Cell
Leading Edge

Review

Genetic and molecular architecture of complex traits

Tuuli Lappalainen,[1,2,*] Yang I. Li,[3,4] Sohini Ramachandran,[5] and Alexander Gusev[6]
[1]New York Genome Center, New York, NY, USA
[2]Science for Life Laboratory, Department of Gene Technology, KTH Royal Institute of Technology, Stockholm, Sweden
[3]Section of Genetic Medicine, University of Chicago, Chicago, IL, USA
[4]Department of Human Genetics, University of Chicago, Chicago, IL, USA
[5]Ecology, Evolution and Organismal Biology, Center for Computational Molecular Biology, and the Data Science Institute, Brown University, Providence, RI 029129, USA
[6]Harvard Medical School and Dana-Farber Cancer Institute, Boston, MA, USA
*Correspondence: tlappalainen@nygenome.org
https://doi.org/10.1016/j.cell.2024.01.023

SUMMARY

Human genetics has emerged as one of the most dynamic areas of biology, with a broadening societal impact. In this review, we discuss recent achievements, ongoing efforts, and future challenges in the field. Advances in technology, statistical methods, and the growing scale of research efforts have all provided many insights into the processes that have given rise to the current patterns of genetic variation. Vast maps of genetic associations with human traits and diseases have allowed characterization of their genetic architecture. Finally, studies of molecular and cellular effects of genetic variants have provided insights into biological processes underlying disease. Many outstanding questions remain, but the field is well poised for groundbreaking discoveries as it increases the use of genetic data to understand both the history of our species and its applications to improve human health.

INTRODUCTION

In the publication describing the initial draft of the human genome,[1,2] the progress during the 20th century in understanding the structure and content of genetic information was divided into four phases. Each of them spanned about a quarter of the century: the discovery of chromosomes; defining the molecular structure of DNA; the discovery of the molecular machinery of gene function; and finally determining the sequence of entire genes, scaffolds, and genomes. These achievements propelled the entire field of genetics into the genomic era in the early 21st century.

As the first quarter of this century soon draws to a close, we can reflect on the crowning achievements of genomics during this period: the characterization of genetic variation in human populations and the discovery of its contribution to phenotypic variation. Since the publication of the draft sequence of the human genome, human genetics has experienced dramatic growth in both the diversity and quality of genetic data, alongside an understanding of how genetic variation is linked to a wide variety of phenotypes (Figure 1A). These developments have demonstrated the fundamental role that genetics has in characterizing human biology, ranging from molecular to physiological levels, as well as the evolutionary history of our species and the evolution of complex traits, as we discuss in this review.

These investments have been also motivated by the potential of human genetic research to enhance human health. This can unfold through two synergistic routes. Accurate prediction of genetic effects on disease risk can improve diagnosis, prognosis, and treatment selection. Although genomic medicine has already had a transformative clinical impact in rare disease,[3,4] analogous applications in complex diseases are only now emerging from polygenic risk scores, as discussed further below[5] (Figure 1B). Beyond prediction, human genetics also empowers the development of new drugs and interventions via identification of causal genes and molecular mechanisms involved in disease[6] (Figure 1C). This paradigm is well supported by the higher success rates for drug targets backed by genetic evidence.[7,8] This requires characterization of functional mechanisms of genetic disease associations, which remains a considerable challenge, with current insights and ways forward discussed further below. A foundation of these goals of genetic prediction and mechanistic understanding is population genetics, which describes the processes that have given rise to and maintain human genetic variation.

ORIGINS AND CONTEMPORARY PATTERNS OF GENETIC VARIATION IN HUMAN POPULATIONS

Population genetics, which originated just over a century ago alongside modern statistics, is the study of the origin and evolution of genetic variation within groups of individuals. Classically, "population" is used in the biological sense of groups of randomly mating individuals. When applied to our own species,

