg270His) in FFA4 (also known as GRP120) (MAF 1%, OR 1.62)\textsuperscript{304} and rs28932472 (p.Arg236Gly) in POMC (MAF 0.5%, OR 4)\textsuperscript{305}. 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.\textsuperscript{306}). As the number of identified disease-associated variants increases, their cumulative predictive value will undoubtedly gain in importance\textsuperscript{307}.

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\textsuperscript{298}. 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
oligogenic variants is more likely to lead to actionable treatments. 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). 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. For example, the value of a PRS in cancer as a tool for stratification in public health has been exemplified in several studies, 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). For coronary artery disease, PRSs have identified individuals with risk equivalent to rare monogenic mutations; these individuals constitute a substantially larger fraction of the population than do individuals with rare monogenic mutations and might derive greater benefit from early lifestyle interventions and initiation of statin therapy than individuals at lower genetic risk. 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.

**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. For example, early genome-wide SNP arrays were designed by selecting tag SNPs from reference panels of predominantly European populations. Because linkage disequilibrium patterns vary across ethnic groups, these arrays often provided poor coverage in non-European populations. 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. 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. 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. 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.

**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. Although empirical evidence suggests that much of the heritability of complex traits can be explained by common variants, ultra-rare variants are also expected to contribute. 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. 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. At the present time, statistical imputation using large reference population panels is reasonably accurate for variants as infrequent as 1 in 1,000 (REF.). 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.

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. 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.

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, 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 and are beginning to show promise in...
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\(^{311}\), 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

---

**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.


This paper provides an overview of the lessons learned from the past decade of GWAS.


This study may be considered the first GWAS to be published.


This reference provides a theoretical groundwork for the omicnic model of inheritance.


In this paper, the author is among the first to suggest that GWAS may eventually implicate most of the genome.


This is the largest GWAS published to date.


This study helps to largely resolve the ‘missing heritability’ problem by demonstrating that a large portion of the heritability can be explained by common SNPs.


60. This study is the first true GWAS for a complex disease that used SNP arrays with exhaustive coverage of the genome.
66. The authors of this paper lay out the theoretical basis for how to do a GWAS.


Liu, L., Zhang, D., Liu, H. & Arendt, C. Robust...  
Farooqi, I. S. et al. Effects of recombinant leptin...  
Kuhnen, P. et al. Proopiomelanocortin deficiency treated...  
Wang, K., Bucan, M., Grant, S. F., Schellenberg, G. & Hakonarson, H. Strategies for genetic studies of...  
Challis, B. G. et al. A missense mutation disrupting...  
Ichimura, A. et al. Dysfunction of lipid sensor GPR120...  
Stutzmann, F. et al. Non-synonymous polymorphisms...  
Precht, G.B. et al. Evaluation of a missense mutation disrupting...  
Ji, H. et al. Editing of exogenous DNA...  
Stutzmann, F. et al. Non-synonymous polymorphisms...  
Kobayashi, Y. et al. A melanocortin-4 receptor...  
Kuhn, P. et al. Proopiomelanocortin deficiency treated...  
Kruhipa, P. et al. Proopiomelanocortin deficiency treated...  
Kruhipa, P. et al. Proopiomelanocortin deficiency treated...  
Kruhipa, P. et al. Proopiomelanocortin deficiency treated...  
Manning, A. et al. A low-frequency inactivating AKT2 variant enriched in the Finnish population...  
Torkamani, A., Winningen, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores...  
Gröneberg, H. Prostate cancer screening in men aged 50–69 years (STHLM5): a prospective population-based diagnostic study...  
Maas, P. et al. Breast cancer risk from modifiable and nonmodifiable risk factors among white women in the United States...  
Khera, A. V. et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations...  
This paper demonstrates the utility of using PRSs to identify individuals with a level of risk equivalent to that of rare monogenic mutations...  
Therault, S. et al. Polyclinogen contribution in individuals with early-onset coronary artery disease...  
Natarajan, P. et al. Polyclinogen risk score identifies subgroup with higher burden of atherosclerotic and greater relative benefit from statin therapy in the primary prevention setting...  
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...  
Mesel, 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...  
Mesel, S. F., Walker, C. & Wardle, J. Psychological responses to genetic testing for weight gain: a vignette study...  
Xing, C. et al. Evaluation of the power of the Illumina HumanOmniSNV+1 BeadChip to detect risk variants for human complex diseases...  
Abecasis, G. R. A map of human genome variation from population-scale sequencing...  
Rosenberg, N. A. et al. Genome-wide association studies in diverse populations...  
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...  
Southam, L. et al. Whole genome sequencing and imputation in isolated populations identify genetic associations with medically-relevant complex traits...  
Mitt, M. et al. Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel...  
Tachmazidou, J. et al. Whole-genome sequencing coupled to imputation discovers genetic signals for anthropometric traits...  
Bonnefond, A. et al. Evaluation of a melanocortin-4 receptor (MC4R) agonist and antagonist in diabetic and obese patients...  
Clement, K. et al. MC4R agonism promotes durable weight loss in patients with leptin receptor deficiency...  
Hendricks, A. E. et al. Rare variant analysis of human and rodent obesity genes in individuals with severe childhood obesity...  
Steinthorodottir, V. et al. Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes...  
Scott, R. A. et al. An expanded genome-wide association study of type 2 diabetes in Europeans...  
Loos, R. J. F. & Janssens, A. Predicting polygenic obesity using genetic information...  
Scott, R. A. et al. An expanded genome-wide association study of type 2 diabetes risk...  
Thorleifsson, G. et al. Genome-wide association yields new sequence variants in a locus that associate with measures of obesity...  
Weedon, M. N. et al. A common variant of HMCA2 is associated with adult and childhood height in the general population...  
Gudbjartsson, D. F. et al. Many sequence variants affecting diversity of adult human height...  
Weedon, M. N. et al. Genome-wide association analysis identifies 20 loci that influence adult height...  
Letouze, G. et al. Identification of ten loci associated with height highlights new biological pathways in human growth...  
Soroiano, N. et al. Meta-analysis of genome-wide scans for human adult stature identifies novel loci and associations with measures of skeletal frame size...  
Lango Allen, H. et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height...  
Wood, A. R. et al. Defining the role of common variation in the genetic and architectural biology of adult human height...  
Heid, I. M. et al. Meta-analysis identifies 15 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution...  
Shungin, D. New genetic loci link adipose and insulin biology to body fat distribution...  
Lotta, L. A. et al. Association of genetic variants related to glucose metabolism with abdominal fat distribution...  
Jönsson, H. et al. Whole genome characterization of sequence diversity of 15,220 Icelanders...  
International HapMap 5 Consortium. Integrating common and rare genetic variation in diverse human populations...  
1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing...  
UK10K Consortium. The UK10K project identifies rare variants in health and disease...  
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...  
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. 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...  
The authors declare no competing interests...  
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations...  
Nature Reviews Genetics thanks S. Chanock, J. Florez and M. Nelson for their contribution to the peer review of this work.