






Dissecting polygenic signals from genome-wide association studies on human behaviour

Abdel Abdellaoui   and Karin J. H. Verweij 

Genome-wide association studies on human behavioural traits are producing large amounts of polygenic signals with significant predictive power and potentially useful biological clues. Behavioural traits are more distal and are less directly under biological control compared with physical characteristics, which makes the associated genetic effects harder to interpret. The results of genome-wide association studies for human behaviour are likely made up of a composite of signals from different sources. While sample sizes continue to increase, we outline additional steps that need to be taken to better delineate the origin of the increasingly stronger polygenic signals. In addition to genetic effects on the traits themselves, the major sources of polygenic signals are those that are associated with correlated traits, environmental effects and ascertainment bias. Advances in statistical approaches that disentangle polygenic effects from different traits as well as extending data collection to families and social circles with better geographical coverage will probably contribute to filling the gap of knowledge between genetic effects and behavioural outcomes.

The road from molecule to complex behaviours starts at the DNA sequence. The small effects of many individual DNA sequence variants travel through many cascades of increasingly complex biological pathways that react to environmental stimuli, resulting in complex behavioural outcomes (Fig. 1). Genome-wide association studies (GWASs) are successfully connecting millions of single DNA bases to a wide range of complex behavioural outcomes on a large scale. It is a bold approach that quantifies the relationships between the DNA sequence and the most complex of human characteristics by examining our genomes with brute force through unprecedented study sample sizes, some of which exceed a million human participants, and without including any of the intermediate processes. This hypothesis-free approach keeps providing us with many new clues on the nature of human behavioural differences, but we struggle with how to read these clues. As GWASs continue to increase their sample sizes and refine their approaches, more and better data are being produced on the association between DNA and behavioural outcomes. Here we summarize the history and current state of the field and discuss how we can make progress in disentangling and interpreting the growing amount of polygenic signals produced by GWASs on human behaviour.

The advent of GWAS

Throughout the last century, twin studies have consistently shown us that most human characteristics that show individual differences have considerable heritable components^{1,2}. Nearly every behavioural outcome that was measured showed significantly higher correlations between identical twins compared with between fraternal twins. Thus, the first law of behavioural genetics was coined, which states that ‘all human behavioural traits are heritable’³. The first studies to link genetic regions to heritable traits at the molecular level were the family linkage studies in the 1980s, in which the segregation of a limited number of genetic variants within families was compared with the segregation of a trait within families (a glossary of key terms is provided in Box 1). This approach was successful only for traits that were influenced by genes with exceptionally large

effect sizes, such as single-gene (monogenic) disorders including Huntington’s disease⁴ and cystic fibrosis⁵. In the early 2000s, the first whole human genomes were sequenced in the Human Genome Project^{6,7}. More genomes followed soon thereafter, enabling the correlational structure of genetic variants to be mapped on a population level^{8,9}, which shifted the focus from sparse genetic variants within families to genome-wide variation at the population level¹⁰. The insights and reference datasets that followed made it affordable to study larger numbers (millions) of genetic variants at larger scales, which resulted in the first GWASs^{11–13}. The microarray chips used in GWAS analyses are designed to capture as much genetic variation as possible by leveraging the correlational structure of the genome in selecting an affordable number of genetic variants to be directly measured. The correlational structure of the genome can then also be used to fill in many of the remaining gaps of unmeasured genetic variation using genotype imputation, which is especially useful when combining studies that use different microarray chips¹⁴. In a GWAS, each of the measured and imputed genetic variants is tested for its association with a complex trait of interest. Complex traits are traits that are influenced by a combination of many genetic variants (polygenic) and environmental factors. Before it was feasible to measure that many genetic variants across the genome, complex traits were being linked to genetic variants that were chosen based on a priori hypotheses in candidate gene studies. These studies were largely unsuccessful due to our limited understanding of the genetic architecture of complex traits and resulted in many false-positive findings in the literature^{15,16}. The hypothesis-free genome-wide scans of the GWAS design were the first to result in many new and replicable associations between genetic variants and complex traits.

Early on, GWAS analyses were successful in detecting associations between genetic variants and physical characteristics, including common diseases such as type 2 diabetes¹⁷, obesity¹⁸ and cardiovascular disease¹⁹, as well as non-disease dimensions of variation such as height²⁰, blood lipids²¹ and body mass index (BMI)²². As sample sizes increased, it became more viable to apply the GWAS approach to behavioural outcomes as well, starting with mental health disorders such as schizophrenia^{23,24}, bipolar disorder^{25,26} and

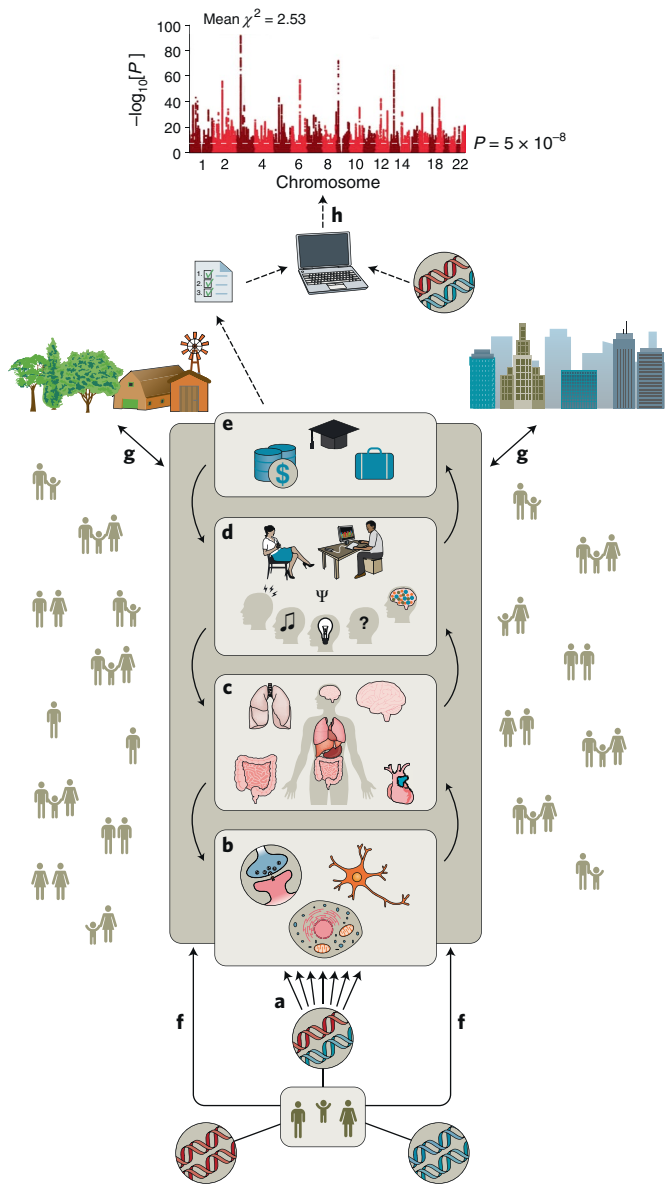


Fig. 1 | The complexity of the associations between DNA sequence variants and human behaviour. **a–h**, The small effects of many individual DNA sequence variants (**a**) travel through many cascades of bidirectional biological processes that react to environmental stimuli to end up associated with complex behavioural outcomes. The DNA sequence contains patterns that encode protein sequences and regions that regulate their expression. Proteins that are built as a result of these instructions perform vital functions at the cellular level (**b**). Complex networks of cells make up organs, a complex network of organs makes up the human body (**c**) and a complex network of humans makes up society (**g**). Individual differences in biological make-up (**b,c**) in combination with environmental differences in exposures from the family (**f**; which are also influenced by the same (genetic) influences) and society (**g**) explain the individual differences in psychological and psychiatric (Ψ) outcomes (**d**). These behavioural outcomes have an important role for people's place in society, which is often measured through socioeconomic outcomes such as educational attainment, occupation and income (**e**). All of these different phenotypic levels influence each other as well as the people and the environment around them. In modern behavioural genetics, to date, we have focused mainly on estimating the associations between genetic variants and relatively crude measures of human behaviour (**h**; the Manhattan plot at the top shows the results of a GWAS on educational attainment³⁴), with relatively little progress on disentangling what processes are happening in-between.