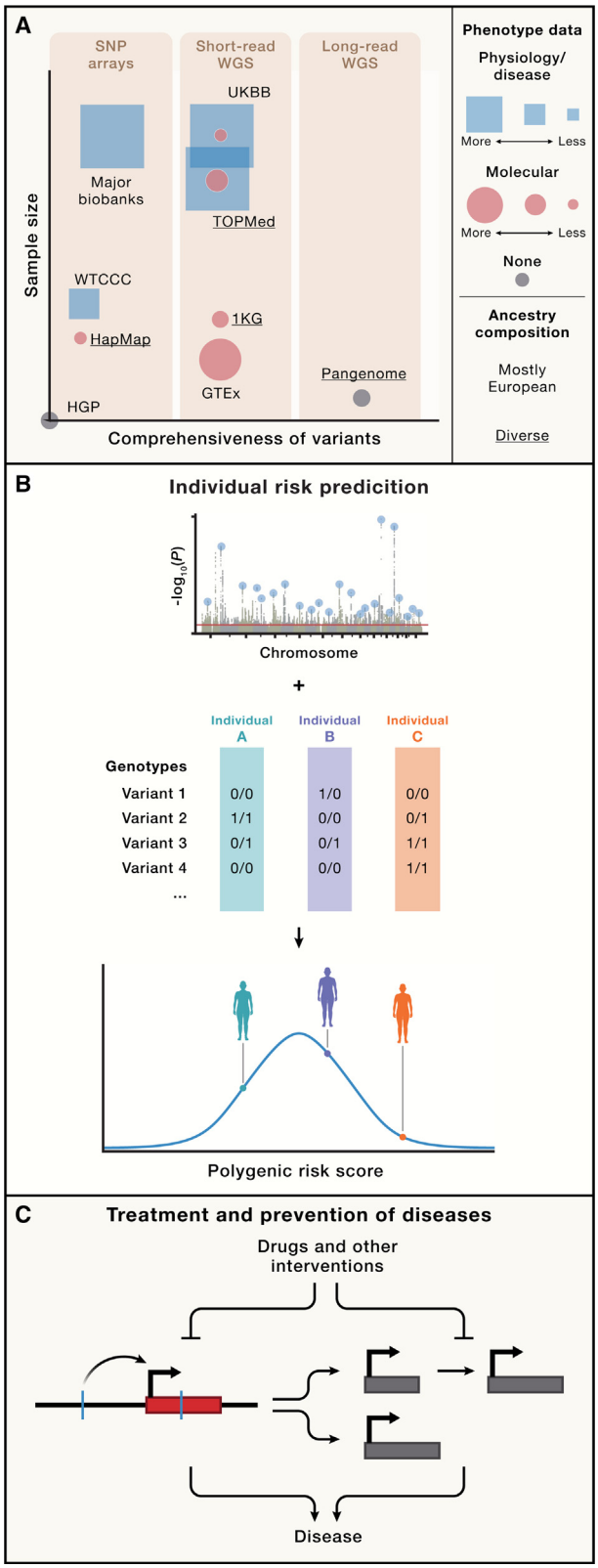

Figure 1. Datasets and motivation for human genetics research

(A) Growth in human genetics dataset as exemplified by properties of selected landmark studies, plotted by the comprehensiveness of the genome analysis (x axis) with the technologies indicated on the top, and the number of donors (y axis). The type and quantity of phenotype data available are indicated by the dots. Underlined project names include a relatively balanced representation of individuals from diverse ancestries. The projects shown are Human Genome Project (HGP), HapMap, Wellcome Trust Case Control Consortium (WTCCC), 1000 Genomes (1KG), UK Biobank (UKBB), Pangenome project, Genotype Tissue Expression (GTEx), and Trans-Omics for Precision Medicine (TOPMed). WGS, whole-genome sequencing.

(B and C) Illustration of the two complementary approaches how human genetics contributes to human health. (B) illustrates how well-powered GWAS can allow building polygenic risk scores that can be used for personalized disease risk prediction. (C) illustrates how understanding the functional mechanisms of GWAS loci can allow targeting these mechanisms with drugs and other interventions to prevent or treat disease.

however, the term is often used to demarcate groups of humans and thereby implies discrete units of human genetic variation that is in stark contrast with the incredible amount of shared variation among all humans.[9,10] Although we will use population throughout this review in the technical biological sense, there is an increasing disciplinary call to shift from labeling human groups as discrete separate units, particularly when social systems have influenced those unit labels.[11,12] In fact, one of the ten "bold predictions for human genomics by 2030" from the US National Human Genome Research Institute's strategic plan is that "Research in human genomics will have moved beyond population descriptors based on historic social constructs such as race."[13]

Quoting Hubby and Lewontin,[14] who were the first to use gel electrophoresis to demonstrate variation at the genetic level in natural populations, "a description of the genetic variation in a population is the fundamental datum of evolutionary studies." Thus, population genetic studies that focus on the genetic variation in a population are crucial to understanding the genetic basis of complex traits, and many future opportunities for research lie at the intersection of human population genetics and statistical genetics, as we discuss later.

Genomic-era datasets for population genetics

Many insights into recent human population histories have been enabled by early projects pursued in tandem with the Human Genome Project (HGP), in particular the Human Genome Diversity Panel (HGDP)[15]; and the International HapMap Project (hereafter "HapMap").[16] The HGDP consisted of lymphoblastoid cell lines from 1,064 individuals from globally distributed populations, collected for characterization of human population genetic variation.[17] The HapMap was an international collaboration that began in 2002 and focused on the development of a haplotype map of the human genome, with a specific motivation to advance genetic association studies. The HapMap ultimately led to the 2009 release of over 1.6 million single-nucleotide variants (SNVs) from 1,301 samples from 11 populations.[16] The 1000 Genomes Project was initiated in 2008 as a continuation of the HapMap project to catalog the variants in the human genome that have a frequency of at least 1% in the populations studied. This was done by expanding focus from only SNVs to include other types of genetic variants and using both low- and high-coverage whole genome and exome sequencing, ushering in a

new phase of population genetics focusing on analyzing whole-genome sequences.[18,19] The 1000 Genomes Project's data include sequences from over 2,500 individuals from 26 populations.

Alongside the projects focused on characterizing genetic variation of the general population, with accessible data resources but no link to donor phenotypes, large case-control datasets created for genetic mapping of complex traits (see below) were also used for population genetic inference. Today, this trend continues with an increasing emphasis on biobank datasets that combine genetic datasets to comprehensive phenotyping from up to hundreds of thousands of individuals, often leveraging healthcare systems and registries. A major benefit of biobanks for population genetic research is that they offer high-resolution insight into gene flow, assortative mating, and population structure over the last few 100 years, enabled due to their scale and the presence of distant and close genetic relatives. For example, the UK Biobank (UKBB) contains over 40,000 first- and second-degree relative pairs,[20] and FinnGen contains over 30,000 first- and second-degree relative pairs.[21]

Although the early disease-focused case-control datasets had sparse data from non-European ancestries,[22] during the biobank era some progress has been made—at least in absolute numbers of non-European donors, such as in BioBank Japan and All of Us. However, European ancestries still dominate most biobanks, such as UKBB, FinnGen, deCODE, and the Estonian Biobank. Unfortunately, African populations remain understudied and underrepresented in genetic datasets, even though they hold particular value in understanding the origins of human genetic variation.[23–26] Furthermore, research on large-scale and biobank datasets often still ignores data from minoritized groups,[27,28] underscoring the importance not just of diverse data for analyses but of methods for handling imbalanced datasets and varying levels of linkage disequilibrium (LD, the correlation of alleles at different genetic variants) in human genetic studies. The need for data and methods is coupled with ethical, legal, and social issues in diversifying genetic research. Many population genetic studies, beginning with the HGDP, have grappled with and continue to grapple with ethical considerations regarding the collection and use of genetic data from participating individuals and their communities. For biobank projects and other genetic studies, informed consent practices are paramount, together with the need for community engagement and release of results to stakeholders.

Insights into human population history from genetic studies

Population genetics methods applied to data of genetic variation in natural populations enable inferences about the past based on four fundamental processes: mutation, recombination, drift (caused by finite population size), and selection. A fifth process of migration (gene flow) is increasingly appreciated as a force shaping human genetic variation at multiple timescales. Altogether, analysis of these processes has provided valuable insights into human population history and its contribution to the contemporary patterns of genetic variation in humans.

One of the major focus areas of population genetic research has been characterization of human migrations across vast timescales. New models for human origins have recently highlighted the complexity of deep population structure in Africa, which in turn offers paths to expand the focus of studies of archaic introgression into modern humans beyond patterns out of Africa.[29] In many geographic regions, local migrations over millennia have produced a high correlation between genetic distance and geographic distance.[30,31] However, historical events such as colonization and chattel slavery further led to the founding and persistence of admixed populations—populations descended from gene flow between two or more previously separated source populations, whose descendant individuals derive ancestry in differing proportions over time from the source populations.[32,33] The advances in sequencing technologies in the genomic era and introduction of large-scale datasets for medical studies have enabled insight into very recent and much more localized gene flow, for example during The Great Migration (1910–1970) of African Americans out of the US South.[34] Although some migration events are relatively well known from archeological and historical records, genetic data that captures biological ancestry has provided unique insights to population movements during human history.

