View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Explore Bristol Research

Timpson, N. J., Greenwood, C. M. T., Soranzo, N., Lawson, D. J., & Richards, J. B. (2018). Genetic architecture: The shape of the genetic contribution to human traits and disease. *Nature Reviews Genetics*, *19*(2), 110-124. https://doi.org/10.1038/nrg.2017.101

Peer reviewed version

Link to published version (if available):
10.1038/nrg.2017.101

Link to publication record in Explore Bristol Research
PDF-document

This is the author accepted manuscript (AAM). The final published version (version of record) is available online via Nature Publishing Group at DOI: 10.1038/nrg.2017.101 . Please refer to any applicable terms of use of the publisher.

University of Bristol - Explore Bristol Research

General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available:
http://www.bristol.ac.uk/pure/about/ebr-terms

Genetic architecture: the shape of the genetic contribution to human traits and disease

Nicholas J. Timpson[1], Celia M. T. Greenwood[2,3], Nicole Soranzo[4,5], Daniel J. Lawson[1] & J. Brent Richards [3,6,7]

1. MRC Integrative Epidemiology Unit, School of Social and Community Medicine, University of Bristol, Oakfield House, Oakfield Grove, Clifton, Bristol, BS8 2BN, UK.
2. Department of Oncology, McGill University, 3755 Cote Ste Catherine, Montreal, Quebec, Canada H3T 1E2
3. Departments of Human Genetics and Epidemiology, Biostatistics and Occupational Health, McGill University, 3755 Cote Ste Catherine, Montreal, Quebec, Canada, H3T 1E2
4. The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton CB10 1HH, Cambridge, UK.
5. Department of Haematology, University of Cambridge, Long Road, Cambridge CB2 0PT, UK.
6. Department of Medicine, Jewish General Hospital, Lady Davis Institute, McGill University, 3755 Cote Ste Catherine, Montreal, Quebec, Canada H3T 1E2
7. Department of Twin Research & Genetic Epidemiology, King's College London, St Thomas' Campus, Lambeth Palace Road, London, SE1 7EH, UK.

Correspondence to N. J. T. or J. B. R.: n.j.timpson@bristol.ac.uk; brent.richards@mcgill.ca

Genetic architecture describes the characteristics of genetic variation that are responsible for heritable phenotypic variability. It depends on the number of genetic variants affecting a trait, their frequencies in the population, the magnitude of their effects, and their interactions with each other and the environment. Defining the genetic architecture of a complex trait or disease is central to the scientific and clinical goals of human genetics, which are to understand disease etiology and aid in disease screening, diagnosis, prognosis and therapy. Recent technological advances have enabled genome-wide association studies (GWAS) and emerging next-generation sequencing studies to begin to decipher the nature of the heritable contribution to traits and disease. Here we describe the types of genetic architecture that have been observed, how architecture can be measured, and why an improved understanding of genetic architecture is central to future advances in the field and the influences that shape it.

[H1] Introduction

It can be argued that most of the challenges and rewards in human genetics are dependent upon scientists understanding genetic architecture so that they can more fully describe what causes disease and translate this information to the clinic. The term ‘genetic architecture’ in human population-based studies describes the characteristics of genetic variation that are responsible for broad sense phenotypic heritability **[G]**.[1] Specifically, genetic architecture comprises the number of variants influencing a phenotype **[G]** , the magnitude of their effects on the phenotype, the population frequency of these variants and their interactions with each other and the environment.[2] Thus, by contrast to narrow sense heritability, which only refers to the impact of additive genetic effects on complex traits **[G]** ,[3] genetic architecture refers broadly to a complete understanding of all genetic contributions to a given trait or disease outcome, as well as to an awareness of the characteristics of this contribution.[4] (**Box 1**)

Human genomes can differ from one another at single genomic positions as single nucleotide variants **[G]** (SNVs), or they can exhibit larger structural changes including copy number variations, translocations and inversions[5] (reviewed elsewhere[6]). To understand genetic architecture, variations in DNA sequence between genomes is tested for an association with phenotypic variability through gene-mapping studies, a field that has enjoyed success over the last decade.[7] These association signal mapping studies have increasingly become genome-wide association studies **[G]** (GWAS), whole-exome sequencing (WES) studies **[G]** and whole-genome sequencing (WGS) studies **[G]**.

GWAS use genome-wide genotyping arrays to measure genetic variation and they are the standard platform to test the association of a phenotype with common genetic variants. In this article, common genetic variants, low genetic variants and rare genetic variants are defined as those with a minor allele frequency **[G]** (MAF) of ≥5%, ≥1% but <5%, and <1%, respectively.[8] However, genotyping arrays can be designed to contain relatively rare variants. Furthermore, deep imputation **[G]** (discussed below) can be used to test phenotypic associations with additional low frequency and rare variants. As the least expensive modern genome-wide gene-mapping method, GWAS has been successfully employed in large human populations and has allowed a much improved understanding of the direct association of common variants (that is, not through interactions) with complex traits and diseases.[7] Many of these associations were found at non-coding variants, and these associations, including some that are driven by rare variants, were enriched at regulatory sites[9,10] Indeed it is now thought that up to 85% of all human common genetic variation is at least nominally associated with the expression of gene transcripts for protein coding genes.[11]

WES can identify rare variants associated with a phenotype, but is restricted to examining the protein-coding content of the genome. Although WES can identify genetic variants that are likely to directly influence gene function, the exome represents only about 1% of the genome,[12] and most disease-associated genetic variants that have been identified lie outside the exome.[7] WGS can measure nearly all genetic variation in the human genome and assess structural variants more accurately than WES,[13] but it costs much more than it. It is currently not possible to sequence all regions of the genome with equal quality. For example, regions with highly repetitive DNA are difficult to assess[14] and, therefore, not all genomic variation is captured by WGS. Furthermore, due to their cost, WGS studies have been limited by sample size and consequently may miss rare variants. However, the recent availability of large-scale cohort resources, such as the UKBiobank[15] and TopMed[16] programmes, combined with concurrent advances in WES, WGS and GWAS, will facilitate a more precise description of the contribution of low frequency and rare genetic variants to the genetic architecture of complex traits and disease.[17] Of note, GWAS, WES and WGS can also be used to estimate the narrow sense heritability of a trait or disease, and the resulting estimates have often been lower than those from classical twin heritability studies, which estimate heritability by contrasting the similarity in phenotypes between monozygotic and dizygotic twins. This has been reviewed previously[18] and this is therefore not a topic of this review.

Genetic architecture is often described as monogenic, oligogenic or polygenic, meaning that one, few or many genetic variants contribute to phenotypic variability, respectively.[19] In addition to this, a recent theoretical development in the modelling of genetic architecture has suggested that all complex traits and diseases share a single ‘omnigenic’ architecture.[20] This model suggests that gene regulatory networks may be sufficiently interconnected to allow all genes expressed in a disease-relevant cell to contribute to the disease phenotype. The omnigenic model posits that thousands of ‘non-core’ or ‘peripheral’ genes exert non-zero effects on essentially all downstream phenotypes.[20] An omnigenic architecture would help explain the complexity of genetic architecture, and it draws parallels to the ‘infinitesimal model’[21], in which all variants have a non-zero but small role in phenotypic variation. The omnigenic model also extends the idea of ‘universal pleiotropy’, which suggests that all characteristics are quantitative since, in principle, variation anywhere in the genome affects processes that are intimately related to all others.[22] These broad labels have been useful in theorizing the nature of genetic architecture, but modern techniques will enable the collection of data that will provide empirical evidence to instruct the description of genetic architecture.

In this Review, rather than categorizing the genetic architecture of many diseases and traits, we describe different types of genetic architecture, how they can be assessed, and why genetic architecture is important in biology and in the clinic. We then highlight some factors that influence genetic architecture before outlining outstanding challenges and opportunities in obtaining a more complete understanding of genetic architecture and translating this to patient care. It should be noted that we can only comment on genetic architectures that have been observed to date and we acknowledge that these observed architectures will change as the field evolves.

[H1] Types of genetic architecture

Whether a trait or disease has a monogenic, oligogenic, polygenic or omnigenic architecture, there is variability in the nature of the genetic contributions to phenotype. This variability is likely to be a function of both differences or deficiencies in phenotypic measurement and genuine biological heterogeneity. Hence, the number of discovered genetic variants, and the variety of other attributes that contribute to genetic architecture, can vary substantially between diseases.

To illustrate this, the genetic architecture of two well-studied diseases, type 1 diabetes mellitus and type 2 diabetes mellitus, can be compared (**Figure 1a**). Both diseases lead to hyperglycaemia, but type 1 diabetes mellitus is a disease of autoimmune dysregulation that leads to pancreatic β-cell dysfunction, whereas type 2 diabetes mellitus results from insulin resistance and relative insulin insufficiency.[23] Type 1 diabetes mellitus is polygenic and associated with low frequency and common variants that have comparatively large effects on disease risk, relative to other complex diseases.[24] By contrast, the risk of type 2 diabetes mellitus is associated with many genetic variants that have small effects on disease susceptibility.[25,26] As recent large-scale sequencing studies have not identified low frequency or rare variants of large effect associated with type 2 diabetes mellitus,[27] current data suggest that type 2 diabetes mellitus has a different genetic architecture to type 1 diabetes mellitus. The distinct pathophysiological mechanisms leading to these diseases may have evolved separately, leading to different architectures (see below). Since type 1 diabetes mellitus has a subset of observed variants that have a large effect on disease risk, genetics may help identify individuals at risk for this disease, providing the opportunity to influence disease progression (see below).

In contrast to diseases, biochemical traits are typically more proximally related to the function of a gene than complex diseases and common single nucleotide polymorphisms **[G]** (SNPs), taken together, are thought to contribute importantly to population-level variance.[28] However, biochemical traits may still have highly divergent observed architectures. For example, the biochemical traits of low density lipoprotein (LDL) cholesterol and 25-hydroxyvitamin D levels are both ~50% heritable **[G]** according to classical twin studies.[29,30] A GWAS of over 33,000 individuals found that only four loci were associated with the level of 25-hydroxyvitamin D.[31] One recent study looked for low-frequency and rare variants associated with low levels of vitamin D in 39,000 individuals through deep imputation[32], and a second recent study searched for additional novel common variants associated with vitamin D levels in 79,366 individuals;[33] these studies identified only four additional loci that are associated with 25-hydroxyvitamin D levels. Thus, the observed architecture of vitamin D level, at current sample sizes, is oligogenic and the associated genetic variants have comparatively large effects. By contrast, the level of LDL cholesterol seems to be influenced by many genetic variants with a broader distribution of effect sizes (**Figure 1b**).[34] In a study from the Global Lipids Consortium involving 188,577 individuals of European ancestry, 52 loci were associated with the level of LDL cholesterol.[34] Although the number of individuals in the study of LDL was larger than the number of individuals in the study of 25-hydroxyvitamin D, a smaller GWAS for LDL cholesterol levels in 19,000 individuals identified 11 LDL-associated loci,[35] again suggesting, within the capabilities of existing studies, that these two biochemical traits with similar heritability have fundamentally different genetic architectures. Importantly, these comparisons between traits are limited by differences in sample size, which can impact observed genetic architecture. For example, schizophrenia had few associated common genetic variants at samples of several hundred cases,[36] but at a sample size of tens of thousands, 113 genome-wide significant loci were observed.[37]

Even the same trait measured at different anatomical sites can have a divergent genetic architecture. For example, bone mineral density, a clinically relevant risk factor for osteoporotic fracture,[38] can be measured at different skeletal sites and is highly heritable.[39] Measuring bone mineral density at the forearm in 5,672 individuals identified only one locus associated with this trait. This locus contained the genes *WNT16* (encoding WNT16) and *CPED1* (encoding Cadherin-like and PC-esterase domain-containing protein 1).[40] Doubling this sample size identified no new loci, but did find a low frequency variant with a large effect size of 0.46 standard deviations per effect allele at the same *WNT16–CPED1* locus.[41] The same trait measured at the lumbar spine produces a contrasting genetic architecture; 19 independent loci were identified from 25,225 individuals, but the largest effect size of a single variant was only 0.22 standard deviations.[41] Although this difference in architecture could be a function of the different sample sizes, the available data suggest that architecture arising from common variants of highly similar traits can be remarkably different, and that these differences can be difficult to predict.