major depression^{27,28}, and eventually also moving on to non-disease dimensions of human behaviour, such as personality traits²⁹, sexual orientation³⁰, substance use^{31,32}, intelligence³³ and educational attainment³⁴. With more advanced statistical genetics methods and growing sample sizes, the last decade has led to growing insights into the genetic architecture of complex behavioural traits. GWAS analyses have helped to expose the highly polygenic nature of these complex behavioural traits, showing that behavioural outcomes are influenced by more common genetic variants with smaller effects than many previously assumed^{35,36}. The distribution of the observed effects of individual genetic variants was largely consistent with the infinitesimal model that was proposed by R. A. Fisher about a century earlier, whereby quantitative traits are influenced by an almost infinite amount of genes with increasingly smaller additive effects³⁷. Non-additive genetic effects (that is, gene–gene interactions and dominance effects) are rarely included in traditional GWAS designs, because they are more difficult to estimate, and both theory and empirical evidence suggest that their contribution is minimal^{38,39}. Predictions from theory and experimental evidence state that, for polygenic traits, the smaller the effects of individual genes, the more nearly additive they are⁴⁰. These increasingly smaller effects of individual genetic variants on behavioural traits make it difficult for GWASs to capture all genetic influences that were predicted to exist by twin studies (Fig. 2). A substantial portion of the expected heritability may also reside in rare genetic variants that are not well captured by current microarray chips^{41,42}. Efforts to increase GWAS sample sizes and genomic coverage to capture more heritability in the polygenic signals are continuously ongoing. However, the proportion of the heritability explained in GWAS analyses to date contains enough signal to do meaningful follow-up studies on the relationships between the DNA sequence and complex behavioural outcomes.

Biology of behaviour

Identifying genetic variants that are associated with complex traits is regarded as an important first step towards deeper insights into underlying biological mechanisms. Studies have shown that therapeutic drug targets for physical diseases with support from GWAS results are more than twice as likely to succeed^{43,44}, which is one of the many signs that GWAS results indeed contain biologically meaningful information. For behavioural traits, the search for biological insights from genetic variants identified by GWASs is still in its infancy. One notable occasion of an individual GWAS association leading to a crucial biological insight occurred when follow-up analyses revealed that the strongest associations for schizophrenia came from a genetic variant on chromosome 6 that increases synaptic pruning during late adolescence⁴⁵. However, to date, more studies have succeeded in learning about the biology of behaviour by looking at polygenic signals (that is, the aggregate of effects of many genetic variants) than by looking at individual genetic variants.

As more GWAS associations lie in intergenic and intronic regions, it is probable that gene regulation has a larger role in individual differences compared with changes to the proteins⁴⁶. If differences within our species are largely due to varying gene expression levels, this would be consistent with the finding that differences in neuronal synapses between different vertebrate species are also largely due to the tuning of protein expression levels rather than changes in the proteome content⁴⁷. Novel methods that allow polygenic signals from behavioural traits to be partitioned into manually curated components confirm that polygenic signals for behavioural traits are strongly enriched for genetic variants that are known to influence gene expression through transcription and gene methylation^{48–50}. The same approaches show that the genes that are being regulated are largely expressed in expected tissues with respect to their associated traits—polygenic signals for height come largely from genes expressed in connective tissues or bone cells; polygenic signals for

Box 1 | Glossary of key terms

Family linkage study. Parents transmit relatively large segments of DNA to their offspring, causing genetic variants that are physically closer to each other to be transmitted together more often than expected by chance (that is, the genetic variants are linked). In a family linkage study, it is investigated whether a disease co-occurs with measured genetic markers in family pedigrees, in which case, the causal genetic variant is probably near (linked to) that genetic marker. This approach has been successful for identifying genetic variants with relatively large effects.

Genotype imputation. Some genetic variants appear together in a population more often than others, especially when they are physically closer to each other. Genotype imputation is a process whereby unmeasured genetic variants are estimated (imputed) with the help of the expected correlations between measured and unmeasured genetic variants, which are estimated from a reference dataset with more densely measured genomes of the respective population.

GWAS. A hypothesis-free study design that estimates the associations between many genetic variants (up to millions) and a heritable trait. The genetic variants that are analysed in a GWAS are generally limited to common single-nucleotide polymorphisms (SNPs; which are substitutions of a single nucleotide at a specific location in the genome). This approach has been successful for identifying many common genetic variants with smaller effects.

Polygenic signal. The aggregate of estimated effects of many genetic variants on a heritable trait. These are generally summarized in GWAS summary statistics, which include the effect sizes and P values for each individual genetic variant.

Complex traits. Traits that are influenced by a combination of many genetic variants (that is, polygenic) with relatively small effects and environmental factors.

Polygenic score. A genetic predictor for a complex trait, which is computed by summing the genetic variants carried by an individual, weighted by their estimated effect size. The effect size

estimates come from a GWAS that excludes the participants for which the polygenic score is computed.

Population stratification. A systematic difference between populations in the frequencies of genetic variants due to different ancestral backgrounds. These can cause false-positive associations in a GWAS if the trait also differs between the populations. This can be controlled for by (1) analysing relatively homogeneous populations and (2) controlling for large patterns of genetic variation that reflect ancestry differences.

Genetic correlation. An estimate of the overlap in genetic effects between two traits. A genetic correlation (r_g) can vary from -1 (100% overlap in the opposite direction) to 0 (no overlap) to 1 (100% overlap in the same direction). Significant genetic correlations are widespread between behavioural traits (Fig. 2).

Mendelian randomization. A study design that uses genetic variants, ideally with a known effect mechanism, as instrumental variables to assess the causal effect of a modifiable exposure (on which the genetic variants have an effect) on an outcome (on which the genetic variants are assumed to not have a direct effect).

Gene–environment correlation. When genetic influences are associated with environmental exposures, which can happen in three ways: (1) passive gene–environment correlation occurs when the rearing environment that parents provide for their offspring is influenced by heritable parental behaviours, resulting in parents who pass down both genes and environment to offspring; (2) active gene–environment correlation comes about when a choice that leads to an environmental exposure is influenced by heritable traits; and (3) evocative gene–environment correlations occur when the environmental exposure results from a response from others to a heritable outcome.