Population founding events that have characterized much of human history lead to dramatic reductions in genetic variation. This, together with the relatively recent origin of our species in Africa, has resulted in a pattern where genetic differences among human individuals' genomes are very small and less than for many other species; common variants are often shared across populations; and most human genetic variation is quite rare and confined to single continental ancestries. Although SNVs occur in 3.1% of sites in the genome,[35] the vast majority of all the cataloged variants are vanishingly rare, and thus any two individuals differ by an average of a few million SNVs, representing less than 0.1% of the genome. Due to serial founder effects, the amount of genetic variation decreases with population distance from Africa: recent efforts to harmonize the HGDP and 1000 Genomes (1KG) high-quality whole-genome sequence data[15,19] counted an average of 6.1 M SNVs per African individual and 5.3 M SNVs in others,[36] with similar patterns for structural variants. In pairwise comparisons, two Yoruba individuals had 4,897,091 pairwise differences at sequenced SNVs, over 38% more than two French individuals or two East Asian individuals.[35] LD is also lowest in African ancestries, followed by European, Asian, and the American ancestries.

GENETIC ARCHITECTURE OF HUMAN DISEASES AND TRAITS

Connecting genetic variation to phenotypes and understanding the underlying biological mechanisms has been a fundamental goal of human genetics, but the means to achieving this goal have changed dramatically over the past decades. Initial efforts focused on genotyping individuals with severe or highly familial conditions to identify the causal pathogenic mutations they shared, under the assumption that these mutations would be highly penetrant and few in number. In parallel, linkage studies collected families with affected and unaffected individuals and traced the genetic segments that were overrepresented in cases, sometimes implicating large haplotypes with many

genes. As the cost of genotyping decreased, the study of common traits shifted toward association studies, wherein large cohorts of unrelated cases and controls were genotyped and individual variants tested for correlation with the trait of interest. An initial period of candidate gene association studies, where only predefined regions were genotyped and tested, led to contradictory findings,[37] with many questioning the contribution of common variants to common disease.[38] However, both theory and practical application of genome-wide association studies (GWASs), together with rigorous multiple test correction, began to yield robust associations that replicated across independent studies.[39] Even these early associations were often surprisingly weak, indicative of either a small contribution of common genetic variation to phenotype or a highly polygenic contribution involving many variants.[40] As GWAS sample sizes grew, evidence for the polygenicity of common traits accumulated,[41] implying that very large studies are necessary to identify the full spectrum of causal genetic variants. This has motivated the rise of large-scale biobanks and propelled the number of genome-wide significant associations into the hundreds of thousands, enabling highly precise estimates of "disease architecture": the number, frequencies, genomic distribution, and disease contributions of causal variants across the genome, discussed in detail below.

Recently, disease genetics has come full circle with large-scale sibling and family-based GWAS, which mirror the early linkage studies but at massive scale.[42] Family-based studies enable the partitioning of disease architecture into so-called direct and indirect effects; the former associated with variants within an individual and the latter associated with variants shared by their relatives (presumably acting through shared environments). Although still relatively small, these studies have demonstrated that many apparent genetic associations are, in fact, correlated with rather than causal for environmental influences on traits, potentially spanning generations and communities.[43] Although the study of common traits has primarily been driven by GWAS of common variants, enabled by inexpensive genotyping arrays, the contribution of rare variants is now being quantified through large-scale exome- and genome-sequencing studies that can capture the full spectrum of genetic variation.[44,45] Direct genetic association studies are often underpowered for rare variants, leading to the use of burden tests that "collapse" all variation in a tested gene. Analogous approaches, albeit with different implementations and standards that often include features of linkage analysis, are applied in Mendelian disease genetics.

Ubiquitous common-variant heritability

Decades before genotyping, the total contribution of genetics to a trait, i.e., the trait heritability, could be estimated through the use of twin and family-based studies. Under certain strict assumptions,[46] the increased correlation in phenotype between monozygotic and dizygotic twins or across family relationships can be decomposed into genetic and environmental components. Large-scale genotyping enabled the application of similar principles to putatively unrelated individuals by contrasting subtle patterns of genetic similarity with phenotypic similarity to estimate the so-called genotype-, SNP-, or chip-heritability. The resulting parameter quantifies the variance in phenotype explained by all genotyped variants and any untyped variation they are correlated with.[47] Multiple methodologies have been devised for estimating SNP heritability, either using individual-level data,[41] polygenic scores,[48] or only summary-level data,[49] but all of these approaches have converged on the general finding that most common traits have a significant SNP heritability. For example, in the UKBB, 551 common phenotypes had a mean SNP heritability of 10.9% and 15.6% across all illness and non-illness traits, respectively[50]; in the Biobank Japan, a mean SNP heritability was estimated to be 8.6% across 58 continuous traits.[51] Indeed, nearly every common biobank phenotype has some correlation with genetics, with 91% of traits in the FinnGen biobank[21] exhibiting at least one genome-wide significant association (for traits with >10,000 cases). The identification of some genetic variants influencing any common trait should thus be the expectation rather than the exception.

Extreme polygenicity of common traits

In addition to SNP heritability, another key parameter driving genetic discoveries is the trait polygenicity: the total number of causal variants influencing the trait and the distribution of their effect sizes. Highly polygenic traits involve many weak causal variants and require large sample sizes to characterize. Because most causal variants are still unknown, various quantifications of trait polygenicity have been proposed, such as the number of non-null effects on a trait,[52] the *effective* number of independent variants,[53] or the minimum number of causal variants explaining a given fraction of heritability.[54] Regardless of the statistical model, polygenicity has been consistently estimated to be very high, ranging from thousands of causal variants for some estimators[52,53] to millions of variants for others.[54] These staggering estimates would imply that, for some traits, many causal variants are acting through nearly every gene in the genome on average and implicate more than half of all common polymorphisms.[54] In general, cellular and pigmentation traits exhibit the lowest polygenicity (hundreds of causal variants[52,54]), whereas anthropometric and cognitive/behavioral traits exhibit some of the highest estimates (>10,000 effective variants[53]). Although traits with similar heritabilities often exhibited different levels of polygenicity,[52] the variance in polygenicity across traits was generally lower than expected, suggesting that selection, one of the factors driving polygenicity, may be acting pleiotropically across traits rather than on any one measured phenotype.[55,56] Recently, a GWAS of height in 5.4 million participants demonstrated that 12,111 jointly significant variants explained 40% of the phenotypic variation (compared with total SNP heritability of 45%), lending the first direct evidence for high trait polygenicity.[57] The evolutionary causes of high polygenicity continue to be actively investigated,[55] but the implications are clear: understanding human traits will require distilling the function of tens of thousands of variants.[58,59]