[H1] How to assess genetic architecture

The optimal conditions for elucidating genetic architecture are only achieved when all variable genotypes in the genome are measured in large populations in parallel with appropriate phenotyping. Although these conditions have not yet been achieved uniformly, studies have progressed towards this goal.[27,42–44] GWAS data make it possible to estimate the number of undiscovered additive genetic associations that contribute to the genetic architecture of a trait.[45] Such estimates can provide guidance when deciding whether to pursue WES and/or WGS after a large GWAS has generally described the effect of common variants. However, for some polygenic and complex traits a multi-pronged analysis is likely to be needed to elucidate genetic architecture.[42] Using triglyceride levels as an example, we describe an approach to partially resolve the architecture of a polygenic, complex trait.

GWAS has identified many common variants of varying effect size for lipid levels.[34] In the case of triglycerides (which are a type of lipid), targeted genotyping,[46] GWAS,[47] WES[48] and WGS[49] have identified common variants of small effect as well as rare variants of larger effect that are associated with triglyceride levels and that are located in and near *APOC3*, the gene encoding apolipoprotein C3. The identification of large and small effect size variants at the same gene allows scientists to create a dose-response curve of genetic variants, where the effect of the genetic variant on protein function is plotted against the effect on phenotype; this curve helps to predict how pharmaceuticals targeted at the phenotype will affect the drug target.[50] Although it is difficult to draw direct comparisons between the predictions made from short-term trials with those made on the basis of genetic effects that are often exerted over a lifetime, the exploitation of information like this for drug development is useful.[51,52] Indeed, APOC3 has been therapeutically targeted using an antisense inhibitor of APOC3 synthesis, resulting in lowered triglyceride levels in humans, as would be predicted from the dose response curve.[53] This suggests that a mixed discovery strategy aimed at identifying small and large effect-sized variants may be beneficial to describing genetic architecture when many variants of varying effect size are implicated.

As noted above, GWAS has been limited to common variants (MAF ≥5%), but variants with lower MAFs can now be estimated from genome-wide genotyping arrays using accurate deep imputation, which leverages the genotyping scaffold available from genome-wide genotyping to impute missing genetic variation at millions of additional genomic sites.[54] This is achieved by comparing the haplotypes **[G]** observed in individuals subjected to genome-wide genotyping to those seen in an imputation reference panel **[G]** , which is a set of haplotypes derived from WGS.[54] Through deep imputation, adequately powered GWAS can capture, in samples from individuals of European ancestry, the contribution to genetic architecture of genotypes with a MAF of approximately ≥0.1%.[42] WES and WGS are generally used to explore the contribution of genetic variants with a MAF lower than 0.1%. However, WES and WGS genetic association mapping strategies suffer from low statistical power since single SNV association test **[G]** power decreases as the minor allele becomes rarer. Furthermore, since WES and WGS are expensive, their sample sizes are generally small. This further decreases the statistical power to reliably identify associations using these study designs.

To overcome the reduced statistical power of WES and WGS studies in assessing the contribution of rare genetic variants to human traits and disease, region-based testing **[G]** is often used to ‘collapse’ information across a genomic region and test the association of the region with the phenotype; this strategy aims to improve statistical power.[55] However, these tests have important limitations (**Box 2**) and thus have not often led to new insights.

Prior to the population-based sequencing era, it was anticipated that low frequency and rare genetic variants would display very large effect sizes and hence explain some of the missing heritability (that is, heritability that cannot be explained by common SNVs).[56] However, testing low frequency and rare genetic variants separately has revealed that they are not always associated with large effect sizes. For example, the UK10K project used single SNV association tests for >13 million SNVs with a MAF ≥0.1% to test their association with more than 30 traits in 3,781 individuals.[41] The study had 80% statistical power to detect associations for alleles with effect sizes of at least ~1.2 standard deviations on the trait at genetic variants with MAF as low as 0.5%. There was little evidence that alleles with a MAF in this range had effects on traits larger than anticipated based on the power curve threshold (**Figure 2;** where the power curve defines the bound of statistical power, given the effect size and MAF). By contrast, several larger studies relying on single SNV association testing together with deep imputation have identified novel associations of large effect between low frequency variants and common traits, such as bone mineral density in the GEFOS Consortium, height in the GIANT Consortium and lipid levels in the Global Lipids Consortium.[41,57–59] As the field progresses towards larger sample sizes, through the availability of resources such as UKBiobank and TopMed, increasingly rare genetic variants with larger effect sizes are likely to be identified from single SNV association testing. We anticipate that, of the methods currently available, this method will enhance our knowledge of genetic architecture the most. Of note, it has been suggested that common genetic variant signals may be explained by their synthetic association with rare genetic variants.[60] Although this is a logical hypothesis, and some synthetic associations between common genetic variant signals and rare genetic variants have been observed,[10] most common variants to date have not been found to be driven by synthetic associations.[61,62]

[H1] When is genetic architecture important?

Genetic architecture is important for screening for and diagnosing disease, enhancing biological understanding, drug development, Mendelian randomization and the scientific pursuit of gene mapping. Here we describe the role of genetic architecture in each of these aspects of human genetics.

[H3] Screening and diagnosis.

The genetic architecture of a disease can influence both an individual's susceptibility to the disease and the variance in the population that can be explained by genetic factors.[63] Here we try to disentangle these concepts, which are often conflated.

An individual's genetic susceptibility to disease is the sum of the effects of independent disease-causing genetic variants and their interactions, and it is independent of the frequency of the disease-causing alleles in the population. However, variance explained **[G]** in the population depends on the number of disease-causing alleles and their frequencies and effect sizes, and it is thus a function of genetic architecture. One commonly-used measure of variance explained assumes that: the variants contribute to the additive genetic variance component only; that variants have small effect sizes, such that linearity approximately holds (that is, the cumulative effect of all variants can be approximated by their sum); and that the variant 'is' the causal variant (that is, its association with the phenotype is not mediated through another variant in linkage disequilibrium **[G]**).[64–66] The proportion of variance explained, under these assumptions, has been expressed for continuous phenotypes as $2p(1-p)\beta^2$, where p is the effect allele frequency, and β is the effect of the allele on a standardized phenotype that has a mean of zero and a variance of 1. Thus, the frequency of the disease-associated allele helps explain variation in the population, even though it is not relevant when describing an individual's susceptibility to disease. This has important implications for the use of genetic architecture in the diagnosis, prognosis and treatment of human diseases.

The utility of a diagnostic test is often evaluated by assessing the area under a receiver operator curve **[G]** (ROC), which combines information from the sensitivity and specificity of a test for a binary outcome. Variance explained influences the specificity and sensitivity of genetic diagnostic tests and this is reflected in ROCs; as more variance in a phenotype is explained the area under the ROC will increase.[63] As the amount of variance explained by genetic factors for most common diseases is currently low, the clinical utility of a ROC based on genetic factors is low. For example, a genetic risk score for osteoporosis in a study employing genetic variants that explained 5.8% of variance in bone mineral density (the clinically-relevant marker of osteoporosis) indicated a risk of osteoporosis that was not importantly different from the risk that would be expected by chance.[67] However, variance explained by genetic factors in rare monogenic diseases such as cystic fibrosis can approach 100%.[68] Consequently, at present, disease-associated genetic variants can be used to diagnose cystic fibrosis but not osteoporosis. [69] The reason for this difference in clinical care is due to the amount of variance explained by the known genetic variants. Further, the accuracy of a diagnostic test will increase with the prevalence of the disease in the population. Note that in this Review we define accuracy as the proportion of all diagnostic test results (both positive and negative) that are correct.

Thus, genetics can aid the diagnosis of rare diseases in which most phenotypic variation is explained by known and highly penetrant genetic variants. However, the genetic architecture of most common diseases does not currently permit the use of genetics in diagnosis and screening, due to low variance explained. This situation will change as the variance explained in common disease risk by SNVs increases as the sample size of gene mapping studies increases, thus enabling the detection of smaller effects from common variants and larger effects from low-frequency and rare variants.[70]

[H3] Biological understanding and drug development.

Some have suggested that the small amount of variance explained for most common SNPs for common diseases precludes their utility in drug target identification. This concept can be misguided in the absence of further information about the genetic architecture of the disease association. In order to understand the relevance of a small effect size SNP to drug development, one must first understand the effect of that SNP on protein level or function. Even if a SNP has a small effect on protein level and disease risk, this protein may still be a suitable target for disease prevention. This is because a small effect on protein level which translates to a small effect on disease risk may be consistent with a large effect on protein level which translates to a large effect on disease risk. Similarly, if a drug has a small effect on a protein level and has a small effect on disease risk, it remains possible that a drug having a larger effect on protein level may have a larger effect on disease risk. Further, such small effect sizes from SNPs may be particularly informative if the SNP has a large effect on protein level and no effect on disease risk. In such situations, the protein would consequently be less attractive as a drug target.