Ascertainment bias. A bias in the data collection in which a (heritable) trait is associated with the probability of being included in the sample.

inflammatory disorders come mostly from genes expressed in the immune system; and polygenic signals for behavioural traits such as substance use, psychiatric disorders or educational attainment are significantly enriched for genes expressed in the central nervous system, especially in the cortex and the cerebellum^{50,51}. Surprisingly, polygenic signals for BMI also mostly come from genes that are expressed in the central nervous system (especially the cortex and the cerebellum), suggesting that BMI is more of a behavioural trait than a metabolic one^{50,51}. These methods enable the polygenic signal to be further delineated into more specific neuronal types, showing, for example, that the polygenic signal for bipolar disorder is significantly enriched for GABAergic neurons (inhibitory), whereas the signals for BMI and schizophrenia seem to be more enriched for glutamatergic neurons (excitatory), and the signals for educational attainment and neuroticism for neither⁵⁰. One could zoom in further on the synaptic level, showing that polygenic signals for educational attainment and bipolar disorder are more enriched for postsynaptic activity, autism for presynaptic activity, and attention deficit hyperactivity disorder for both⁵². However, as behavioural traits are more distal and less directly under biological control than physical traits, the delineation of their polygenic signals into such specific biological categories becomes more difficult to interpret. These polygenic signals are probably made up of a composite of signals from different sources that are a result of different combinations of underlying biological processes. If a GWAS analysis is the first step towards a deeper understanding of the biology of

behaviour, the steps in between GWAS analyses and biology have to include disentangling what those different sources are (Box 2).

Prediction of behaviour

Understanding the underlying biology leading to an association between polygenic signals and their outcome is not always regarded as a strong prerequisite to using polygenic signals in practice. A promising potential for polygenic signals lies in the ability to leverage them to produce genetic risk predictors. These genetic risk predictors (hereafter, polygenic scores), represent the aggregated genetic effects across the genome and can be computed by summing the alleles that an individual carries and weighting them by the allelic effect estimates from GWASs. Despite the gap in our knowledge about the processes between the polygenic signals and the phenotypic outcomes, these polygenic scores are predictive of heritable outcomes, which has proven to be useful in research and holds potential for clinical use^{53–55}. Although each genetic variant generally explains a very small proportion of variance, the combined effect of all variants can be substantial, although it is presently considerably lower than the total heritability estimates from twin studies. One of the first successful applications of polygenic scores in research was in 2009, in which it served as evidence for the polygenic nature of schizophrenia⁵⁶. The polygenic score for schizophrenia then explained ~2–3% of the individual differences in schizophrenia in an independent sample, which increased to ~10% after increasing the effective GWAS sample size from ~6,900

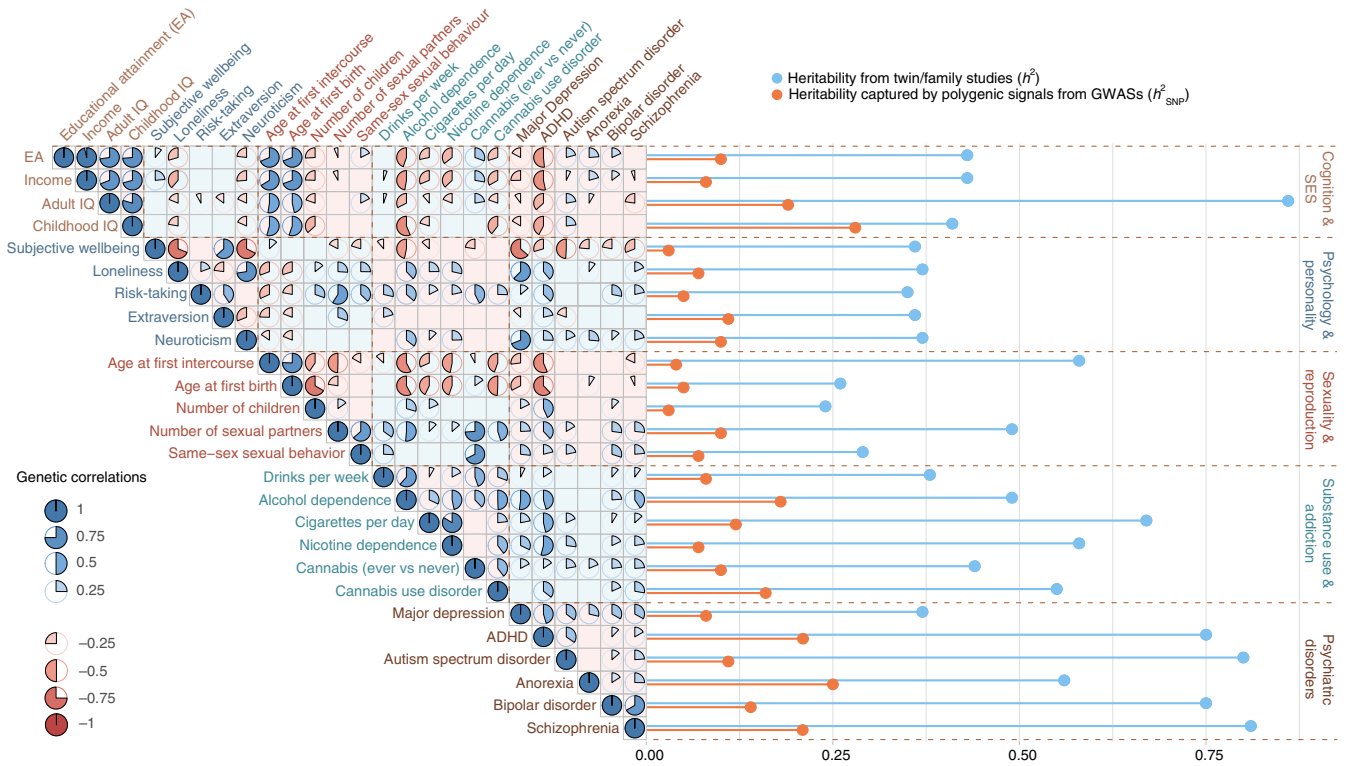


Fig. 2 | Genetic correlations, SNP-based heritabilities and twin/family-based heritabilities. The genetic correlations as captured by the polygenic GWAS signals (left; computed using linkage disequilibrium score (LDSC) regression⁸⁶). The empty boxes show non-significant genetic correlations. Statistical significance is based on false-discovery-rate-corrected *P* values. Right: the heritability estimates as obtained from twin/family studies and the heritability as captured by the polygenic GWAS signal (that is, SNP-based heritability), computed using LDSC regression⁸¹ (references for the GWASs and the twin/family studies are provided in Supplementary Table 1). ADHD, attention deficit hyperactivity disorder; SES, socio-economic status.

to ~214,000 about a decade later⁵⁷. The polygenic score for the less-heritable outcome of educational attainment went through a comparable trajectory, first explaining 2% of the variation using a GWAS of ~125,000 participants⁵⁸, which increased to ~12% when the GWAS sample size reached 1.1 million participants³⁴. As sample sizes keep expanding, the prediction accuracy of polygenic scores is expected to further increase with an expected r^2 (that is, the proportion of variance explained) of $h^2_{\text{SNP}} / (1 + (M/N \times h^2_{\text{SNP}}))$, where h^2_{SNP} is the heritability captured by the GWAS, M is the number of independent genetic variants and N is the discovery GWAS sample size⁵⁹. This expected increase may eventually yield clinically actionable predictions from disease-related polygenic scores⁶⁰. Polygenic scores can improve the predictive power of clinical models, as has been shown for cardiovascular disease, breast cancer, prostate cancer and type 1 diabetes^{61–65}. With enough predictive power, polygenic scores could potentially be applied to identify individuals at risk before symptoms manifest and to identify who could be most responsive to specific treatments. Even without a deeper knowledge on the origin of the predictive value, contributions to research and the clinic can be useful provided that the prediction is valid and informative for the research question or clinical outcome. However, prediction of more complex behavioural outcomes with polygenic scores comes with substantial caveats, particularly when the interpretation of the prediction is relevant to the research question or the intervention. When the nature of the predictive ability of polygenic scores for behavioural traits is not fully understood, applying them in either scientific research or the clinic can lead to incorrect interpretations and conclusions. The ease with which polygenic scores can be applied for prediction should increase the urgency with which we need to invest in disentangling the sources of the polygenic signals that these scores are based on.