Functional partitioning of disease polygenicity

Similar to partitioned SNP heritability, polygenicity can also be partitioned to quantify whether a given functional annotation contains variants with strong or weak effects on disease. Strikingly, estimates of partitioned polygenicity exhibit very high

correlation with partitioned heritability (r^2 = 0.88).[53] SNPs in conserved regions, for example, are enriched 13× for heritability and likewise enriched 14× for polygenicity relative to other SNPs, implying that their outsized contribution to heritability may be due in part to an increase in the number of causal variants rather than an increase in the absolute effect sizes. This model, referred to as a "flattening" of heritability, posits that natural selection has distributed (or "flattened") causal variation in functionally important regions to be more polygenic. Because higher polygenicity also leads to decreased GWAS power, the most significant GWAS associations (and those identified in smaller GWASs) may thus not reside in the most functionally "important" regions. The flattening of genetic effects may also explain why many complex traits appear to be "omnigenic"[58]: governed by a small number of "core" genes with direct effects on the trait, which are in turn disproportionately dampened by negative selection, thus increasing the relative contribution of "peripheral" genes with no direct connection to the trait.[60] One interpretation of both the flattening and omnigenic models is that mapping core genes from top GWAS hits alone may be difficult, with large variant effect sizes not implicating the most biologically relevant genes or drug targets. Intriguingly, recent analyses have shown that approved drug target genes are enriched for GWAS association evidence, regardless of the effect size, allele frequency, or year of GWAS.[7] Larger, better powered GWAS may thus continue to yield important insights into disease mechanisms and therapeutics or even increase in relevance.

Rare-variant heritability

The emergence of large whole genome and whole exome sequencing studies has started to enable the characterization of rare- and low-frequency-variant disease architectures. Initially, studies relied on genotype imputation from reference data to explore the heritability of low-frequency variants (0.5%–5%). For example, across 40 UKBB traits, coding variants explained a greater proportion of low-frequency SNP heritability (17%, 38× enriched) than of common SNP heritability (2%, 7.7× enriched), consistent with the action of negative selection in keeping large effect (typically coding) variants at lower frequencies.[61] Nevertheless, all variants in coding and untranslated regions (i.e., those captured by exome-sequencing) still explained only 26.8% of low-frequency SNP heritability, indicating that whole-genome sequencing would be necessary to identify most low-frequency effects.

Recently, whole-genome sequencing data from 25,465 unrelated individuals was leveraged to estimate total SNP heritability, including rare variants.[62] These total heritability estimates were 68% for heights and 30% for body mass index, contrasted with a common SNP heritability of 48% and 24%, respectively. Rare variants may thus increase the explained trait variance by 1.25–1.4× relative to common variants alone. A major contribution to the heritability of height came from very rare variants in low LD, which are particularly difficult to impute from reference panels. Although a fundamental advance, the study had limitations: the use of complex heritability partitioning to account for allele frequency and LD biases, and the use of conventional common-variant approaches to account for population structure (which can fail for rare variation). More data, more traits, and novel methodological approaches will continue to shed light on the question of whole-genome heritability. Intriguingly, both estimates were significantly lower than prior estimates from twin studies, implying either the existence of additional untyped genetic variation (for example, due to structural variants) or systematic biases in the twin cohort analyses.

Emerging methods such as burden heritability regression[63] have expanded the estimation of genome-wide partitioned heritability to rare variation. Under the assumption that rare alleles are likely to have consistent effects within a gene, this approach quantifies the total variance in a trait that can be explained by gene burdens across all genes. When applied to 6.9 million coding variants across 22 common traits in the UKBB, the average burden heritability was estimated to be 1.3% (for loss-of-function and missense variants below 0.001 frequency) and significantly non-zero for each trait. Notably, genes that were individually significant in an independent analysis of the same cohort often explained a large fraction of the burden heritability: for example, *APOB* alone explained 39% of the burden heritability for LDL cholesterol, and 172 known tumor suppressor genes explained 48% of the burden heritability for a composite cancer phenotype. If accurate, these estimates would imply that the rare-variant trait architecture is much less polygenic than the extreme polygenicity often observed for common variants. We caution that the characterization of rare-variant disease architecture is still in its infancy, larger cohorts and orthogonal methods needed to understand these parameters and to move beyond relatively simple burden models.

Several large-scale exome-wide association studies have now been conducted and have yielded novel rare-variant associations. Exome sequencing data from ~450,000 individuals in the UKBB were tested for association with ~4,000 traits, identifying 8,865 significant associations across 564 genes.[45] Multiple insights into disease architecture were observed. First, rare coding associations were significantly enriched near common GWAS loci, with an enrichment of 59.3× for the nearest gene to a GWAS association, decreasing to 11.4× for genes within one megabase of a GWAS association. These findings show a striking convergence of rare- and common-variant effects on common diseases. Second, target genes for approved drugs were 3.6× more likely to exhibit an association, consistent with prior findings that drug targets with human genetics evidence are more likely to be approved. Third, 77% of associations could only be identified using a burden analysis and not single-variant associations, underscoring burden heritability as a major driver of discoveries at this sample size. Fourth, although disease lowering associations are potentially the most attractive drug targets, only five such associations were identified and all were previously known, indicative of low power to detect protective effects in unascertained cohorts. Although most rare-variant associations have been deleterious, an exome study of smoking behavior in n = 749,459 individuals, one of the largest to date, identified rare variants in *CHRNB2* associated with a 35% decreased odds for smoking heavily.[64] This finding highlights the growing opportunities for discovering new levers into the treatment of common phenotypes.