The clinical effect of drugs on LDL cholesterol level and cystic fibrosis illustrate the dichotomy between variation explained and its utility to drug development. The activity of 3-hydroxy-3-methylglutaryl-coenzyme A reductase (HMG-CoA reductase) partially determines the level of circulating LDL cholesterol.[71] Pharmacological inhibition of HMG-CoA reductase reduces the level of LDL cholesterol by approximately 30–40%, which reduces the incidence of coronary heart disease.[72] The common SNP most strongly associated with LDL cholesterol level near *HMGCR*, the gene encoding HMG-CoA reductase, explains 0.26% of the variance in LDL cholesterol level, which is clearly a small amount (rs12916, MAF 0.4).[34] Thus, even though the HMG-CoA reductase locus harbours a common genetic variant that explains only a small amount of phenotypic variation, the pharmacologic inhibition of HMG-CoA reductase is clinically beneficial. Indeed there are many other reported cases in which genetic variants near the drug target have small effects on the phenotype, yet pharmacological manipulation of the drug target has profound effects on phenotype.[73] For example, common variants near *PCSK9*, the gene encoding proprotein convertase subtilisin/kexin type 9 (PCSK9), have small effects on LDL cholesterol level[74] whereas pharmacological inhibition of PCSK9 has large effects.[75] Furthermore, *RANKL,* the gene encoding receptor activator of nuclear factor κ-B ligand (RANKL), harbours common variants of small effect on bone mineral density,[67] yet pharmacological inhibition of RANKL has large effects on bone mineral density.[76] Thus, small effect size SNPs can serve to highlight proteins that, when targeted with large effect size pharmaceuticals, can have large effects on disease risk.

As a contrasting example, nearly all patients with cystic fibrosis have mutations in *CFTR*, thus the variance explained by genetic variation in *CFTR* approaches 100%.[68] However, despite the discovery of the association between *CFTR* and cystic fibrosis in 1989, the only drug targeting cystic fibrosis transmembrane conductance regulator (CFTR) was approved 23 years later and is partially efficacious in only 4.4% of cystic fibrosis patients.[77] Therefore, even when nearly 100% of the phenotypic variance **[G]** is explained by a few genetic variants at a single gene, pharmacologic therapies against the identified gene may not immediately advance patient care.

The amount of variance explained by a genetic variant does not always correlate with the suitability of the gene as a therapeutic target because drugs work on proteins; the base pair associated with the disease serves to help identify the causal protein. The relevance of the variation explained to the clinic should be measured by assessing the effect of pharmacological agents on the protein and its resultant effect on disease. Furthermore, a gene which has no variants that are associated with a disease, perhaps because natural selection makes such perturbing genetic variants so rare that they lack statistical power for such an association, could still be a good drug target.

How would a truly omnigenic architecture affect drug development? If all expressed genes in a cell that influences a phenotype had equivalent effects on phenotypic variance, then pharmacological manipulation of any of the expressed proteins would have an impact on the phenotype. This situation is unlikely as most drugs fail in drug development pipelines because they do not affect the phenotype, despite evidence of their engagement with the drug target.[78] Thus, there must be a gradient of effect of the impact of different proteins on phenotype, where some genes have large effects and other genes have smaller effects. This suggests that a set of ‘core’ genes must have a more pronounced effect on phenotype and that the proteins derived from these genes will drive pharmaceutical development.

[H3] Mendelian randomization.

Mendelian randomization is an established genetic epidemiology method that can provide evidence supporting, or contradicting, the causality of a risk factor in disease.[79] This method uses SNVs as proxies for risk factors and can help to address confounding **[G]** and reverse causation **[G]**. Confounding is theoretically prevented in this method since SNVs are randomly allocated at conception, thereby breaking their association with factors not in the causal pathway. This situation is similar to the randomization process which prevents confounding in a randomized trial. Reverse causation is eliminated in this method since SNV allocation always precedes disease onset and cannot be altered by it. One of the main assumptions of Mendelian randomization is an absence of horizontal pleiotropy **[G]** , in which the genetic variant influences the outcome in a manner independent of the risk factor. Horizontal pleiotropy is distinct from vertical pleiotropy **[G]** ; the latter is defined as the association of the genetic variant with other traits in the same pathway due to its effect on the risk factor.[80,81] Mendelian randomization studies rely upon vertical pleiotropy, but can be biased by horizontal pleiotropy. Knowledge of genetic architecture can help to detect the presence of pleiotropy and to guard against it.

A polygenic architecture provides the opportunity to undertake sensitivity testing to identify the presence of horizontal pleiotropy through the Mendelian randomization-Egger (MR-Egger) test,[82] (reviewed elsewhere).[83] MR-Egger aims to account for, and address, the presence of unbalanced horizontal pleiotropy by assessing whether the intercept is different from the origin when plotting the relationship between the SNV on the outcome versus the SNV on the exposure. Unbalanced horizontal pleiotropy would lead to SNVs with a systematically higher or lower effect on the outcome than on the exposure, as they act upon the outcome through exposure-independent pathways. An omnigenic genetic architecture has implications for Mendelian randomization because it suggests universal pleiotropy in the human genome, however whether omnigenic pleiotropy is horizontal and balanced, horizontal and unbalanced, or vertical must be considered.

It has been suggested that omnigenic pleiotropy violates Mendelian randomization assumptions when two phenotypes with omnigenic architectures are influenced by the same tissue type as this situation would result in horizontal pleiotropy[18]. However in one example for bone tissue (Figure 3), current data strongly suggest that unbalanced horizontal pleiotropy does not exist in two highly polygenic (and possibly omnigenic) traits, bone mineral density and height, both of which are influenced by the same tissue.

Beyond the omnigenic model, it is clear that the expression of certain genes in some cells causes meaningful biological changes in other cells. Indeed, signalling molecules dominate in endocrinology, whereby complex homeostatic feedback systems regulate many central biological processes. For example, insulin secreted by the pancreas causes glucose uptake in different cell types, such as skeletal muscle.[85] These are examples of vertical pleiotropy, but not horizontal pleiotropy, and therefore do not violate the assumptions of Mendelian randomization.

[H1] What influences genetic architecture?

Here we describe some of the major factors that influence genetic architecture and discuss how understanding these determinants of architecture can help to improve our understanding of the genetic determinants of common diseases and traits.

[H3] Phenotype.

Phenotypes vary in how they relate to underlying genetic variation, their interaction with the environment, and by the quality of their measurement; all of these parameters contribute to observed genetic architecture. In contrast to genuinely polygenic complex traits, some molecular traits or medical conditions can have relatively large portions of variance predicted by one or a few relatively large genetic contributions. Examples of these molecular traits include the levels of C-reactive protein and[86] uric acid,[87] and age-related macular degeneration.[88] In cardiovascular disease, rare variants of large effect can lead to severe monogenically controlled lipid disorders[89], and there are several other polygenic traits for which heritability is high but the number of major contributing loci is relatively low (reviewed in reference 86).[90] Measurement can also introduce complexity. For example, when measuring educational attainment,[91] the observable phenotype (that is, years of schooling or college attendance) is likely to capture many factors marking a collection of biological pathways that contributed to the analysed outcome. Consequently, in the presence of adequate analytical power, the architecture of this trait will have polygenic characteristics as a result of the broad-spectrum measurement.

This relationship between the measurement of phenotype and genetic architecture has implications for the interpretation and utilization of genetic variation in applied genetic and epidemiological analyses. The extent of horizontal pleiotropy[92] can be estimated and analytical methods can use genetic associations to assess the overlap in heritable contribution between traits; for example linkage disequilibrium score regression[93] can assess shared, narrow-sense heritability. However, these analyses **[Au:OK? Or, please clarify 'these']** cannot change what has been measured. Our limited approaches to population-based phenotyping are likely to produce situations whereby apparently different traits in any given study are actually distal measures of the same underlying biological events. In this case, it is measurement that has shaped our interpretation of shared genetic architecture and a perceived phenotypic dependence is a consequence of the difficulty in directly measuring biology.

[H3] Selection.

Selection is the evolutionary process by which the frequency of genetic variation changes in response to a fitness consequence in the local environment. We will show examples where genetic architecture may have been influenced by the nature of the trait of interest, the relative age and effect of the mutations that explain its variation, and the characteristics of the population being assessed, contribute to selection and will influence genetic architecture. As a motivating example, common genetic variants with large effect could not exist if purifying selection removed them from the population[42] (**Figure 2**). Although it is difficult to prove how selection has directly influenced genetic architecture, several natural experiments inform the relationship between complex trait genetic architecture and selection.

Firstly, effective population size may influence observed genetic architecture by reducing the strength of selection[94,95], allowing deleterious variants to increase in frequency by genetic drift (**Figure 4a**). Small population sizes allow variants of any frequency to change more rapidly — akin to a founder effect **[G]** — and by chance, the frequency of some functional alleles can drift upwards so that they provide sufficient statistical power to detect their effect.[96] Drift has been exploited by GWAS using isolated populations to enhance analytical power in otherwise limited sample sizes. An example of this is the Kosrae Pacific island population, individuals of which have a high prevalence of type 2 diabetes mellitus, thought to be a consequence of a founder event.[97] Further, the frequencies of variants that cause type 2 diabetes mellitus had likewise changed in this population compared with populations on nearby islands. However, little evidence of different effect sizes from individuals of European descent was found, meaning that any individual possessing a given variant of the allele has the same increase in risk of type 2 diabetes mellitus, regardless of ancestry.[97] Furthermore, genetic drift may contribute to differences in longevity between Greek island populations, despite similarities in culture.[98] SNPs that both increase or decrease longevity can be found at corresponding frequency in such drifted populations.

Secondly, populations share different histories, which can affect genetic architecture (**Figure 4b**). For example, infection prevalence naturally varies, resulting in differential selection over time[99], and the genetic architecture of infectious disease resistance varies from Mendelian to highly complex.[100] This genetic architecture can theoretically be linked to the diverse evolutionary responses of the immune system[101]**.** An important example is the human leukocyte antigen (HLA) locus, which encodes the major histocompatibility complex that allows the immune system to distinguish 'self' from 'non-self'.[102] As individuals with the same HLA variant will be susceptible to similar infectious disease strains, recombination at this locus (and observed variation at the level of the population) is structured to provide offspring with a different resistance phenotype to their parents.[103] However, this has not occurred for resistance to human malaria, which is caused by the same variant that causes sickle cell disease.[104] In this case, antagonistic horizontal pleiotropy has allowed sickle cell disease to be maintained at relatively high frequency in populations exposed to malaria.[105] Horizontal pleiotropy appears to maintain phenotypic diversity across many culturally regulated human phenotypes.[106] Over evolutionary timescales, we therefore expect genetic architecture to change where selection is strong and mutations arise.

Determining the extent by which selection influences genetic architecture requires adequate measurement of variants under selection. Strategies to detect selection in the genomes of contemporary populations include examination of functional variation, allele frequency variation, population differences and haplotype profiles (reviewed elsewhere).[107] These methods agree that strong and recent signatures of selection have radically altered the frequency profiles of specific variants; for example, lactase persistence **[G]** and haemoglobinopathy-linked malarial resistance have both been detected in this way[108–111]. Signals as young as 2,000 years may be detected in the patterns of singleton **[G]** variation.[112] Strong selection acting on traits influences genetic architecture because the anticipated phenotypic effect for a SNP of a given frequency is distorted (**Figure 4b**) relative to the phenotypic effects that are predicted by the dose-response curve.

Despite difficulties in measuring selection, many genetic variants might be subject to subtle selection mechanisms acting in a polygenic model. Evidence has emerged for a 'coordinated shift in allele frequency'[113] in, for example, height[112] and educational attainment[114] across different populations. Specifically, GWAS for complex traits in large population-based collections have yielded evidence for polygenic contributions to complex traits which also demonstrate detectable and trait specific differences in allele frequency across populations (**Figure 5**).[115,116] Together, these observations are suggestive of polygenic selection, where even in the presence of relatively small phenotypic effects, coordinated action across many loci will ultimately have an effect on the genetic architecture of the trait in question.