Dissecting the polygenic signal

Substantial parts of genetic signals from GWASs contain signals that are not directly related to the biology of the trait of interest itself. The largest patterns of genetic variation in a population reflect ancestry differences⁶⁶, which are strongly correlated with geography^{67–70} and are therefore in line with many environmental and cultural differences between populations/subpopulations. Not controlling for these systematic ancestry differences results in spurious associations (population stratification)⁷¹. In the earlier days of GWAS analyses, confounding due to ancestry differences that was not accounted for accordingly could have easily gone undetected and become part of the signal^{72,73}. More recently, methods have been developed to successfully disentangle which part of the GWAS signal is due to confounding due to ancestry differences and which part reflects polygenic effects⁷⁴. This was an important step in extracting the genetic signals of interest from GWAS studies, but blindly extracting all associated signals that reflect polygenicity is still probably not sufficient for a fuller understanding of the genetic architecture of the trait. Besides polygenic effects on the trait of interest itself, the part of the GWAS signal that reflects polygenicity is expected to contain significant amounts of signal from at least three other sources that need to be disentangled—polygenic effects on correlated traits, environmental effects that are correlated with polygenic effects and polygenic effects reflecting systematic biases in the ascertainment of study participants.

Genetic correlations. At the end of the nineteenth century, the correlation coefficient was invented in an attempt to quantify inherited similarities between individuals from the same family and to quantify the covariation between different traits within the same

Box 2 | The next steps in behavioural genetics

GWAS analyses of behavioural traits produce polygenic signals that are difficult to interpret. The next phase in the field of behavioural genetics includes disentangling the complex mixture of sources that make up these polygenic signals, including the following steps:

- **Partitioning polygenic signals into their subcomponents.** Polygenic signals from behavioural traits include a composite of a large number of lower-level outcomes. Studies need to extend their focus beyond single behavioural traits, and combine polygenic signals from multiple measures using methodological advancements in statistical genetics that (1) enable polygenic signals to be decomposed into their subcomponents and (2) enable the modelling of the (causal) relationships between those components.
- **Dealing with environmental effects in polygenic signals.** Heritable traits are associated with environmental exposures for a variety of reasons, which results in polygenic signals that contain environmental effects. Analysing genetic data within families and within geographical regions would mitigate this type of confounding, but would also require larger genotyped datasets that extend to families and social circles, preferably with a good geographical coverage of the population.
- **Dealing with ascertainment bias.** Whether someone participates in a study is a heritable behavioural trait. This affects polygenic signals for a wide range of behavioural traits. To overcome this bias, the data collection needs to be extended to unmeasured parts of the population, and/or additional

individuals^{75,76}. It helped to further illustrate the heritability of human traits and it helped in exposing that correlations between human traits and diseases are widespread. For example, people with a mental health disorder show a higher risk of developing other mental health problems⁷⁷, but are also at a higher risk of developing additional physical health problems⁷⁸. There are several ways in which traits can become correlated with each other, mostly due to one trait causing the other or due to shared underlying causal factors. Two traits are genetically correlated with each other when a significant portion of the correlation between them can be explained by the same underlying genetic influences. Significant genetic correlations have been detected between many behavioural outcomes by measuring and modelling multiple outcomes within identical and fraternal twins^{79–85}. More recently, methods have been developed that make it possible to compute genetic correlations between traits without having to measure the different traits in the same individuals. These methods compare polygenic signals extracted from GWAS results on different traits⁸⁶. This has greatly increased our potential to map out the genetic overlap between a wide range of complex traits and diseases. Results so far suggest that significant genetic correlations between complex traits are widespread, both between behavioural traits (Fig. 2) and between behavioural and physical traits^{86,87}. Genetic correlations can reflect different types of relationships⁸⁸, mostly reflecting some form of pleiotropy (that is, the same genes influencing different traits), the following two in particular: (1) horizontal pleiotropy, where the same genetic variants influence both traits either directly or indirectly through an intermediate phenotype; and (2) vertical pleiotropy, where the genetic variants influence the first trait, and the first trait has a causal effect on a second trait. Methods that distinguish between horizontal and vertical pleiotropy indicate that many of the genetic correlations result from a combination of both^{89,90}. Another potential source for genetic correlations is cross-trait assortative mating, whereby

representative genetic ‘census’ datasets need to be assembled that can be used to model and account for participation bias.

A more global form of ascertainment bias is the strong Eurocentric focus of most large-scale GWASs. Systematic differences in genetic and environmental influences between ancestries make current polygenic signals substantially less predictive in individuals of non-European descent¹³². Data collection needs to be expanded to a wider range of ancestral backgrounds to better reflect global and increasingly diversifying societies.

- **Increasing sample sizes.** Despite the enormous increase in GWAS sample sizes throughout the last decade, we are still capturing only a fraction of the expected heritability of behavioural traits. To increase the heritability captured by GWASs, sample sizes need to be further increased. However, note that the biases discussed here will also become more pronounced in the GWAS signal with an increased sample size, so it is essential that those are dealt with accordingly, particularly through more inclusive sampling strategies.
- **Sharing GWAS summary statistics.** The development of methodology and analysis of polygenic signals requires expertise from a wide array of disciplines, including statistics, genetics, bioinformatics and psychometrics. Public access to GWAS results is an essential stimulant for this cooperative scientific effort. Fortunately, it is already standard practice among many groups that conduct large-scale GWASs to publicly share their GWAS summary statistics, and it is important to continue to stimulate and expand this tradition.

people who score high on one trait tends to choose a mate who scores highly on another trait, causing genes for the two different traits to be passed down and inherited together. This has been suggested to partly explain the genetic correlation between educational attainment and height⁹¹. Assortative mating on the same trait, which is especially strong for educational attainment^{92,93}, can also result in the trait becoming genetically correlated with itself, meaning that genetic variants that influence educational attainment, but are not usually transmitted together, will appear together in the offspring more often than expected by chance, which can inflate heritability estimates⁹⁴.

One way to learn more about the nature of widespread pleiotropy among behavioural traits is to dissect the polygenic signals into polygenic effects that are shared between traits and polygenic effects that are more trait specific. Several methods have recently been developed and applied that take important first steps towards that goal. Some approaches leverage overlapping polygenic signals to improve the statistical power to study the genetic architecture of shared underlying dimensions; these approaches have been applied to better capture polygenic effects that overlap between mood-related traits (wellbeing, major depression and neuroticism)^{95,96}, between psychiatric disorders⁹⁷, and between intelligence and educational attainment⁹⁸. A recently developed approach called case–case GWAS (CC-GWAS)⁹⁹ uses polygenic signals from two different disease GWASs (that is, disease cases versus healthy controls) to create new polygenic signals that reflect what separates the two groups of disease cases from each other. To illustrate the effectiveness of CC-GWAS, the approach was applied to identify novel loci that reflect genetic differences between eight psychiatric disorders⁹⁹. Psychiatric disorders are, with their enormous societal burden, important candidates for these applications, because to date we understand little about why the substantial genetic overlap between polygenic signals for psychiatric disorders is not in line with current

clinical boundaries^{100,101}. Another recently developed method called genomic structural equation modelling (genomic SEM)¹⁰² can be used to model the joint genetic architecture of complex traits. Genomic SEM has recently been applied to subtract polygenic signals for cognitive ability from polygenic signals of educational attainment¹⁰³. The cognitive and non-cognitive parts were estimated to make up 43% and 57% of the educational attainment polygenic signal, respectively. These analyses further revealed that the cognitive and non-cognitive skills that make up the genetic predisposition for educational attainment show genetic correlations in the opposite directions for conscientiousness, extraversion, agreeableness, empathy, and the risk of schizophrenia and bipolar disorder (these all show negative genetic correlations with cognitive skills and positive genetic correlations with non-cognitive skills). Additional differences between the cognitive and non-cognitive part of the polygenic signal include that the non-cognitive part shows stronger associations with traits that are related to risk tolerance, delayed gratification and healthier behaviour, as well as weaker associations with regional brain volume and stronger associations with white matter tracts¹⁰³. The same approach has been applied to subtract polygenic signals associated with socioeconomic status from polygenic signals from GWASs on mental health problems, which substantially altered the genetic correlations between various mental health problems, most strongly for attention deficit hyperactivity disorder and substance use¹⁰⁴.