Trans-ancestry genetic architecture

Although most of the above analyses focused on genetic architecture within presumptively homogeneous populations, progress is being made toward understanding genetic architecture across different populations. Theoretical and data-driven studies demonstrated that individual variant associations and aggregates of associations in polygenic scores are likely to translate poorly to genetically distant populations even if the underlying causal variants are shared.[65,66] This lack of transferability can be driven by a mixture of differences in causal variant allele frequency, LD patterns to non-causal variants, and the true underlying effects (modulated, for example, by gene-gene or gene-environment interactions). A strikingly linear relationship between genetic distance and polygenic risk score predictive accuracy was recently demonstrated in a large, admixed biobank across 84 traits (mean Pearson correlation of −0.95 between genetic distance and accuracy).[67] Importantly, although the mean risk score value also correlated significantly with genetic distance, the strength and direction of the correlation varied substantially across traits and populations, highlighting the challenges of correcting trans-ancestry score estimates. Beyond demonstrating lack of portability, the contributions of frequency, LD, and effect size are also now being quantified. A recent study of admixed individuals used *local* ancestry to quantify the correlation in causal effect sizes between African and European ancestry segments.[68] The estimate was remarkably high, with a mean causal effect-size correlation of 0.95 ± 0.02 across 38 traits and three very different biobanks. This high genetic correlation was also consistent with prior work showing that poor polygenic score portability may be largely explained by frequency and LD differences between populations rather than different causal variants.[69] Intriguingly, the genetic correlation was significantly lower (0.50 ± 0.07) when estimated across non-admixed individuals from different populations in the same study, with prior studies also showing trans-ancestry correlations ranging from 0.46 to 0.85 and generally well below 1.0[70,71] Given the dearth of existing multi-ancestry cohorts,[27] these findings and open questions further emphasize the importance of designing association studies to maximize population-level and individual-level genetic diversity. Indeed, large multi-ancestry biobanks have already demonstrated increased ability to identify and refine causal variants.[72,73]

Multiple approaches are emerging for maximizing the utility of genetic data across populations and ancestries. Many studies have shown that both polygenic prediction and variant fine-mapping can be improved by incorporating functional annotations.[74,75] Notably, the trans-ancestry gains in polygenic prediction accuracy are often substantially larger than the within-population gains, suggesting that better identification of causal variants can mitigate some of the heterogeneity due to frequency or LD differences across populations. Furthermore, power can be increased by aggregating (potentially heterogeneous) variant-level effects into sets such as genes and then combining the effects of these sets across populations.[28,76,77] Such variant sets could in principle be aggregated at various biological scales—genes, pathways—and their effects further propagated through biological networks. These approaches highlight how deeper understanding of the causal biological network across traits and populations can be incorporated back into multi-ancestry analyses to further improve power.

Opportunities at the interface of human population genetics and statistical genetics

As the HGP outlined from its inception, research in human genomics is motivated by gaining understanding of the genetic basis of human disease and complex traits. Moving forward, such research must draw on population genetic models and data from the full diversity of our species. Here we outline a series of opportunities for research at the intersection of statistical and population genetics.

As discussed above, trans-ancestry transferability of genetic associations and polygenic scores remains a key challenge in the field, exacerbated by the ongoing lack of well-powered datasets for many non-European ancestries. Although initial studies in admixed populations suggest that causal effect sizes may be largely shared across populations, the full extent of gene-gene and gene-environment interactions, as well as their population or trait specificity, remains to be quantified. Understanding these parameters will further inform the optimal design of accurate predictive scores across the full range of human diversity. Population genetic summaries and models of ancestry are needed to increase transferability of association results, especially for individuals of mixed ancestry for whom local ancestry-aware score construction may improve predictive accuracy.[78,79] Multiple studies have shown that increased genetic diversity improves the resolution of statistical fine-mapping, which in turn increases the accuracy of polygenic scores. Thus, greater diversity of data from underrepresented individuals will yield immediate benefits for all individuals.

Additionally, environmental heterogeneity pervades human genetic studies and confounds our understanding of the genetic basis of human disease and complex traits,[80] leading to poor prediction of traits from genetic data alone. Models of assortative mating and consanguinity highlight how violations of standard population genetic assumptions of random mating and equilibrium population dynamics can inflate observed correlations between human traits.[81] Recent clustering[82] and contrastive learning approaches[83] highlight confounding factors that bias the downstream estimation of genetic effects. Although family-based studies offer the ability to estimate direct and indirect effects of genetic variation on a given trait,[42,84] family-based estimates of direct effects can be biased by genetic confounding,[85] in ways that are compounded when estimating susceptibility using genome-wide association results. The scale of biobanks, increasing detail of metadata and environmental covariates, and the development of longitudinal follow-up efforts will enable more awareness and better controlling for environmental confounders, as well as allow for leveraging distant genetic relatives for estimating genetic effects and understanding recent population genetic processes such as pedigree collapse. Modeling the full complexity of human relationships, environmental correlations, and interactions will increase the causal validity and generalizability of genetic discoveries.

In order to prioritize traits for risk prediction and genetic studies, evolutionary models for complex trait architecture are key. Recent work on stabilizing selection suggests that trait

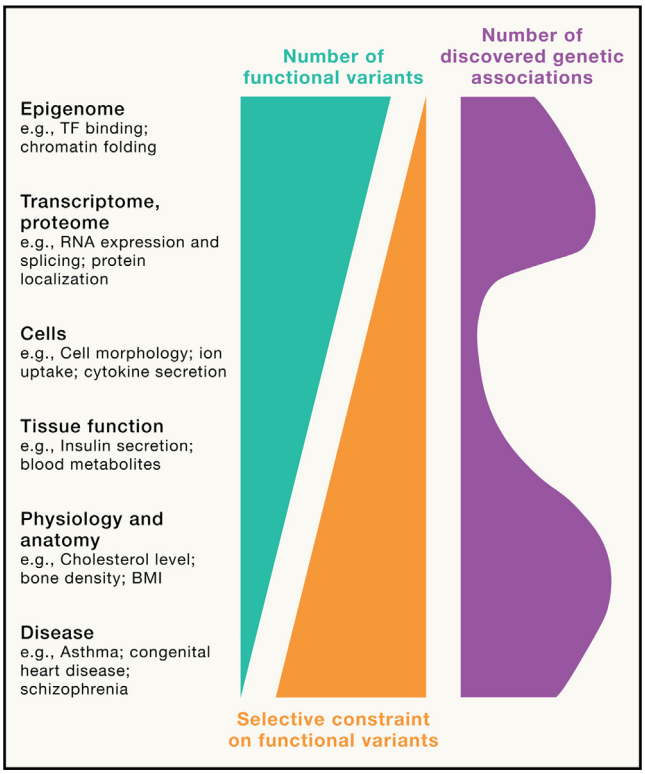

Figure 2. An illustration of genetic effects on functions at different levels

There are large numbers of variants affecting molecular functions of the genome and the cell, many of which have no or smaller effects downstream. Variants affecting physiological, anatomical, and disease traits can be under direct natural selection. The purple graph indicates the success in discovery of genetic associations for molecular traits (captured by molQTL mapping) and for physiological and disease traits (captured by classical GWAS), with a gap in our knowledge of genetic associations for cellular and tissue-level traits.