[H3] Decanalization.

Canalization[117] maintains physiologic homeostasis through plastic responses to environmental or endogenous perturbations. Important cellular systems with long evolutionary histories are likely to be canalized; for example, body temperature is regulated to remain constant regardless of environment in humans, but not in all species.[118] Decanalization is the hypothetical process whereby well-canalized systems can be destabilized by changes in environment or by the introduction of large-effect size genetic variants.[117] The decanalizing effect generated by strong perturbations of long-standing homeostatic processes can lead to disease.[117] Here we provide examples of genetic and environmental decanalization that have led to specific genetic architectures.

Genetic variants of large effect, which hypothetically should become rare through negative selection, may cause perturbations that cannot be physiologically adapted to, thereby creating decanalization events. The effect size of variants which drive decanalization events can, in fact, be substantially larger than the effect size predicted by their MAF[41,42,57,119]. For example, the changes in bone mineral density owing to low-frequency genetic variants associated near *EN1* (the gene encoding homeobox protein engrailed-1) are four-fold larger than the mean changes caused by common variants and are in excess of that expected for the frequency of the associated variant; this may therefore represent an example of genetic decanalization.[41] An example of environmental decanalization is the large change in carbohydrate intake in Inuit (a group of culturally similar indigenous individuals inhabiting Arctic regions) after the introduction of Western diets over the past 60 years, which is thought to have precipitated the discovery of common alleles with a large effect on the risk of type 2 diabetes mellitus and glucose dysregulation (**Box 3**).[120,121]

The interaction between canalization and selection can be used to understand complex traits (**Figure 4c**). Differences in selection across different populations enables admixture mapping **[G]** ,[122] which provides an opportunity to further understand genetic architecture. Since admixed individuals carry different proportions of their ancestral population genomes, it is possible to explore whether the effect size at a causal locus varies as a function of the ancestry proportions across the rest of the genome. This enables inference about whether causal variants act independently and additively, or if more complex relationships are likely. For example, consider a variant that affects a canalized phenotype. If other variants across the genome also affect the phenotype and further vary in frequency by population, then the detected effect size should depend on the ancestry mixture.[121]

[H1] Gene and environment interactions [Au:OK to reduce length to <39 characters, including spaces?]

The effect of a genetic variant may vary depending on the level of an environmental determinant of the trait (gene x environment interactions), or by the number of alleles at another genetic variant (gene x gene interactions). This is not a focus of this Review and it has been discussed elsewhere[123]; however, it should be noted that there is not yet strong evidence that gene x environment or gene x gene interactions play a predominant role in determining most complex phenotypes. For example, recent evidence suggests that amongst the potential environmental determinants of body mass index (BMI), which together explain 14% of phenotypic variation, there is only evidence for interactions between genotype and smoking.[124] All other environmental determinants of BMI had genetic interactions effects of 1% or less of the total phenotypic variance. This finding is supported by a general deficit of replicated gene x environment interactions in the literature. Consequently, although some interactions must exist, as yet these do not explain a large proportion of phenotypic variance and therefore do not strongly influence observed genetic architecture.

Migration studies are important for comparing genetic and environmental risk factors and their interaction. For diseases primarily related to lifestyle and diet, including obesity,[125–127] heart disease,[128] inflammatory bowel disease,[129] tuberculosis,[130] and several cancers,[131,132] migrants transition between the risk associated with their original population and their assimilating population. Studying individuals that migrate allows researchers to explore the relative role of environment in each disease, and hence the conclusion varies depending on the genetic architecture and contribution of each.

[H1] Summary and conclusions

The scientific drive behind exploring and understanding genetic architecture follows a desire to explain and understand all of the genetic contributions to phenotypic variance, which has been a goal in quantitative genetics for more than a century.[21] It will become possible to empirically describe near-complete genetic architecture for some traits. Alongside a growing collection of analytical approaches addressing the phylogenetic relationships between complex traits and diseases,[93] the availability of genetic and phenotypic data in increasingly large population-based studies, such as UKBiobank[15], will inevitably add to our understanding of relative genetic contributions.

A more complete understanding of the genetic architecture of complex traits and diseases will maximize the utility of human genetics in disease screening, diagnosis, prognosis and therapy. Importantly, variance explained is strongly related to genetic architecture but it is not essential for drug development and individual-level risk prediction. The past decade of gene mapping in complex traits and diseases has shown that their genetic architectures are highly variable and difficult to predict. Nonetheless, clear trends have emerged, demonstrating that phenotypes that are reliably and inexpensively measured and more proximal to the effects of genetic variation are more amenable to the tools used to dissect their genetic architecture. Subject to measurement, the ultimate architecture of many traits may well be infinitesimal[21], and this will affect the clinical goals of genetics; however, some genes have more important roles in disease causation than others, and some of these genes can be targeted for drug development.[51] Drug developers should always consider the effect of the SNV on the function of the encoded protein when assessing the magnitude of the SNV's effect on disease risk. Small effect sizes of SNVs on disease can be highly relevant to drug development when they have large effects on protein level or function, suggesting that the protein target is not appropriate for that disease. Furthermore, small effect sizes of SNVs can also highlight proteins, the pharmacological manipulation of which has large effect sizes on disease.

Finally, understanding how the forces of natural selection and decanalization have influenced differing architectures across populations will be particularly helpful as the field moves to more fully characterize architectures in non-European ancestries. Architecture can be most easily measured through single-base pair testing and this approach has produced most of the loci associated with traits and common diseases. By contrast, rare variant collapsing tests are difficult to define, interpret and compare across traits. Thus, most advances in understanding allelic architecture will likely arise in the short-term through single-base pair testing in very large populations. Many of the greatest challenges and rewards in human genetics over the next decade will rely upon understanding genetic architecture to more fully appreciate the biologic mechanisms that translate varying architectures to disease susceptibility.

References:

1. Mackay, T. F. C. The Genetic Architecture of Quantitative Traits. *Annu. Rev. Genet.* **35,** 303–339 (2001).
2. Gratten, J., Wray, N. R., Keller, M. C. & Visscher, P. M. Large-scale genomics unveils the genetic architecture of psychiatric disorders. *Nat. Neurosci.* **17,** 782–90 (2014).
3. Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics era--concepts and misconceptions. *Nat. Rev. Genet.* **9,** 255–266 (2008).
4. Hansen, T. F. The Evolution of Genetic Architecture. *Annu. Rev. Ecol. Evol. Syst.* **37,** 123–157 (2006).
5. Frazer, K. A., Murray, S. S., Schork, N. J. & Topol, E. J. Human genetic variation and its contribution to complex traits. *Nat. Rev. Genet.* **10,** 241–251 (2009).
6. Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* **12,** 363–376 (2011).
7. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **101,** 5–22 (2017).
8. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42,** D1001-6 (2014).
9. Schork, A. J. *et al.* All SNPs Are Not Created Equal: Genome-Wide Association Studies Reveal a Consistent Pattern of Enrichment among Functionally Annotated SNPs. *PLoS Genet.* **9,** e1003449 (2013).
10. Astle, W. J. *et al.* The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell* **167,** 1415–1429.e19 (2016).
11. Aguet, F. *et al.* Genetic effects on gene expression across human tissues. *Nature* **550,** 204–213 (2017).
12. Clark, M. J. *et al.* Performance comparison of exome DNA sequencing technologies. *Nat. Biotechnol.* **29,** 908–914 (2011).
13. Zarrei, M., MacDonald, J. R., Merico, D. & Scherer, S. W. A copy number variation map of the human genome. *Nat. Rev. Genet.* **16,** 172–183 (2015).
14. Treangen, T. J. & Salzberg, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.* **13,** 36–46 (2011).

15. Allen, N. E., Sudlow, C., Peakman, T. & Collins, R. UK Biobank Data: Come and Get It. *Sci. Transl. Med.* **6,** 224ed4-224ed4 (2014).

16. NIH. Trans-Omics for Precision Medicine (TOPMed) Program. (2017).

17. Day, F. R. *et al.* Physical and neurobehavioral determinants of reproductive onset and success. *Nat. Genet.* **48,** 617–623 (2016).

18. Tenesa, A. & Haley, C. S. The heritability of human disease: estimation, uses and abuses. *Nat. Rev. Genet.* **14,** 139–149 (2013).

19. Badano, J. L. & Katsanis, N. Beyond Mendel: an evolving view of human genetic disease transmission. *Nat. Rev. Genet.* **3,** 779–789 (2002).

20. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169,** 1177–1186 (2017).

21. Fisher, R. A. The correlation between relatives on the supposition of Mendelian inheritance. *Proc. Roy. Soc. Edinburgh* **52,** 99–433. (1918).

22. Kacser, H. & Burns, J. A. THE MOLECULAR BASIS OF DOMINANCE. *Genetics* **97,** 639-666 (1981). **[Au: Page numbers OK?]**

23. Harris, M. I. Impaired glucose tolerance in the U.S. population. *Diabetes Care* **12,** 464–74 (1989).

24. Polychronakos, C. & Li, Q. Understanding type 1 diabetes through genetics: advances and prospects. *Nat. Rev. Genet.* **12,** 781–92 (2011).

25. Morris, A. D. P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat. Publ. Gr.* **44,** 981–990 (2012).

26. Scott, R. A. *et al.* An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans. *Diabetes* db161253 (2017). doi:10.2337/db16-1253

27. Fuchsberger, C. The genetic architecture of type 2 diabetes. *Nature* (2016). doi:10.1038/nature18642

28. Hindorff, L. a *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 9362–9367 (2009).

29. Pilia, G. *et al.* Heritability of Cardiovascular and Personality Traits in 6,148 Sardinians. *PLoS Genet.* **2,** e132 (2006).

30. Shea, M. K. *et al.* Genetic and non-genetic correlates of vitamins K and D. *Eur. J. Clin. Nutr.* **63,** 458–64 (2009).

31. Wang, T. J. *et al.* Common genetic determinants of vitamin D insufficiency: a genome-wide association study. *Lancet* **376,** 180–188 (2010).

32. Manousaki, D. *et al.* Low-Frequency Synonymous Coding Variation in CYP2R1 Has Large Effects on Vitamin D Levels and Risk of Multiple Sclerosis. *Am. J. Hum. Genet.* **101,** 227–238 (2017).

33. Jiang, X. *et al.* The Genetic Architecture of Vitamin D. in *American society of Human Genetics*

34. Willer, C. J. *et al.* Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45,** 1274–1283 (2013).

35. Kathiresan, S. *et al.* Common variants at 30 loci contribute to polygenic dyslipidemia. *Nat. Genet.* **41,** 56–65 (2009).

36. O'Donovan, M. C. *et al.* Identification of loci associated with schizophrenia by genome-wide association and follow-up. *Nat. Genet.* **40,** 1053–1055 (2008).