While identifying which parts overlap between traits can be leveraged to increase the power of a GWAS by combining different traits, identifying which parts of the polygenic effects are more trait specific may help to get us closer to the biological mechanisms for a specific trait of interest. Furthermore, the production of 'cleaner' polygenic signals opens the door to a better investigation of the causal relationships that underlie the widespread genetic correlations between complex behavioural outcomes. Methods that leverage polygenic signals to investigate causal relationships are in place: Mendelian randomization (MR) is an instrumental variable approach that uses genetic markers that are robustly predictive of an 'exposure' variable as an instrument to test causal effects on an 'outcome' variable of interest. Assuming that genes are randomly transmitted from parents to offspring and that an outcome cannot alter a person's genes, MR suffers less from confounding and reverse causality compared with conventional observational research¹⁰⁵. By picking a gene ideally with a known effect on the exposure of interest (for example, a nicotinic receptor gene affecting smoking behaviour) and testing for its association with an outcome variable (for example, lung cancer), MR can reveal the causal effect of the exposure on the outcome (for example, whether smoking causes lung cancer), because the known effect of the gene can travel only through the exposure (smoking) to the outcome (lung cancer), and not the other way around. Indeed, MR has been successfully applied to establish the causal relationships between smoking and lung cancer¹⁰⁶, and has successfully shown the absence of causal effects of high-density lipoprotein cholesterol¹⁰⁷ and hormone replacement¹⁰⁸ on heart disease. However, MR can suffer from confounding in the presence of horizontal pleiotropy¹⁰⁹, which could potentially be addressed by the extraction of non-overlapping polygenic signals. Another assumption of the MR approach is that the genetic instrument is not associated with confounders that influence the two traits under investigation¹⁰⁹, which could be violated in the presence of gene–environment correlations.

Gene–environment correlations. Complex traits and diseases are influenced by a combination of genetic and environmental factors, which are not independent sources of variation. Environmental effects can be correlated with polygenic effects due to a variety of reasons. The effects that genes have on people's behaviour also influence which environment they actively expose themselves to

(active gene–environment correlations). For example, it was recently shown that people who leave economically disadvantaged neighbourhoods have a higher polygenic score for educational attainment than the people they leave behind, causing a correlation between genes associated with educational attainment and environmental factors associated with regional wealth and health¹¹⁰. Rearing environments are influenced by the genetic make-up of parents, which results in a correlation between the parental genes that offspring inherit and the environment in which they grow up (passive gene–environment correlations). Polygenic scores for educational attainment built from the half of the parental genes that are not transmitted to their offspring have recently been shown to be predictive of offspring education and a range of other offspring health outcomes¹¹¹. These familial indirect genetic effects are significant for both the cognitive part and the non-cognitive part of the polygenic signal of educational attainment¹¹². This shows that effects that are often considered to be environmental (parental rearing environment) can also be under genetic influence.

When environmental effects are correlated with polygenic effects, they become part of the polygenic signal in a traditional GWAS design. One way to reduce confounding due to gene–environment correlations is through family-based GWAS designs. These approaches generally compare transmitted and non-transmitted alleles between siblings, which has the added benefit of protection against confounding due to population stratification, as the transmitted and non-transmitted alleles between siblings have the same ancestry⁷¹. These designs are expected to remove a great deal of inflation for traits such as educational attainments, for which polygenic scores have been estimated to have a predictive power of ~1.6 times greater between families than within families¹¹³ and twice as large in non-adopted children than in adopted children¹¹⁴. Recently, one of the first large-scale within-family GWAS efforts estimated the genetic effects for 25 traits in up to ~160,000 siblings, and found a decreased polygenic signal for a range of traits, most strongly reduced for educational attainment, and strongly reduced genetic correlations between educational attainment and physical traits such as BMI and height¹¹⁵. The genetically informative nature of twins has motivated the recruitment of large numbers of twins and their family members by twin registries throughout the globe^{116,117}, which will increase in value once again when within-family designs are more broadly implemented in a GWAS setting.

Family-based designs will not be sufficient to get rid of all inflation when gene–environment correlations extend beyond the family. The environments that shape us are probably influenced by people outside the family as well, such as neighbours, teachers, colleagues and peers. Indirect genetic effects from individuals outside of the family have received little attention so far. Expanding family cohorts to include genotypes and behavioural measures of people in their social networks will probably add valuable information for new research designs that could take these influences into account. The physical and social environments that we live in are built by the communities we live in, and the people within those communities may also react to us differently depending on our own genetic make-up (evocative gene–environment correlations). If the alleles a person carries are correlated with the alleles of people close by, these indirect genetic effects should also lead to an overestimation of the effects of an individual's alleles in GWASs. It was recently described how migration leads to geographical clustering of alleles that influence complex traits and disease risk in Great Britain¹¹⁰. Regional clustering is strongest for alleles that are associated with educational attainment, but is also significant for alleles that are associated with other behavioural traits such as personality and mental health. The geographical clustering of risk alleles is in line with regional measures of environmental risk factors for health and socioeconomic adversity¹¹⁰. Controlling for geographical location or conducting within-region GWAS analyses has recently been shown to reduce

this type of confounding¹¹⁸, but may not eliminate it completely, as geographical location alone is a relatively crude approximation of someone's social environment. To more fully account for genetic correlates from one's physical and social environment, we would need deeper longitudinal characterizations of the participants' lives and social circles.

Ascertainment bias. Even when defining the phenotype perfectly and accounting for all gene–environment correlations, a GWAS probably still contains polygenic signals from unintended phenotypes, namely phenotypes related to whether participants were able and willing to be measured. The people who donate their DNA and phenotypic measurements to the datasets that we study are rarely representative of the general population.