architectures for many well-studied complex quantitative traits are similar in their polygenicity[86] and that this mode of selection may lead to less cross-population transferability of association results.[87] Additionally, traits with smaller effects on fitness produce less transferable associations, with weak negative selection producing more population-specific trait architectures.[88] Quantifying the extent of polygenic selection and adaptation on complex traits remains a great challenge, in part due to the complexities of disentangling subtle population stratification.[89,90] Improved methods to detect fine-scale population structure,[91] larger within-family analyses,[43] and comprehensive models of selection will enable a more complete understanding of genome-wide evolutionary processes. Beyond answering fundamental questions in human evolution, these findings will also have practical implications: how to mine the thousands of trait-associated loci for the most disease relevant genes and drug targets, and how to integrate the findings from rare and common variation.

Beyond understanding the mechanisms of known associations, there are many opportunities to incorporate novel or difficult-to-collect variation. Large-scale biobanks have highlighted the important role of structural variation, including copy-number variants and tandem repeats with some of the largest effect sizes on traits seen to date.[92] Structural variants are typically not directly genotyped and often not imputed, leaving a gap in our knowledge of disease mechanisms beyond single variants. Methods for identification of complex structural changes directly from data,[93] as well as improvements in genome assembly,[94] may reveal entirely new classes of disease relevant variation. Similarly, although the role of additive and dominance variation has been well characterized through large-scale biobank and heritability analyses,[95] the influence of epistatic effects remains largely a mystery. Although genetic interactions and hotspots are widespread in model organisms[96] they have been challenging to characterize in humans due to the breadth of the search space and statistical limitations.[97] This is especially true for more complex relationships beyond simple pairwise interactions, which may be impossible to even enumerate in human populations. Integration of population genetic modeling[98] and functional studies[99] may push through the statistical limitations and expand our understanding of trait effects into higher orders.

Continued research at the nexus of population and statistical genetics, as well as the increased ability to study traits in biobanks using family-based and genealogical approaches, will help make gains toward improved variant discovery and risk prediction while identifying traits with large environmental influences for which additional studies, data, and approaches will be needed for risk assessment, treatment, and prevention.

MOLECULAR AND CELLULAR EFFECTS OF GENETIC VARIATION

Genetic variants affecting complex physiological traits and diseases must have proximal effects on molecular functions, which then impact subsequent molecular processes at the cellular level. Deciphering these molecular and cellular mediators of genetic associations has emerged as a central focus of contemporary human genetics, as it can offer insights into molecular understanding of causal processes of disease. The significance of this extends beyond fundamental biology since these processes serve as potential intervention targets[7,8] (Figure 1C). Furthermore, although many molecular effects of variants have no impact on physiological phenotypes (Figure 2), they represent a natural experiment of variations in the genome sequence, which can contribute to understanding the biology of genome function.

Methods for analyzing the functional effects of genetic variants

Although the analysis of molecular effects of variants has been part of molecular genetics from its inception, it was not until the DNA hybridization array technology in the 2000s that genome-wide analysis became feasible. This technological advancement led to expression quantitative trait locus (eQTL) mapping to identify variants associated with gene expression levels, an approach first applied to humans about 15 years ago (Figure 3A). Since then, this method has evolved to cover molecular phenotypes from epigenomic measurements to splicing and protein levels, collectively often referred to as molecular QTLs (molQTLs). Large-scale projects have constructed molQTL

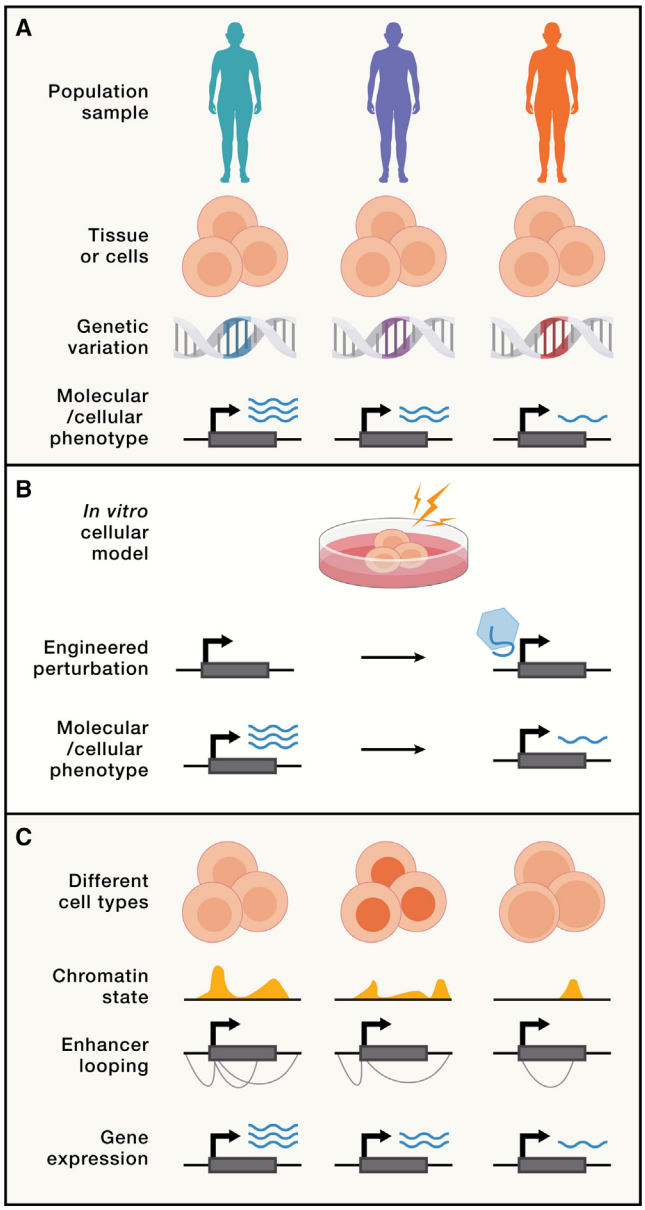

Figure 3. Approaches for understanding molecular effects of genetic variants at scale

(A) molQTL mapping.