37. Li, Z. *et al.* Genome-wide association analysis identifies 30 new susceptibility loci for schizophrenia. *Nat. Genet.* (2017). doi:10.1038/ng.3973

38. Kanis, J. A. *et al.* Interpretation and use of FRAX in clinical practice. *Osteoporos. Int.* **22,** 2395–2411 (2011).

39. Arden, N. K., Baker, J., Hogg, C., Baan, K. & Spector, T. D. The heritability of bone mineral density, ultrasound of the calcaneus and hip axis length: a study of postmenopausal twins. *J. Bone Miner. Res.* **11,** 530–534 (1996).

40. Zheng, H.-F. F. *et al.* WNT16 Influences Bone Mineral Density, Cortical Bone Thickness, Bone Strength, and Osteoporotic Fracture Risk. *PLoS Genet.* **8,** e1002745 (2012).

41. Zheng, H. *et al.* Whole-genome sequencing identifies EN1 as a determinant of bone density and fracture. *Nature* **526,** 112–7 (2015).

42. Walter, K. *et al.* The UK10K project identifies rare variants in health and disease. *Nature* **526,** 82–90 (2015).

43. Sidore, C. *et al.* Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. *Nat. Genet.* **47,** 1272–1281 (2015).

44. Gudbjartsson, D. F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* (2015). doi:10.1038/ng.3247

45. Yang, J. *et al.* Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat. Genet.* (2015). doi:10.1038/ng.3390

46. Jørgensen, A. B., Frikke-Schmidt, R., Nordestgaard, B. G. & Tybjærg-Hansen, A. Loss-of-function mutations in APOC3 and risk of ischemic vascular disease. *N. Engl. J. Med.*

371, 32–41 (2014).

47. Kathiresan, S. *et al.* Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nat. Publ. Gr.* **40,** 189–197 (2008).
48. Heart, N. & The TG and HDL Working Group of the Exome Sequencing Project, National Heart, Lung, and Blood Institute. Loss-of-Function Mutations in APOC3,Triglycerides, and Coronary Disease. *N. Engl. J. Med.* 140618140014007 (2014). doi:10.1056/NEJMoa1307095
49. Timpson, N. J. *et al.* A rare variant in APOC3 is associated with plasma triglyceride and VLDL levels in Europeans. *Nat. Commun.* **5,** 4871 (2014).
50. Plenge, R. M., Scolnick, E. M. & Altshuler, D. Validating therapeutic targets through human genetics. *Nat. Rev. Drug Discov.* **12,** 581–594 (2013).
51. Nelson, M. R. *et al.* The support of human genetic evidence for approved drug indications. *Nat. Genet.* **47,** 856–860 (2015).
52. Ference, B. A., Majeed, F., Penumetcha, R., Flack, J. M. & Brook, R. D. Effect of naturally random allocation to lower low-density lipoprotein cholesterol on the risk of coronary heart disease mediated by polymorphisms in NPC1L1, HMGCR, or both: a 2 × 2 factorial Mendelian randomization study. *J. Am. Coll. Cardiol.* **65,** 1552–61 (2015).
53. Gaudet, D. *et al.* Antisense Inhibition of Apolipoprotein C-III in Patients with Hypertriglyceridemia. *N. Engl. J. Med.* **373,** 438–47 (2015).
54. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* (2016). doi:10.1038/ng.3643
55. Ladouceur, M., Dastani, Z., Aulchenko, Y. S., Greenwood, C. M. T. & Richards, J. B. The Empirical Power of Rare Variant Association Methods: Results from Sanger Sequencing in 1,998 Individuals. *PLoS Genet.* **8,** e1002496 (2012).
56. Bodmer, W. & Bonilla, C. Common and rare variants in multifactorial susceptibility to common diseases. *Nat. Genet.* **40,** 695–701 (2008).
57. Styrkarsdottir, U. *et al.* Nonsense mutation in the LGR4 gene is associated with several human diseases and other traits. *Nature* **497,** 517–520 (2013).
58. Tachmazidou, I. *et al.* Whole-Genome Sequencing Coupled to Imputation Discovers Genetic Signals for Anthropometric Traits. *Am. J. Hum. Genet.* **100,** 865–884 (2017).
59. Iotchkova, V. *et al.* Discovery and refinement of genetic loci associated with cardiometabolic risk using dense imputation maps. *Nat. Genet.* **48,** 1303–1312 (2016).
60. Dickson, S. P., Wang, K., Krantz, I., Hakonarson, H. & Goldstein, D. B. Rare Variants

Create Synthetic Genome-Wide Associations. *PLoS Biol.* **8,** e1000294 (2010).

61. Wray, N. R., Purcell, S. M., Visscher, P. M., Richardson, A. & Sisay-Joof, F. Synthetic Associations Created by Rare Variants Do Not Explain Most GWAS Results. *PLoS Biol.* **9,** e1000579 (2011).

62. Anderson, C. A., Soranzo, N., Zeggini, E., Barrett, J. C. & Lim, X. L. Synthetic associations are unlikely to account for many common disease genome-wide association signals. *PLoS Biol.* **9,** e1000580 (2011).

63. Wray, N. R. *et al.* Pitfalls of predicting complex traits from SNPs. *Nat. Rev. Genet.* **14,** 507–15 (2013).

64. Witte, J. S., Visscher, P. M. & Wray, N. R. The contribution of genetic variants to disease depends on the ruler. *Nat. Rev. Genet.* **15,** 765–776 (2014).

65. Chapman, J. M., Cooper, J. D., Todd, J. A. & Clayton, D. G. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum. Hered.* **56,** 18–31 (2003).

66. Spencer, C. C. A. *et al.* Designing Genome-Wide Association Studies: Sample Size, Power, Imputation, and the Choice of Genotyping Chip. *PLoS Genet.* **5,** e1000477 (2009).

67. Estrada, K. *et al.* Genome-wide meta-analysis identifies 56 bone mineral density loci and reveals 14 loci associated with risk of fracture. *Nat. Genet.* **44,** 491–501 (2012).

68. Collins, F. S. Cystic fibrosis: molecular biology and therapeutic implications. *Science* **256,** 774–9 (1992).

69. Yankaskas, J. R., Marshall, B. C., Sufian, B., Simon, R. H. & Rodman, D. Cystic fibrosis adult care: consensus conference report. *Chest* **125,** 1S–39S (2004).

70. Kemp, J. *et al.*. Identification of 153 new loci associated with heel bone mineral density and functional involvement of GPC6 in osteoporosis. *Nat. Genet.* **49**, 1468–1475 (2017). **[Au: Updated reference OK?]**

71. Istvan, E. S. & Deisenhofer, J. Structural mechanism for statin inhibition of HMG-CoA reductase. *Science* **292,** 1160–4 (2001).

72. Illingworth, D. R. *et al.* Comparative effects of lovastatin and niacin in primary hypercholesterolemia. A prospective trial. *Arch. Intern. Med.* **154,** 1586–95 (1994).

73. Richards, J. B., Zheng, H.-F. & Spector, T. D. Genetics of osteoporosis from genome-wide association studies: advances and challenges. *Nat. Rev. Genet.* **13,** 672–672 (2012).

74. Willer, C. J. *et al.* Newly identified loci that influence lipid concentrations and risk of

coronary artery disease. *Nat. Genet.* **40,** 161–9 (2008).

75. Sullivan, D. *et al.* Effect of a monoclonal antibody to PCSK9 on low-density lipoprotein cholesterol levels in statin-intolerant patients: the GAUSS randomized trial. *JAMA* **308,** 2497–506 (2012).

76. McClung, M. R. *et al.* Denosumab in postmenopausal women with low bone mineral density. *N. Engl. J. Med.* **354,** 821–831 (2006).

77. Jones, A. M. & Helm, J. M. Emerging treatments in cystic fibrosis. *Drugs* **69,** 1903–10 (2009).

78. Arrowsmith, J. Trial watch: phase III and submission failures: 2007-2010. *Nat. Rev. Drug Discov.* **10,** 87 (2011).

79. Smith, G. D. & Ebrahim, S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol* **32,** 1–22 (2003).

80. Burgess, S., Bowden, J., Fall, T., Ingelsson, E. & Thompson, S. G. Sensitivity Analyses for Robust Causal Inference from Mendelian Randomization Analyses with Multiple Genetic Variants. *Epidemiology* **28,** 30–42 (2017).

81. Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M. & Smoller, J. W. Pleiotropy in complex traits: challenges and strategies. *Nat. Rev. Genet.* **14,** 483–95 (2013).

82. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* 512–525 (2015). doi:10.1093/ije/dyv080

83. Holmes, M. V, Ala-korpela, M. & Smith, G. D. Mendelian randomization in cardiometabolic disease: challenges in evaluating causality. *Nat. Rev. Cardiol.* (2017). doi:10.1038/nrcardio.2017.78

84. Wood, A. R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **46,** 1173–86 (2014).

85. Leto, D. & Saltiel, A. R. Regulation of glucose transport by insulin: traffic control of GLUT4. *Nat. Rev. Mol. Cell Biol.* **13,** 383–96 (2012).

86. Dehghan, A. *et al.* Meta-analysis of genome-wide association studies in &gt;80 000 subjects identifies multiple loci for C-reactive protein levels. *Circulation* **123,** 731–8 (2011).

87. Merriman, T. R. An update on the genetic architecture of hyperuricemia and gout. *Arthritis Res. Ther.* **17,** 98 (2015).

88. Maller, J. *et al.* Common variation in three genes, including a noncoding variant in CFH,

strongly influences risk of age-related macular degeneration. *Nat. Genet.* **38,** 1055–1059 (2006).

89. Sasidhar, M. V., Reddy, S., Naik, A. & Naik, S. Genetics of coronary artery disease - A clinician's perspective. *Indian Heart Journal* **66,** 663–671 (2014).

90. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *Am. J. Hum. Genet.* **90,** 7–24 (2012).

91. Rietveld, C. A. *et al.* GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science* **340,** 1467–71 (2013).

92. Hodgkin, J. Seven types of pleiotropy. *International Journal of Developmental Biology* **42,** 501–505 (1998).

93. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47,** 291–295 (2015).

94. Hartl, D. L. & Clark, A. G. *Principles of population genetics*. (1997).

95. Fu, Y. X. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* **147,** 915–925 (1997).

96. Ramachandran, S. *et al.* Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc. Natl. Acad. Sci.* **102,** 15942–15947 (2005).

97. Lowe, J. J. K. *et al.* Genome-Wide Association Studies in an Isolated Founder Population from the Pacific Island of Kosrae. *PLoS Genet.* **5,** e1000365 (2009).

98. Panoutsopoulou, K. *et al.* Genetic characterization of Greek population isolates reveals strong genetic drift at missense and trait-associated variants. *Nat. Commun.* **5,** 5345 (2014).

99. Dowell, S. F. Seasonal Variation in Host Susceptibility and Cycles of Certain Infectious Diseases. *Emerg. Infect. Dis.* **7,** 369–374 (2001).