Participation bias is detectable through polygenic risk: people with a higher genetic risk of mental health problems, such as schizophrenia, were less likely to complete questionnaires or attend data collection in a population cohort of ~15,000 participants¹¹⁹. Such participation bias can also introduce artificial (genetic) correlations between two traits through collider bias when those two traits influence the variable that influences participation¹²⁰. These confounding effects do not get solved by increasing sample sizes; to the contrary, with larger sample sizes, the statistical power increases to detect polygenic effects associated with behavioural differences between people who participate in studies versus the general population. One of the most widely used large-scale genetic datasets is the UK Biobank ($n_{\text{researchers}} \approx 15,000$, $n_{\text{participants}} \approx 500,000$), which depends on voluntary participation. The ~500,000 volunteers make up 5.5% of the ~9.2 million people who were approached to participate in the study and are not a perfect reflection of the general population; on average, they tend to live healthier lives, are more educated and live in less economically deprived areas of the country^{121,122}. In the UK Biobank, e-mail contact and the completion of an online mental health survey show significant genetic correlations with educational attainment and physical and mental health¹²³. Continued engagement in follow-up measurements in the UK Biobank is associated with polygenic signals for complex traits such as intelligence, Alzheimer's disease, neuroticism and schizophrenia¹²⁴. Another large contributor to many GWASs is 23andMe, which has access to DNA from millions of their customers who participate in their studies in return for feedback on their genetic make-up. Both the UK Biobank and 23andMe show sex differences in participation bias that are detectable at the polygenic level. For example, females with a higher genetic predisposition for higher BMI are less likely to have participated in the study than their male counterparts¹²⁵. Non-random misreports related to disease ascertainment can introduce additional biases; for example, in the UK Biobank, people with a higher disease burden tend to underreport their alcohol intake, which significantly affects estimates of genetic correlations between alcohol use and complex disease outcomes, sometimes even changing the direction of the correlation¹²⁶.

Most human populations on Earth are underrepresented in GWASs, as ~88% of the participants up to 2017 were of European descent¹²⁷. The ethnic homogeneity of GWAS datasets is partly intended to prevent systematic ancestry differences from being mistaken for associations due to causal genetic variants (that is, false positives due to population stratification)⁷¹. Even within the large European datasets, participation bias is difficult to overcome when you do not measure the entire population. There are a few examples of studies with a population approach, such as the Danish iPSYCH¹²⁸ study, which is based on nationwide neonatal dried blood spots, or the Icelandic deCODE¹²⁹ dataset, which consists of genotypes and phenotypes from approximately half of the Icelandic population. Such population-level genetic databases could provide important population-level estimates of genetic variation that could potentially be used to model and account for participation bias in GWAS studies¹²⁵. Genotyping

the majority of a population is not yet a viable option for most countries but, given the relatively strong geographical clustering of genome-wide ancestry and complex trait variation^{67–69,110,130}, we are more likely to get closer to population-level estimates if we aim for higher geographical coverage.

Conclusions

The pathways between our genetic code and complex behavioural outcomes contain many possible cascades of biological and social processes with intermediate (endo)phenotypic outcomes. Mapping these pathways in their entirety is a daunting task that will probably not be completed soon. We have the ability to, within a reasonable timeframe, substantially narrow the gap between the polygenic signals that we currently capture and behavioural traits of interest, which will be helpful for studies that rely on GWAS signals to investigate causal relationships or underlying biological mechanisms. We summarized the three major sources of polygenic signal that are being mixed with the polygenic signals of interest and described ways in which we can better characterize them or account for them. However, these additional sources are not just noise that we need to exclude from our signals; these unintentionally captured signals can teach us more about the genetic architecture of human behaviour and the way we study them. Before we are ready to dive deeper into the biological processes behind the polygenic signals, or their utility in the clinic or social policies, the way forward is to keep increasing sample sizes, ideally targeting families and extended social circles with good geographical coverage, to increase the signal while simultaneously refining the polygenic signals by better delineating their origin.

Received: 17 August 2020; Accepted: 31 March 2021;

Published online: 13 May 2021

References

- Boomsma, D., Busjahn, A. & Peltonen, L. Classical twin studies and beyond. *Nat. Rev. Genet.* **3**, 872–882 (2002).
- Polderman, T. J. et al. Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nat. Genet.* **47**, 702–709 (2015).
- Turkheimer, E. Three laws of behavior genetics and what they mean. *Curr. Dir. Psychol. Sci.* **9**, 160–164 (2000).
- Gusella, J. F. et al. A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* **306**, 234–238 (1983).
- Tsui, L.-C. et al. Cystic fibrosis locus defined by a genetically linked polymorphic DNA marker. *Science* **230**, 1054–1057 (1985).
- Venter, J. C. et al. The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
- Lander, E. S. et al. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
- Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J. & Lander, E. S. High-resolution haplotype structure in the human genome. *Nat. Genet.* **29**, 229–232 (2001).
- Gabriel, S. B. et al. The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229 (2002).
- The International HapMap Consortium The international HapMap project. *Nature* **426**, 789–796 (2003).
- DeWan, A. et al. HTRA1 promoter polymorphism in wet age-related macular degeneration. *Science* **314**, 989–992 (2006).
- The 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).
- Klein, R. J. et al. Complement factor H polymorphism in age-related macular degeneration. *Science* **308**, 385–389 (2005).
- Li, Y., Willer, C., Sanna, S. & Abecasis, G. Genotype imputation. *Annu. Rev. Genomics Hum. Genet.* **10**, 387–406 (2009).
- Duncan, L. E., Ostacher, M. & Ballon, J. How genome-wide association studies (GWAS) made traditional candidate gene studies obsolete. *Neuropsychopharmacology* **44**, 1518–1523 (2019).
- Border, R. et al. No support for historical candidate gene or candidate gene-by-interaction hypotheses for major depression across multiple large samples. *Am. J. Psychiatry* **176**, 376–387 (2019).