(B) Engineered perturbations of the genome.

(C) Inference from multi-layered functional omics data.

resources for various tissues and cells, including under *in vitro* stimuli, with an increasing use of single-cell technologies.[100] Most molQTLs robustly identified to date are in *cis*, i.e., affecting a nearby target gene via *cis*-regulatory mechanisms, because *trans*-QTLs between variants and genes across the genome can be reliably identified only with large sample sizes and careful control of confounders.[101] molQTL methodology is reviewed in detail elsewhere.[102]

Experimental genome perturbations in *in vitro* cellular systems have rapidly become popular tools for scalable mapping of molecular effects of genetic variants (Figure 3B). These approaches include episomal assays such as massively parallel reporter assay (MPRA) and perturbations of the genome using the CRISPR toolkit, coupled with diverse readouts of molecular effects. Ongoing efforts such as Impact of Genomic Variation on Function (IGVF) and Atlas of Variant Effects (AVE) pursue more systematic application of these tools toward both noncoding and coding variation. A key prerequisite for most of these approaches is high-quality fine-mapping to target the likely causal variant(s) at associated loci, and the improving methods and resources from the GWAS community will thus greatly enhance these experimental efforts.

Furthermore, the vast functional genomics datasets from projects such as ENCODE provide a powerful foundation for predictive inference of genetic variant effects even when genome variation is not directly assayed (Figure 3C). Development of these methods is a highly active area of research, with progress particularly for predicting the effects of coding and splice-affecting variation,[103–105] while predicting the effects of transcriptional regulatory variants[106] has proven to be challenging.[107]

Molecular architecture of complex trait loci

There are likely dozens of different molecular mechanisms by which genetic variation can impact organismal phenotypes. Among these mechanisms, perhaps the most easily interpretable is that of coding variants, which directly impact protein coding sequence and function. However, unlike early mapping of Mendelian disorder variants that found causal SNPs to nearly always affect coding sequences, GWASs have revealed very early on already that genetic variants underlying complex trait associations are often noncoding and impact gene expression.[108,109] These discoveries motivated an explosion of interest in understanding how genetic variants impact gene regulation, and particularly how they impact gene expression levels.

Over the last decade, several statistical methods have been developed and deployed on different datasets to identify functional enrichments of GWAS loci. The most compelling GWAS functional enrichments identified thus far have been in regions with high chromatin accessibility or in regions marked by histone modifications associated with enhancers and promoters.[109–112] In fact, the majority of SNP heritability for a variety of common traits can be localized to regulatory rather than coding regions, with estimates of up to 79% of SNP heritability residing in DNAse I hypersensitive sites (spanning 16% of variants, a 4.9× enrichment) across 11 diseases,[113] or 15% of SNP heritability residing in enhancer elements (spanning 0.4% of variants, a 37.5× enrichment) across 17 common traits.[111] Additionally, regions conserved in mammals were estimated to harbor 35% of SNP heritability (spanning 2.6% of variants, a 9.6× enrichment),[111] consistent with the expected role of evolutionary constrained elements in shaping disease architecture.

GWAS loci have also been reported to be enriched among variants associated with multiple types of regulatory variants. As expected, these include genetic variants that impact gene expression level regulation, e.g., by affecting DNA methylation,[114] histone modification levels,[115] and chromatin accessibility.[116] The enrichment of trait heritability among eQTL fine-mapped SNPs is similar to that in enhancer regions (about 5× for

non-specific eQTLs/enhancers and 20× for eQTLs or enhancers specifically identified in trait-relevant cell types).[111,117] However, the total trait heritability estimated to be explained by common variants overlapping eQTL SNPs (averaging 11%[118] or 14%[117] across traits estimated by mediation or enrichment analysis, respectively) tend to be much smaller than that overlapping enhancer and promoter regions (23.9%–79.2%[111]). Although the 11%–14% and 80% estimates likely represent conservative and optimistic estimates for heritability explained by eQTLs and variants in enhancers or promoters, respectively, these observations suggest that the quality of our maps of eQTLs lag far behind that of enhancer and promoter regions and that more work is needed to understand how regulatory variants impact gene expression.

In addition to variants that impact gene expression levels, GWAS loci are also enriched in many other types of molQTLs such as those with effects on mRNA splicing[119,120] and other effects on transcript structure[121] and posttranscriptional modifications.[122] The estimated enrichment of GWAS signals in these molQTLs is highly variable and may be trait-dependent. For example, several studies have reported a higher enrichment of neuropsychiatric GWAS loci among variants that impact RNA splicing QTLs (sQTLs) than compared with that among eQTLs[123]; and a recent study found higher enrichment of autoimmune GWAS signal in RNA editing QTLs than compared with both eQTLs and sQTLs.[122] To date, the total heritability explained by different posttranscriptional molQTLs pale in comparison with that explained by eQTLs or variants in enhancer and promoter regions. Though this may simply reflect the fact that eQTLs and enhancers have been studied at much larger scales and in a wider number of cell types compared with other regulatory mechanisms.

In addition to functional enrichments that indicate *cis*-regulatory mechanisms of immediate molecular drivers of GWAS loci, GWAS offers a unique causality anchor to identify cell types and cell states where the causal molecular processes contributing to traits are taking place. Understanding the cell type specificity of these can inform disease biology as well as potential targets of interventions that minimize off-target side effects. For most complex traits, it is far from trivial to infer causal cell types from clinical characteristics, as symptoms of a disease can pinpoint different tissues, cell types, or developmental stages than where processes that are causal to disease take place. The most fruitful approach to addressing this challenge has been to analyze GWAS heritability enrichment in genes and regulatory elements that are active in specific tissues and cell types.[112,124] Notably, a major finding has been that the enrichment of GWAS loci in enhancer/promoter regions is highest in cell types or tissue types that make intuitive sense. For example, autoimmune disease loci are most enriched in enhancers active in immune cell types (e.g., T cells and B cells), while neuropsychiatric disease loci are most enriched in neuronal cell types. Still, these enrichments often only give us a coarse-grain idea of which cell types contribute to a trait or disease, and more research is needed to find out whether the genetic signal is strong enough to identify more precise causal cell types. In fact, previous work observed that although genes or enhancers with cell-type-specific patterns of activity were highly enriched in trait heritability, the bulk of the heritability was found in genes or enhancers that were broadly active in many or most cell types.[58] These observations suggest that most of the genetic signal will be in enhancers with broad, rather than cell-type-specific activity. Thus, it is possible that the pleiotropic nature of functional enhancers limits our ability to use genetic signals to "fine-map" causal cell types.