100. Cooke, G. S. & Hill, A. V. S. Genetics of susceptibitlity to human infectious disease. *Nat. Rev. Genet.* **2,** 967–977 (2001).

101. Kirschner, M. & Gerhart, J. Evolvability. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 8420–8427 (1998).

102. Martin, M. P. & Carrington, M. Immunogenetics of viral infections. *Current Opinion in Immunology* **17,** 510–516 (2005).

103. Jeffreys, a J., Kauppi, L. & Neumann, R. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat. Genet.* **29,** 217–222 (2001).

104. Flint, J., Harding, R. M., Boyce, A. J. & Clegg, J. B. The population genetics of the

haemoglobinopathies. *Baillieres. Clin. Haematol.* **11,** 1–51 (1998).

105. Carter, A. J. & Nguyen, A. Q. Antagonistic pleiotropy as a widespread mechanism for the maintenance of polymorphic disease alleles. *BMC Med. Genet.* **12,** 160 (2011).

106. Laland, K. N., Odling-Smee, J. & Myles, S. How culture shaped the human genome: bringing genetics and the human sciences together. *Nat. Rev. Genet.* **11,** 137–148 (2010).

107. Sabeti, P. C. *et al.* Positive natural selection in the human lineage. *Science* **312,** 1614–20 (2006).

108. Gerbault, P. *et al.* Evolution of lactase persistence: an example of human niche construction. *Philos. Trans. R. Soc. B Biol. Sci.* **366,** 863–877 (2011).

109. Charlesworth, D. Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genet.* **2,** 379–384 (2006).

110. Hamblin, M. T. & Di Rienzo, A. Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. *Am. J. Hum. Genet.* **66,** 1669–79 (2000).

111. Currat, M. *et al.* Molecular analysis of the beta-globin gene cluster in the Niokholo Mandenka population reveals a recent origin of the beta(S) Senegal mutation. *Am. J. Hum. Genet.* **70,** 207–223 (2002).

112. Field, Y. *et al.* Detection of human adaptation during the past 2000 years. *Science* **354,** 760–764 (2016).

113. Berg, J. J. & Coop, G. A Population Genetic Signal of Polygenic Adaptation. *PLoS Genet.* **10,** (2014).

114. Kong, A. *et al.* Selection against variants in the genome associated with educational attainment. *Proc. Natl. Acad. Sci.* **114,** E727–E732 (2017).

115. Turchin, M. C. *et al.* Evidence of widespread selection on standing variation in Europe at height-associated SNPs. *Nat. Genet.* **44,** 1015–9 (2012).

116. Robinson, M. R. *et al.* Population genetic differentiation of height and body mass index across Europe. *Nat. Genet.* **47,** 1357–1362 (2015).

117. Gibson, G. Decanalization and the origin of complex disease. *Nat. Rev.* **10,** 134–140 (2009).

118. Flatt, T. The Evolutionary Genetics of Canalization. *Q. Rev. Biol.* **80,** 287–316 (2005).

119. Marouli, E. *et al.* Rare and low-frequency coding variants alter human adult height. *Nature* (2017). doi:10.1038/nature21039

120. Manousaki, D. *et al.* Toward precision medicine: TBC1D4 disruption is common among the inuit and leads to underdiagnosis of type 2 diabetes. *Diabetes Care* **39,** 1889–1895

(2016).

121. Moltke, I. *et al.* A common Greenlandic TBC1D4 variant confers muscle insulin resistance and type 2 diabetes. *Nature* **512,** 190–193 (2014).

122. Smith, M. W. *et al.* A High-Density Admixture Map for Disease Gene Discovery in African Americans. *Am. J. Hum. Genet.* **74,** 1001–1013 (2004).

123. Clayton, D. G. Prediction and interaction in complex disease genetics: Experience in type 1 diabetes. *PLoS Genet.* **5,** 1–6 (2009).

124. Robinson, M. R. *et al.* Genotype-covariate interaction effects and the heritability of adult body mass index. *Nat. Publ. Gr.* **49,** 1174–1181 (2017).

125. Curb, J. D. & Marcus, E. B. Body fat and obesity in Japanese Americans. *Am. J. Clin. Nutr.* **53,** 1552S–1555S (1991).

126. Delavari, M., Sønderlund, A. L., Swinburn, B., Mellor, D. & Renzaho, A. Acculturation and obesity among migrant populations in high income countries – a systematic review. *BMC Public Health* **13,** 458 (2013).

127. Murphy, M., Robertson, W. & Oyebode, O. Obesity in International Migrant Populations. *Curr. Obes. Rep.* (2017). doi:10.1007/s13679-017-0274-7

128. Patel, J. V *et al.* Impact of migration on coronary heart disease risk factors: Comparison of Gujaratis in Britain and their contemporaries in villages of origin in India. *Atherosclerosis* **185,** 297–306 (2006).

129. Ko, Y., Butcher, R. & Leong, R. W. Epidemiological studies of migration and environmental risk factors in the inflammatory bowel diseases. *World J. Gastroenterol.* **20,** 1238–47 (2014).

130. Pareek, M., Greenaway, C., Noori, T., Munoz, J. & Zenner, D. The impact of migration on tuberculosis epidemiology and control in high-income countries: a review. *BMC Med.* **14,** 48 (2016).

131. Ziegler, R. G. *et al.* Migration patterns and breast cancer risk in Asian-American women. *J. Natl. Cancer Inst.* **85,** 1819–1827 (1993).

132. Le, G. M., Gomez, S. L., Clarke, C. A., Glaser, S. L. & West, D. W. Cancer incidence patterns among Vietnamese in the United States and Ha Noi, Vietnam. *Int. J. Cancer* **102,** 412–417 (2002).

133. Bearn, A. G. & Miller, E. D. Archibald Garrod and the development of the concept of inborn errors of metabolism. *Bull Hist Med* **53,** 315–28 ST–Archibald Garrod and the development (1979).

134. Xu, C. *et al.* Estimating genome-wide significance for whole-genome sequencing studies.

Genet. Epidemiol. **38,** 281–290 (2014).

135. Moutsianas, L. *et al.* The power of gene-based rare variant methods to detect disease-associated variation and test hypotheses about complex disease. *PLoS Genet.* **11,** e1005165 (2015).

136. BANSAL, V., LIBIGER, O., TORKAMANI, A. & SCHORK, N. J. in *Biocomputing 2011* 76–87 (WORLD SCIENTIFIC, 2010). doi:10.1142/9789814335058_0009

137. Bansal, V., Libiger, O., Torkamani, A. & Schork, N. J. Statistical analysis strategies for association studies involving rare variants. *Nat. Rev. Genet.* **11,** 773–785 (2010).

138. Basu, S. & Pan, W. Comparison of statistical tests for disease association with rare variants. *Genet. Epidemiol.* **35,** 606–19 (2011).

139. Styrkarsdottir, U. *et al.* Severe osteoarthritis of the hand associates with common variants within the ALDH1A2 gene and with rare variants at 1p31. *Nat. Publ. Gr.* **46,** 498–502 (2014).

140. Do, R. *et al.* Exome sequencing identifies rare LDLR and APOA5 alleles conferring risk for myocardial infarction. *Nature* **518,** 102–6 (2015).

141. Ladouceur, M., Zheng, H.-F., Greenwood, C. M. T. & Richards, J. B. Empirical power of very rare variants for common traits and disease: results from sanger sequencing 1998 individuals. *Eur. J. Hum. Genet.* **21,** 1027–1030 (2013). **[Au: Page numbers OK?]**

142. Mancuso, N. *et al.* Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits. *Am. J. Hum. Genet.* **100,** 473–487 (2017).

143. The Emerging Risk Factors Collaboration. Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. *Lancet* **375,** 2215–2222 (2010).

144. Krogh, A. & Krogh, M. A study of the diet and metabolism of Eskimos undertaken in 1908 on an expedition to Greenland. *Meddelelser om Gronl.* **41,** 165–173 (1914).

145. Mouratoff, G. J., Carroll, N. V & Scott, E. M. Diabetes mellitus in Eskimos. *JAMA* **199,** 107–112 (1967).

146. Jorgensen, M. E. *et al.* Diabetes and impaired glucose tolerance among the inuit population of greenland. *Diabetes Care* **25,** 1766–1771 (2002).

147. Sladek, R. *et al.* A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445,** 881–5 (2007).

148. Scott, R. a. *et al.* Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. **44,** 991-1005 (2012).

[Au: Page numbers OK?]

149. Bradfield, J. P. *et al.* A genome-wide meta-analysis of six type 1 diabetes cohorts identifies multiple associated loci. *PLoS Genet.* **7,** e1002293 (2011). **[Au: Page numbers OK?]**

150. Li, J. Z. *et al.* Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation. *Science (80-.).* **319,** 1100–1104 (2008).

151. Rosenberg, N. A. *et al.* Genetic structure of human populations. *Science* **298,** 2381–5 (2002).

Acknowledgements: The authors wish to acknowledge Vincenzo Forgetta for his help drawing the figures. NJT is a Wellcome Trust Investigator (202802/Z/16/Z), a programme lead in the MRC Integrative Epidemiology Unit (MC_UU_12013/3) and works within the University of Bristol NIHR Biomedical Research Centre (BRC). CG has received funding from NSERC. DJL is funded by the Wellcome Trust under grant number WT104125MA. JBR receives support from the Canadian Institutes of Health Research, The Lady Davis Institute of the Jewish General Hospital and the Fonds du Recherche Santé Québec.

Author Contributions

N. J. T., C. M. T. G., D. J. L and J. B. R. researched data for article, contributed to discussion of the content, wrote the article and reviewed and/or edited the manuscript before submission. N.S. contributed to discussion of the content and reviewed and/or edited the manuscript before submission.

Competing interests statement

The authors declare no competing interests.

Publisher's note

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

Databases [CE/PE: Please add hyperlinks to the databases where they first appear in the main text]

UKBiobank
http://www.ukbiobank.ac.uk/

TopMed
https://www.nhlbi.nih.gov/research/resources/nhlbi-precision-medicine-initiative/topmed

GEFOS Consortium
http://www.gefos.org

UK10K
https://www.uk10k.org/

GoT2D
http://www.type2diabetesgenetics.org/projects/got2d

T2D-GENES
http://www.type2diabetesgenetics.org/projects/t2dGenes

GIANT-Consortium
http://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium

deCODE
https://www.decode.com/

Box 1: What is genetic architecture?