17. Morris, A. P. et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat. Genet.* **44**, 981–990 (2012).
18. Locke, A. E. et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).
19. Ehret, G. B. et al. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature* **478**, 103–109 (2011).
20. Lango, H. A. et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–838 (2010).
21. Teslovich, T. M. et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).
22. Yang, J. et al. FTO genotype is associated with phenotypic variability of body mass index. *Nature* **490**, 267–272 (2012).
23. Ripke, S. et al. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
24. Ripke, S. et al. Genome-wide association study identifies five new schizophrenia loci. *Nat. Genet.* **43**, 969–976 (2011).
25. Stahl, E. A. et al. Genome-wide association study identifies 30 loci associated with bipolar disorder. *Nat. Genet.* **51**, 793–803 (2019).
26. Sklar, P. et al. Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near *ODZ4*. *Nat. Genet.* **43**, 977–983 (2011).
27. Wray, N. R. et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat. Genet.* **50**, 668–681 (2018).
28. Ripke, S. et al. A mega-analysis of genome-wide association studies for major depressive disorder. *Mol. Psychiatry* **18**, 497–511 (2013).
29. Lo, M.-T. et al. Genome-wide analyses for personality traits identify six genomic loci and show correlations with psychiatric disorders. *Nat. Genet.* **49**, 152–156 (2017).
30. Ganna, A. et al. Large-scale GWAS reveals insights into the genetic architecture of same-sex sexual behavior. *Science* **365**, eaat7693 (2019).
31. Liu, M., Jiang, Y. & Wedow, R. Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nat. Genet.* **51**, 237–244 (2019).
32. Pasmán, J. A. et al. GWAS of lifetime cannabis use reveals new risk loci, genetic overlap with psychiatric traits, and a causal effect of schizophrenia liability. *Nat. Neurosci.* **21**, 1161–1170 (2018).
33. Sniekers, S. et al. Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence. *Nat. Genet.* **49**, 1107–1112 (2017).
34. Lee, J. J. et al. Gene discovery and polygenic prediction from a 1.1-million-person GWAS of educational attainment. *Nat. Genet.* **50**, 1112–1121 (2018).
35. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *Am. J. Hum. Genet.* **90**, 7–24 (2012).
36. Visscher, P. M. et al. 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
37. Fisher, R. A. XV.—The correlation between relatives on the supposition of Mendelian inheritance. *Earth Environ. Sci. Trans. R. Soc. Edinb.* **52**, 399–433 (1919).
38. Hivert, V. et al. Estimation of non-additive genetic variance in human complex traits from a large sample of unrelated individuals. *Am. J. Hum. Genet.* <https://doi.org/10.1016/j.ajhg.2021.02.014> (2021).
39. Hill, W. G., Goddard, M. E. & Visscher, P. M. Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genet.* **4**, e1000008 (2008).
40. Crow, J. F. On epistasis: why it is unimportant in polygenic directional selection. *Philos. Trans. R. Soc. B* **365**, 1241–1244 (2010).
41. Wainschtein, P. et al. Recovery of trait heritability from whole genome sequence data. Preprint at *bioRxiv* <https://doi.org/10.1101/588020> (2019).
42. Kaiser, J. ‘Landmark’ study resolves a major mystery of how genes govern human height. *Science* (3 November 2020).
43. Nelson, M. R. et al. The support of human genetic evidence for approved drug indications. *Nat. Genet.* **47**, 856–860 (2015).
44. King, E. A., Davis, J. W. & Degner, J. F. Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. *PLoS Genet.* **15**, e1008489 (2019).
45. Sekar, A. et al. Schizophrenia risk from complex variation of complement component 4. *Nature* **530**, 177–183 (2016).
46. Ward, L. D. & Kellis, M. Interpreting noncoding genetic variation in complex traits and human disease. *Nat. Biotechnol.* **30**, 1095–1106 (2012).
47. Emes, R. D. et al. Evolutionary expansion and anatomical specialization of synapse proteome complexity. *Nat. Neurosci.* **11**, 799–806 (2008).
48. Ip, H. F. et al. Characterizing the relation between expression QTLs and complex traits: exploring the role of tissue specificity. *Behav. Genet.* **48**, 374–385 (2018).
49. Qi, T. et al. Identifying gene targets for brain-related traits using transcriptomic and methylomic data from blood. *Nat. Commun.* **9**, 2282 (2018).
50. Finucane, H. K. et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621–629 (2018).
51. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
52. Koopmans, F. et al. SynGO: an evidence-based, expert-curated knowledge base for the synapse. *Neuron* **103**, 217–234 (2019).
53. Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* **19**, 581–590 (2018).
54. Ikeda, M., Saito, T., Kanazawa, T. & Iwata, N. Polygenic risk score as clinical utility in psychiatry: a clinical viewpoint. *J. Hum. Genet.* **66**, 53–60 (2020).
55. Wray, N. R. et al. From basic science to clinical application of polygenic risk scores: a primer. *JAMA Psychiatry* **78**, 101–109 (2021).
56. The International Schizophrenia Consortium. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).
57. The Schizophrenia Working Group of the Psychiatric Genomics Consortium, Ripke, S., Walters, J. T. R. & O’Donovan, M. C. Mapping genomic loci prioritises genes and implicates synaptic biology in schizophrenia. Preprint at *medRxiv* <https://doi.org/10.1101/2020.09.12.20192922> (2020).
58. Rietveld, C. A. et al. GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science* **340**, 1467–1471 (2013).
59. Wray, N. R. et al. Pitfalls of predicting complex traits from SNPs. *Nat. Rev. Genet.* **14**, 507–515 (2013).
60. Lambert, S. A., Abraham, G. & Inouye, M. Towards clinical utility of polygenic risk scores. *Hum. Mol. Genet.* **28**, R133–R142 (2019).
61. 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).
62. Schumacher, F. R. et al. Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nat. Genet.* **50**, 928–936 (2018).
63. Sharp, S. A. & Rich, S. S. Development and standardization of an improved type 1 diabetes genetic risk score for use in newborn screening and incident diagnosis. *Diabetes Care* **42**, 200–207 (2019).
64. Sparano, J. A. et al. Clinical and genomic risk to guide the use of adjuvant therapy for breast cancer. *N. Engl. J. Med.* **380**, 2395–2405 (2019).
65. Inouye, M. et al. Genomic risk prediction of coronary artery disease in 480,000 adults: implications for primary prevention. *J. Am. Coll. Cardiol.* **72**, 1883–1893 (2018).
66. Price, A. L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
67. Abdellaoui, A. et al. Population structure, migration, and diversifying selection in the Netherlands. *Eur. J. Hum. Genet.* **21**, 1277–1285 (2013).
68. Kerminen, S. et al. Fine-scale genetic structure in Finland. *G3* **7**, 3459–3468 (2017).
69. Novembre, J. et al. Genes mirror geography within Europe. *Nature* **456**, 98–101 (2008).
70. Leslie, S. et al. The fine-scale genetic structure of the British population. *Nature* **519**, 309–314 (2015).
71. Price, A. L., Zaitlen, N. A., Reich, D. & Patterson, N. New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet.* **11**, 459–463 (2010).
72. Berg, J. J. et al. Reduced signal for polygenic adaptation of height in UK Biobank. *eLife* **8**, e39725 (2019).
73. Cardon, L. R. & Palmer, L. J. Population stratification and spurious allelic association. *Lancet* **361**, 598–604 (2003).
74. Bulik-Sullivan, B. K. et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
75. Galton, F. Typical laws of heredity. III. *Nature* **15**, 512–514 (1877).
76. Galton, F. I. Co-relations and their measurement, chiefly from anthropometric data. *Proc. R. Soc. Lond.* **45**, 135–145 (1889).
77. Plana-Ripoll, O. et al. Exploring comorbidity within mental disorders among a Danish national population. *JAMA Psychiatry* **76**, 259–270 (2019).
78. Momen, N. C. et al. Association between mental disorders and subsequent medical conditions. *N. Engl. J. Med.* **382**, 1721–1731 (2020).
79. Polderman, T. J. et al. A genetic study on attention problems and academic skills: results of a longitudinal study in twins. *J. Canadian Acad. Child Adolesc. Psychiatry* **20**, 22–34 (2011).
80. Cardno, A. G., Rijsdijk, F. V., Sham, P. C., Murray, R. M. & McGuffin, P. A twin study of genetic relationships between psychotic symptoms. *Am. J. Psychiatry* **159**, 539–545 (2002).
81. Polderman, T., Hoekstra, R., Posthuma, D. & Larsson, H. The co-occurrence of autistic and ADHD dimensions in adults: an etiological study in 17 770 twins. *Transl. Psychiatry* **4**, e435–e435 (2014).