Genetic architecture refers to the landscape of genetic contributions to a given phenotype. It comprises the number of genetic variants that influence a phenotype, the size of their effects on the phenotype, the frequency of those variants in the population and their interactions with each other and the environment.[2] This is fundamentally different to the absolute amount of phenotypic variability that is accounted for by heritable factors (**Figure 1**). We can illuminate this concept by comparing two extremely different heritable phenotypes: height and phenylketonuria. Height has a polygenic, or even omnigenic, architecture,[20] the latter of which is similar in concept to an infinitesimal architecture.[21] Height must be sufficiently distal from the genome and inclusive of many biological processes and causal genetic variants to have such a polygenic architecture. By contrast, phenylketonuria has a monogenic architecture: although heritability is high, the shape of this heritability is singular. In phenylketonuria, one 'inborn error'[133] is responsible for the heritable phenotypic variability and thus the trait measured must be proximal to that genetic change to guard it from other potential contributions. These two contrasting examples of genetic architecture differ in the tools needed to discover and describe them and in how they can be used in a research or clinical setting. Traits like height may reflect the existence of many, common, ancient and small contributions to a complex phenotype, which require large population based collections and genome-wide common variant data to detect and which may have use in studies of risk factor exposure through techniques such as Mendelian randomization. In contrast, phenylketonuria may reflect relatively recent and thus rare mutations that have avoided the rigor of time and selection and which require huge samples of sequence data or familial designs to detect, but they may also have immediate clinical or pharmaceutical implications. Importantly, the architecture of these traits cannot be reliably predicted by the assessment of heritability alone. We aim to explore genetic architecture here given lessons from both the genome-wide association study and next generation sequencing eras. Our aim is to highlight that there is likely to be great variability in the genetic architecture of given traits of interest and that this should be considered for three reasons. First, architecture should be a motivating factor for comprehensive genetic studies of many phenotypes with unlimited size. Second, study designs should be tailored to observed genetic architecture and finally, understanding architecture and its limitations directly informs the clinical goals of human genetics, which are to assist in diagnosis, prognosis and the identification of therapeutic targets.

Box 2: Limitations of region-based single nucleotide variant testing.

Region-based testing, which is motivated by a desire to improve statistical power to detect rare SNVs, tests the association of a trait with genetic variation across a genomic region, rather than the association of a trait with specific genetic variant. Region-based have important limitations, including difficulty with replication since, in a region of interest, the number of genetic variants observed and their allele frequencies can differ substantially across cohorts, even in populations of similar ancestry. Use of these tests also presents other practical problems. First, region-based tests should be optimized for a specific genetic architecture, but the complete genetic architecture of a trait or disease is not often known and will vary between populations, and the genetic loci will also exhibit differences in local allelic architecture for a given trait. Second, most of the genetic variants identified through whole-genome sequencing studies (WGS) may have no discernible effect on the selected phenotype, and the inclusion of large numbers of variants with no effect in a region-based test will reduce power.[134] Third, the direction of effect of rare variants (that is, whether they increase or decrease the risk of the disease) is not always known and this reduces the power of some region-based tests. Last, the relative performance of different tests can vary across significance thresholds; hence, if only a small number of candidate loci are being examined, the optimal test statistic is likely to be different from a genome-wide analysis.[135] Previous work has outlined additional assumptions built into region-based testing.[136–138]

These challenges were apparent in the UK10K cohort project as neither a burden test **[G]** nor a variance component test **[G]** could identify a single instance, across 60 traits, where a region-based test could identify a region not already highlighted through single SNV testing[42]. This region-based testing included several strategies to combine variants across a genomic region: it combined variants <1 and <5% MAF separately and included only protein-coding regions and only regions with evidence of evolutionary conservation. These tests may have yielded null results because UK10K used low-coverage sequencing and imputation, which captured rare variants with high fidelity, but as it had lower sensitivity for singletons and doubletons **[G]** it could have missed contributions from these SNVs.[42] Findings from the UK10K cohort project are also limited by the bounds of statistical power, given the study sample size. Nonetheless, region-based testing did not contribute to our understanding of genetic architecture in this study.

Other large sequencing-based studies have employed region-based tests with limited success. The GoT2D and T2D-GENES consortia undertook WGS in 2,657 individuals of European descent with and without type 2 diabetes mellitus, and whole-exome sequencing (WES) in 12,940 individuals.[27] No rare variants or regions were found to be associated with type 2 diabetes mellitus in this programme. A recent assessment by the GIANT-Consortium of coding genetic variants associations with height in 711,428 individuals identified rare variants at 83 loci associated with height using single SNV testing, but only three novel regions were identified through region-based testing.[119] Similarly, although WGS-based studies from deCODE have identified single-SNV associations, clear associations from region-based tests were not identified,[139] and a WES program in 9,983 patients with early-onset myocardial infarction, identified only single-SNV associations.[140] This is consistent with previous work demonstrating that the empirically observed power of region-based tests is low.[141] A promising avenue of region-based testing is transcriptome-wide association testing,[142] although over 80% of the region-based findings using this method were identified using simple single SNV association testing.

It is not apparent how region-based testing could assess the presence of an omnigenic architecture, although a region-based test revealed an enrichment of association signal in different types of variants, stratified by presumed functional effect, which showed a stronger signal from SNPs residing in active chromatin.[20]

The success of region-based testing methods may increase as larger studies can capture and annotate very rare variants in large groups of individuals. However, in our experience the most profitable strategy for finding low frequency or rare genetic variants with previously undiscovered contributions to genetic architecture is currently the use of single SNV association tests in large cohorts.[119]

Box 3: Decanalization can identify unusual genetic architectures.

Decanalization occurs when there is a large environmental change that influences a biological system that is strongly canalized. Circulating glucose levels are strongly canalized and show little variation in a healthy state.[143] However, a large environmental change in Inuit may have led to decanalization of glucose control, which subsequently provided an opportunity to identify an unusual genetic architecture for type 2 diabetes mellitus; in the Inuit, this genetic architecture includes a common variant (minor allele frequency (MAF) 0.17) that has a large effect on the risk of developing type 2 diabetes mellutis.[121] Living in a reduced-carbohydrate environment, Inuit had a relatively low intake of carbohydrates prior to the introduction of Western diets (see the figure, part a). Recent estimates demonstrate a much higher proportion of carbohydrate intake in Inuit contemporary diets.[144] This large environmental change may have resulted in a decanalization of glucose regulation, which may have contributed to a dramatic increase in the prevalence of type 2 diabetes mellitus amongst Inuit between 1967 and 2002 (see the figure, part b).[145,146] A recent metabochip genome-wide association study for glucose levels 2 hours after the intake of glucose in an oral glucose tolerance test found that a common premature termination codon in *TBC1D4*, the gene encoding TBC1 domain family member 4, had a large effect on this phenotype in Inuit (see the figure, part c, which demonstrates the strong association signal with glucose levels after an oral glucose tolerance test arising on chromosome 13). This information led to different diagnostic strategies in this population, which aim to use oral glucose tolerance testing to more accurately diagnose type 2 diabetes mellitus in this population.[121] By contrast, there are no common genetic variants of similarly large effect for type 2 diabetes mellitus in the European population.[147] The unusually high effect-size common variant for glucose levels 2 hours after oral glucose identified in Inuit (in red; see the figure, part d) contrasts with the small effect size common variants identified for this phenotype in a European-ancestry population (in blue; see the figure, part d).[121,148] These data suggest that decanalization can lead to unusual genetic architectures, particularly in historically isolated populations, such as Inuit. Graphs in parts a, b and d were generated using data published in references xxx, xxx and xxx, respectively. **Part c was reproduced, with permission, from reference 121.**

Figure 1. Contrasting the observed genetic architecture of common diseases and biomedical traits. a) Genome-wide significant single nucleotide variants (SNVs) for type 1 diabetes mellitus and type 2 diabetes mellitus are shown. Large genome-wide association studies (GWAS) show that these common diseases have contrasting observed genetic architectures. Type 1 diabetes mellitus is associated with common and low frequency genetic variants, some of which have relatively large effects on genetic architecture; these effects are measured in odds ratios, which give the odds of the outcome given exposure to one risk allele, compared to the odds of the outcome given exposure to no risk alleles.[149] The genetic architecture of type 2 diabetes mellitus is shaped by, what are in general, smaller effect size common variants (which have a higher minor allele frequency (MAF)). **[Au: higher than what? MAFs of larger effect size common variants?]** The different architectures for diabetes type 1 mellitus and type 2 diabetes mellitus impact the development of diagnostic tests and biologic validation of therapeutic targets for these diseases.[25] b) Genome-wide significant SNVs for the biochemical traits Vitamin D (25OHD) and LDL cholesterol. Vitamin D, as measured by 25-hydroxyvitamin D (25OHD) is associated with few genetic variants, some of which have relatively large effects.[32] Only two new loci have been identified as being associated with 25OHD, despite increasing the discovery sample size five-fold.[33] Low density lipoprotein (LDL) cholesterol is associated with many more loci than 25OHD and these loci have a broader distribution of effect sizes than those associated with 25OHD.[34] Beta is the additive effect of the minor alleles on the phenotype in standard deviations. Graphs in parts a and b were generated using data pubished in references xxx and xxx, respectively.

Figure 2. Allelic spectrum for single marker association results for selected traits. A variant's effect (absolute value of Beta, expressed in standard deviation units) is given as a function of minor allele frequency (MAF). The effect of each variant was assessed using single variant association tests, providing the effect of each variant, in standard deviation units, on the trait. Note that N is sample size and α is the multiple-testing corrected threshold to declare significance. Error bars are proportional to the standard error of the beta. Variants identifying known loci are shown in dark blue and variants identifying novel signals that have been replicated in independent studies are shown in light blue. The red and orange lines indicate 80% power at experiment wide significance level (p value ≤ 4.62×10^{-10}) for the maximum theoretical sample size for the whole genome sequencing sample (red) and whole genome sequencing and genome-wide genotyping samples (orange) in the UK10K project.[42] The observed deficit of

large-effect size rare variants will likely be overcome through larger sample sizes, as already observed for traits like bone mineral density.[70] The main messages of this graph are that effect sizes increase with decreasing MAF, and that identified variants are not dramatically above what would be expected, given the power of the study. **Figure reproduced, with permission, from reference 42**

Figure 3. Lack of unbalanced horizontal pleiotropy between bone mineral density and height using Mendelian randomization. The Mendelian randomization (MR)-Egger[82] plot tests for the presence of unbalanced horizontal pleiotropy, which would violate a core assumption of Mendelian randomization. The two traits studied by genome-wide association studies (GWAS) are height[84] and bone mineral density[70]. Both are highly polygenic and dependent upon the same tissue, bone. The omnigenic hypothesis suggests that widespread network pleiotropy would violate the pleiotropy assumption of Mendelian randomization if both the exposure and the outcome are complex traits dependent on the same tissue. To test whether omnigenic pleiotropy violates Mendelian randomization assumptions, we assessed the evidence for horizontal pleiotropy between bone mineral density and adult height, two polygenic phenotypes that are influenced by bone tissue and that have been subjected to large-scale GWAS. Using the largest published **[Au:OK?]** GWAS for bone mineral density (n = 142,487, using 169 biallelic conditionally independent genome-wide significant independent SNVs)[70] and adult height (n = 253,288),[84] we tested for horizontal pleiotropy using MR-Egger, treating bone mineral density as the exposure and height as the outcome.[82] However, this MR-Egger plot shows that there is no evidence of unbalanced horizontal pleiotropy, strongly suggesting a lack of network unbalanced horizontal pleiotropy that would violate Mendelian randomization assumptions. (The MR-Egger intercept is not different from zero: -0.002, 95% confidence intervals: -0.005, 0.001). For contrast, the inverse variance weighted results are shown, which constrain the line to intersect with the origin. GEFOS, genetic factors for osteoporosis; GIANT, genetic investigation of anthropometric traits.

Figure 4. Hypothetical departures from the ‘dose-response’ curve: Diagrams show the expected relationships between minor allele frequency (MAF) and the effect that variants may have on complex traits; this relationship defines the variation in genetic architecture between populations. **a**) Common single nucleotide polymorphisms (SNPs) are not expected to show strong effects on a phenotype as they would likely be deleterious and selected against to become rare. Therefore, a characteristic ‘dose-response’ curve, above which there are no variants, is expected. The shape of this curve can be determined by the effective population size and the number of samples in the dataset. A lower effective population size reduces the efficacy of selection, allowing greater variation in MAF. The points on the curve represent SNPs; the arrows show how they might move in a smaller population. **b)** Genomic architecture may differ by population. Here, population A experienced strong negative selection for the disease, reducing its incidence. Populations B and C retained the same mean trait but changed their genetic architecture by, for example, drift or pleiotropy. The selective origins of these differences may be inferred using historical allele frequencies. **c)** Most populations, for most traits, have the same effect size. However, some populations (shown here as X) may experience a higher measured effect size as a result of decanalization due to environmental pressure or because a small population size creates drift in the genetic structure that regulates the trait of interest. The different colours represent alternative states in each of the scenarios.

Figure 5: Difference between GWAS height loci across populations. a) A representation of allele frequency difference for height associated single nucleotide polymorphisms (SNPs) between North Europe and South Europe and African samples by their rank of effect size compared to the expected. This is a summary of previous work[115] for **[Au:OK?]** sets of 500 independent ($r^2 < 0.1$) SNPs across the genome, sorted by GIANT height-association *P* value. Differences in population based minor allele frequency for many loci presents the potential for polygenic selection, which is shaping the genetic architecture of height. b) The relationship between genetic scores derived from the sum of sample allele frequencies, weighted by minor allele frequency for height based on existing GWAS data and composed in populations from the Human Gene Diversity Panel.[150] Solid bars represent the actual genetic score for height calculated in each population in comparison to that predicted under a neutral model (with no marked population specific differences) and based on related populations (dashed bars).[113] Coloured areas represent the spread of sub-population specific estimates for genetic score nested within established population groupings.[151] For an exemplar polygenic trait, these differences in genetic score illustrate potential evidence for polygenetic selection/adaptation. **Figure in part a adapted, with permission, from reference 115. Figure in part b adapted, with permission, from reference 113.**

Glossary:

Heritable: A characteristic or trait that has a portion of variability that is accounted for by genetic factors.

Phenotype: A measurable characteristic of an individual.

Broad sense phenotypic **heritability:** The proportion of trait variance that is due to all genetic factors, including dominant and recessive factors as well as the interactions between genetic factors. Narrow sense heritability is the proportion of trait variance that is due to additive genetic factors.

Complex traits: A trait that does not follow Mendelian inheritance patterns and is derived from any combination of multiple genetic factors, environmental factors and their interactions.

Minor allele frequency: The frequency of the less frequent allele at a genetic variant in a population. The less frequent allele is referred to as the minor allele.

Deep imputation: The use of large imputation reference panels to accurately estimate most low-frequency ($1\% \leq \text{MAF} \leq 5\%$) and rare ($\text{MAF} < 1\%$) unobserved genetic variation in individuals who have been genome-wide genotyped.

Phenotypic variance: The variance in a phenotype, which is often assumed to be a function of environmental and genetic factors, as well as their interactions.

Single nucleotide polymorphisms (SNPs): Single base pair positions in the genome where two or more nucleotides occur commonly in the population. 'Common' is usually defined by at least 1% of the population carrying an alternative allele. Most often SNPs are biallelic, which means that the nucleotide will be one of two different alleles.

Single nucleotide variants (SNV): Single base pair positions in the genome where there is variation across individuals. SNVs need not be biallelic or common.

Genome-wide association studies (GWAS): Studies that test the association of all measured genetic variation across the genome with a trait or disease. GWAS usually tests the association of a phenotype with genetic variants that have a minor allele frequency (MAF) ≥1%, but deep imputation methods allow GWAS to test associations with variants at a lower MAF.

Whole-exome sequencing (WES) studies: Studies that tests the association between genetic variation (usually SNVs) across the measured coding sequence of the genome with a trait or disease. WES can measure most coding genetic variants, regardless of minor allele frequency.

Whole-genome sequencing (WGS) studies: Studies that test the association of genetic variation across the entire variable genetic sequence of the genome with a trait or disease.

WGS can measure most genetic variants present in the genome, regardless of minor allele frequency. However, certain regions are not usually measurable via sequencing, such as highly repetitive regions.

Imputation reference panel: A dataset containing genetic information on a large number of individuals who have been whole-genome sequenced and had their haplotypes reconstructed. These haplotype panels enable accurate imputation of non-genotyped genetic variants in individuals who have undergone genome-wide genotyping.

Haplotypes: A section of commonly varying or linked chromosomal material said to be in gametic phase, i.e. not punctuated by recombination at an appreciable population based frequency.

Region-based testing: A single test of association between many genetic variants in a chosen region of the genome and a phenotype.

Burden test: A class of region-based testing that collapses genetic variation into a single genetic score by measuring the total number of minor alleles across a genomic region.

Variance component test: A single test of whether the phenotypic variance explained by a set of chosen genetic variants across a genomic region is zero. For example, a variance component test could be used to test whether all single nucleotide variants in a gene contribute to the variability in a phenotype.

Single SNV association test: A genetic association test that tests variation at a single nucleotide variant with variation in a phenotype. This is the most common genetic association test and is frequently used for genome-wide genotyping data.

Variance explained: The proportion of variance in a phenotype that is explained by a mathematical model.

Linkage disequilibrium: The non-random association of alleles in a population.

Receiver operator curve: A method to evaluate the performance of a diagnostic test for a binary outcome that plots the test’s sensitivity (the true positive rate) against one minus the test’s specificity (the false positive rate).

Confounding: When the association between an exposure and an outcome is distorted by their associations with a third variable. A confounding variable is a variable that is associated with both the exposure and the outcome, but is not in the causal pathway between the two. A confounding variable could include a common cause of both the exposure and the outcome.

Reverse causation: The phenomenon whereby the outcome influences the exposure.

Horizontal pleiotropy: In a Mendelian randomization study, horizontal pleiotropy is when the genetic variant influences the outcome in a manner independent of the risk factor. This is a violation of Mendelian randomization assumptions.

Vertical pleiotropy: In a Mendelian randomization study, vertical pleiotropy is when the genetic variant influences the outcome through multiple biomarkers in the same pathway. This is not a violation of Mendelian randomization assumptions.

Founder effect: Reduced genetic diversity that results when a population is descended from a small number of founders.

Singleton: Genetic variant that is observed only once within the population studied.

Doubletons: Genetic variants that are observed twice within the population studied.

Lactase persistence: The continued activity of the enzyme lactase in adulthood in humans.

Admixture mapping: A method of genetic association testing that relies on the admixture of populations, which occurs when individuals from two or more historically isolated populations interbreed.

ONLINE ONLY

Author biogs

Nicholas J. Timpson is a Wellcome Trust Investigator and programme lead at the MRC Integrative Epidemiology Unit at the University of Bristol. He co-leads work in the CRUK Integrative Cancer Epidemiology Programme and the NIHR Biomedical Research Centre, Bristol. His research focuses on genetic association and genetic epidemiology analysis for complex traits.

Celia M.T. Greenwood is a Senior Investigator at the Lady Davis Research Institute and Professor at McGill University. Her research involves development of statistical methodology for analysis of genetic and genomic data, and she has contributed to many articles discussing analytic strategies for DNA sequencing studies.

Nicole Soranzo is a Senior Group Leader at the Wellcome Trust Sanger Institute and professor of Human Genetics at the University of Cambridge. Her research focuses on the study of genetic influences on complex human traits and diseases.

Daniel J. Lawson is a Sir Henry Dale Wellcome Trust and Royal Society Research Fellow at the MRC Integrative Epidemiology Unit at the University of Bristol. He develops methodology to understand the relationships between population structure, demography and population health.

J. Brent Richards is a Senior Investigator, William Dawson Scholar and Endocrinologist at the Lady Davis Research Institute and Associate Professor at McGill University. His research focuses on identifying the genetic determinants of common endocrine disease and translating this information to improved patient care.

Figure permissions statement/s

Box 3 Part c was reproduced, with permission, from reference 121.
Figure 2 was reproduced, with permission, from reference 42.
Figure 5 part a was adapted, with permission, from reference 115.
Figure 5 part b was adapted, with permission, from reference 113.

Key points

Genetic architecture of common diseases is central to the scientific and clinical goals of human genetics, because it directly impacts biology, disease screening diagnosis, prognosis and treatment.

Genetic architecture is currently assessed by exploiting the differences in types of genetic variants measured through GWAS, WES and WGS. Each of these has its own merits and disadvantages, but all are subject to the limitations of sample size. Gene mapping studies should thus be tailored to the unique contributions of each of these technologies.

To date, the observed genetic architecture of highly heritable diseases and traits differs markedly and cannot be reliably predicted. Where large sample sizes are available there still exist differences in detectable architecture.

The concept of variance explained is not always relevant to individual-level risk prediction or drug development, whereas the genetic architecture of a given trait or disease can be more pertinent.

Genetic architecture is variable in time and place and can be theoretically influenced by phenotypic measurement, selection and decanalization.

Interactions between genetic determinants of a trait or environmental influences contribute to genetic architecture. To date, few such interactions have been identified for most common diseases and traits, but this will likely change with increasing sample sizes.

Subject categories

Biological sciences / Genetics / Population genetics / Genetic variation [URI /631/208/457/649]

Biological sciences / Genetics / Clinical genetics / Disease genetics [URI /631/208/2489/144]

Biological sciences / Genetics / Heritable quantitative trait [URI /631/208/729]

Biological sciences / Genetics / Genetic association study / Genome-wide association studies [URI /631/208/205/2138]

Biological sciences / Genetics / Sequencing / Next-generation sequencing [URI /631/208/514/2254]

Biological sciences / Genetics / Genomics / Medical genomics [URI /631/208/212/2301]

Health sciences / Health care / Public health / Epidemiology [URI /692/700/478/174]

ToC blurb

Genetic architecture: The shape of genetic contribution to human traits and disease

Nicholas Timpson, Celia M. T. Greenwood, Nicole Soranzo, Daniel J. Lawson, & J. Brent Richards

Genetic architecture describes the characteristics of genetic variation that are responsible for phenotypic variability. This Review discusses the types of genetic architecture that have been observed, how they can be measured, and how genetic architecture informs the scientific and clinical goals of human genetics.