82. Bartels, M. et al. The five factor model of personality and intelligence: a twin study on the relationship between the two constructs. *Pers. Individ. Dif.* **53**, 368–373 (2012).
83. Plomin, R. & DeFries, J. Multivariate behavioral genetic analysis of twin data on scholastic abilities. *Behav. Genet.* **9**, 505–517 (1979).
84. Verweij, K. J., Huizink, A. C., Agrawal, A., Martin, N. G. & Lynskey, M. T. Is the relationship between early-onset cannabis use and educational attainment causal or due to common liability? *Drug Alcohol Depend.* **133**, 580–586 (2013).
85. Zietsch, B., Verweij, K., Bailey, J., Wright, M. & Martin, N. Genetic and environmental influences on risky sexual behaviour and its relationship with personality. *Behav. Genet.* **40**, 12–21 (2010).
86. Bulik-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
87. Watanabe, K. et al. A global overview of pleiotropy and genetic architecture in complex traits. *Nat. Genet.* **51**, 1339–1348 (2019).
88. van Rheenen, W., Peyrot, W. J., Schork, A. J., Lee, S. H. & Wray, N. R. Genetic correlations of polygenic disease traits: from theory to practice. *Nat. Rev. Genet.* **20**, 567–581 (2019).
89. Zhu, Z. et al. Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nat. Commun.* **9**, 224 (2018).
90. Verbanck, M., Chen, C.-y., Neale, B. & Do, R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat. Genet.* **50**, 693–698 (2018).
91. Keller, M. C. et al. The genetic correlation between height and IQ: shared genes or assortative mating? *PLoS Genet.* **9**, e1003451 (2013).
92. Hugh-Jones, D., Verweij, K. J. H., Pourcain, B. S. & Abdellaoui, A. Assortative mating on educational attainment leads to genetic spousal resemblance for polygenic scores. *Intelligence* **59**, 103–108 (2016).
93. Robinson, M. R. et al. Genetic evidence of assortative mating in humans. *Nat. Hum. Behav.* **1**, 0016 (2017).
94. Kemper, K. E. et al. Phenotypic covariance across the entire spectrum of relatedness for 86 billion pairs of individuals. *Nat. Commun.* **12**, 1050 (2021).
95. Turley, P. et al. Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat. Genet.* **50**, 229–237 (2018).
96. Baselmans, B. M. et al. Multivariate genome-wide analyses of the well-being spectrum. *Nat. Genet.* **51**, 445–451 (2019).
97. Lee, P. H. et al. Genomic relationships, novel loci, and pleiotropic mechanisms across eight psychiatric disorders. *Cell* **179**, 1469–1482 (2019).
98. Hill, W. et al. A combined analysis of genetically correlated traits identifies 187 loci and a role for neurogenesis and myelination in intelligence. *Mol. Psychiatry* **24**, 169–181 (2019).
99. Peyrot, W. J. & Price, A. L. Identifying loci with different allele frequencies among cases of eight psychiatric disorders using CC-GWAS. *Nat. Genet.* **53**, 445–454 (2021).
100. Anttila, V. et al. Analysis of shared heritability in common disorders of the brain. *Science* **360**, eaap8757 (2018).
101. Lee, P. H. et al. Genomic relationships, novel loci, and pleiotropic mechanisms across eight psychiatric disorders. *Cell* **179**, 1469–1482 (2019).
102. Grotzinger, A. D. et al. Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nat. Hum. Behav.* **3**, 513–525 (2019).
103. Demange, P. A. et al. Investigating the genetic architecture of noncognitive skills using GWAS-by-subtraction. *Nat. Genet.* **53**, 35–44 (2021).
104. Marees, A. T. et al. Genetic correlates of socio-economic status influence the pattern of shared heritability across mental health traits. *Nat. Hum. Behav.* <https://doi.org/10.1038/s41562-021-01053-4> (2021).
105. Lawlor, D. A., Harbord, R. M., Sterne, J. A., Timpson, N. & Davey Smith, G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Stat. Med.* **27**, 1133–1163 (2008).
106. Munafò, M. R. et al. Association between genetic variants on chromosome 15q25 locus and objective measures of tobacco exposure. *J. Natl Cancer Inst.* **104**, 740–748 (2012).
107. Voight, B. F. et al. Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. *Lancet* **380**, 572–580 (2012).
108. Rossouw, J. E. et al. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results from the Women's Health Initiative randomized controlled trial. *JAMA* **288**, 321–333 (2002).
109. Davey Smith, G. & Ebrahim, S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.* **32**, 1–22 (2003).
110. Abdellaoui, A. et al. Genetic correlates of social stratification in Great Britain. *Nat. Hum. Behav.* **3**, 1332–1342 (2019).
111. Kong, A. et al. The nature of nurture: effects of parental genotypes. *Science* **359**, 424–428 (2018).
112. Demange, P. A. et al. Parental influences on offspring education: indirect genetic effects of non-cognitive skills. Preprint at *bioRxiv* <https://doi.org/10.1101/2020.09.15.296236> (2020).
113. Selzam, S. et al. Comparing within- and between-family polygenic score prediction. *Am. J. Hum. Genet.* **105**, 351–363 (2019).
114. Cheesman, R. et al. Comparison of adopted and nonadopted individuals reveals gene–environment interplay for education in the UK Biobank. *Psychol. Sci.* **31**, 582–591 (2020).
115. Howe, L. J. et al. Within-sibship GWAS improve estimates of direct genetic effects. Preprint at *bioRxiv* <https://doi.org/10.1101/2021.03.05.433935> (2021).
116. Hur, Y.-M. & Craig, J. M. Twin registries worldwide: an important resource for scientific research. *Twin Res. Hum. Genet.* **16**, 1–12 (2013).
117. Hur, Y.-M. et al. Twin family registries worldwide: an important resource for scientific research. *Twin Res. Hum. Genet.* **22**, 427–437 (2019).
118. Abdellaoui, A., Verweij, K. J. H. & Nivard, M. G. Geographic confounding in genome-wide association studies. Preprint at *bioRxiv* <https://doi.org/10.1101/2021.03.18.435971> (2021).
119. Martin, J. et al. Association of genetic risk for schizophrenia with nonparticipation over time in a population-based cohort study. *Am. J. Epidemiol.* **183**, 1149–1158 (2016).
120. Munafò, M. R., Tilling, K., Taylor, A. E., Evans, D. M. & Davey Smith, G. Collider scope: when selection bias can substantially influence observed associations. *Int. J. Epidemiol.* **47**, 226–235 (2018).
121. Batty, G. D., Gale, C. R., Kivimäki, M., Deary, I. J. & Bell, S. Comparison of risk factor associations in UK Biobank against representative, general population based studies with conventional response rates: prospective cohort study and individual participant meta-analysis. *BMJ* **368**, m131 (2020).
122. Fry, A. et al. Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *Am. J. Epidemiol.* **186**, 1026–1034 (2017).
123. Adams, M. J. et al. Factors associated with sharing e-mail information and mental health survey participation in large population cohorts. *Int. J. Epidemiol.* **49**, 410–421 (2020).
124. Tyrrell, J. et al. Genetic predictors of participation in optional components of UK Biobank. *Nat. Commun.* **12**, 886 (2021).
125. Pirastu, N. et al. Genetic analyses identify widespread sex-differential participation bias. *Nat. Genet.* <https://doi.org/10.1038/s41588-021-00846-7> (2021).
126. Xue, A. et al. Genome-wide analyses of behavioural traits biased by misreports and longitudinal changes. *Nat. Commun.* **12**, 20211 (2021).
127. Mills, M. C. & Rahal, C. A scientometric review of genome-wide association studies. *Commun. Biol.* **2**, 9 (2019).
128. Pedersen, C. B. et al. The iPSYCH2012 case-cohort sample: new directions for unravelling genetic and environmental architectures of severe mental disorders. *Mol. Psychiatry* **23**, 6–14 (2018).
129. Stefansson, K. Letters from Iceland. *Nat. Genet.* **47**, 425 (2015).
130. Kerminen, S. et al. Geographic variation and bias in the polygenic scores of complex diseases and traits in Finland. *Am. J. Hum. Genet.* **104**, 1169–1181 (2019).
131. Bulik-Sullivan, B. K., Loh, P. R., Finucane, H. K., Ripke, S. & Yang, J. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
132. Martin, A. R. et al. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).

Acknowledgements

A.A. and K.J.H.V. are supported by the Foundation Volksbond Rotterdam. A.A. is also supported by ZonMw grant no. 849200011 from The Netherlands Organisation for Health Research and Development.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41562-021-01110-y>.

Correspondence should be addressed to A.A.

Peer review information *Nature Human Behaviour* thanks Anders Borglum and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

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

© Springer Nature Limited 2021