Interpreting complex trait loci using molQTLs

molQTL mapping is fundamentally a genetic method, while the other approaches showcased in Figure 3 are rooted in molecular and computational biology. Thus, we will discuss molQTL mapping in more detail below, with further discussion of history and methodology provided e.g., in Aguet et al.[102] and Albert and Kruglyak.[125]

In the early 2010s, straightforward analyses that overlap significant GWAS and eQTL SNPs were used to identify variants associated with traits that also had a functional impact on gene expression levels, helping to identify potential causal gene. However, with the increasing number of GWAS and eQTL signals, it became evident that new statistical methods were needed, in particular, to address the scenario where a large proportion of variants are associated with some molecular phenotypes due to LD.[126] As a result, several advanced statistical methods were developed to assess "colocalization": whether the same variant(s) were likely *causal* drivers for both a GWAS signal and a molQTL signal in a specific genetic locus.[127,128] An alternative approach is the estimation of "molecular association" such as in transcriptome-wide association studies (TWASs), which test for an association between a predicted molecular phenotype (e.g., expression) and the trait,[129] and can similarly be applied to summary-level data.[130,131] This approach relaxes the requirement of colocalization that causal variants are shared between the molecular and disease traits—they merely need to be correlated—while increasing sensitivity through the use of multivariable predictive models.[132,133] Although neither approach can guarantee that the particular gene's expression is causally related to disease etiology, colocalization removes spurious overlaps due to LD, and molecular association enables sensitive quantification of the correlated effect and direction.

A notable observation from employing these methods is the low fraction of GWAS loci that colocalize with eQTLs[134] or can be explained by molecular associations.[135] This is evident even for many immune or blood-related traits where the available eQTL data from relevant cell types is assumed to be comprehensive. For example, only about 25% of autoimmune trait GWAS loci colocalize with an eQTL from different immune cell types.[136] Adding other molQTLs such as sQTLs can increase the colocalization rate, but still leaves the majority of autoimmune-associated loci without a colocalization.[137] More generally across complex traits, genetic effects mediated by *cis*-eQTLs account for an average of just 11% of trait heritability.[135] An additional complication is that regulatory elements and variants can regulate multiple genes, and the one picked up by a colocalizing eQTL is not necessarily the truly causal disease gene in the locus.[138]

Several reasons may account for the relatively modest rate of GWAS colocalization. First, genetic variants that affect gene regulation independent of gene expression level may play a

larger role than we previously anticipated. Although most research has concentrated on how genetic variants regulate gene expression levels, genetic variants can influence cellular biology through various other regulatory mechanisms, as previously discussed. Nevertheless, because the rate of colocalization between GWAS and expression QTLs is the highest across nearly all complex traits and among all molQTLs, the prevailing opinion is that the majority of trait variants operate by affecting protein expression levels. Supporting this, several recent studies found that genetic variants that impact chromatin activity (histone mark QTLs, or hQTLs) or accessibility (chromatin accessibility QTLs, or caQTLs) colocalize at much higher rates (sometimes ~50% more) than eQTLs from the same cell- or tissue types.[139,140] These observations imply that trait-associated variants often regulate gene expression levels by modulating enhancer or promoter activity. Yet, the ability to statistically detect their impact on gene expression might be weaker than on chromatin-level phenotypes.[141] This aligns with the idea that enhancer activity has a simpler genetic architecture than gene expression level, as steady-state mRNA expression levels are affected by co- and post-transcriptional mRNA processing mechanisms in addition to mechanisms that impact transcription initiation. Another explanation for the higher chromatin QTL colocalizations compared with eQTLs is a difference in the discovery thresholds: for chromatin QTLs to be detected, the enhancer must be active in a given cell- or tissue type,[142] whereas for the same variant to be an eQTL, the enhancer must both be active and also drive gene transcription. For example, "primed" enhancers have been found to harbor caQTLs in multiple types of naive immune cells, but they appear to be eQTLs only in cells that were stimulated by cytokines or pathogens.[143]

Several theories have been proposed to explain the limited overlap between molQTLs and GWAS hits. Genes involved in complex traits may have redundant enhancers that buffer the impact of genetic variants on gene expression levels, making those eQTLs that are relevant to complex traits more difficult to identify.[144] Along similar lines, features of eQTL mapping may favor discovery of loci and genes with lower selective constraint, regulatory complexity, and functional importance, thus biasing the results away from identification of genes that underlie trait variation.[145] Another possible explanation is that for most traits, we have not studied gene expression in the cell types or cell states that are most relevant for disease.[134] Indeed, despite substantial sharing of molQTLs across tissues, it is possible that many dynamic QTLs dependent on temporal context of cellular state can only be identified in some as yet unexamined rare cell type or developmental trajectory.[146]

A pessimistic interpretation may be that the value of eQTL studies in interpreting complex trait-associated variants is and will continue to be modest in the future. However, it may be wise to recall that early discoveries from GWASs with limited sample sizes were also very modest.[38] As the sample size of GWAS grew in the tens and hundreds of thousands, transformational insights emerged, many of which now shape our understanding of human traits and biology. Analogous scaling up of sample sizes on molQTL studies to the hundreds of thousands has not been done, and the largest studies[101] come from bulk tissue samples, which limits the power, resolution, and interpretability of detecting regulatory effects that may operate and drive disease in specific cell types and cell states. Although even larger and context-specific molQTL maps covering all relevant cell types are unlikely to provide a singular complete solution to molecular interpretation of GWAS loci, it remains as the only approach that allows interrogation of genetic variant effects in diverse primary cell types. Thus, we foresee that expanding molQTL studies will continue to have value in the future, alongside other approaches that use *in vitro* perturbations and computational inference from epigenomic data (Figure 3).

Cellular programs and physiological effects of disease-associated loci

During the past 10 years, the main focus in functional interpretation of GWAS has been on identifying the causal driver genes in the locus. Although the toolkit for this inference is still incomplete, in hundreds if not thousands of GWAS loci, the causal gene in *cis* has been identified with a reasonable confidence.[147] However, these studies have so far provided limited information about the cellular programs and downstream physiological mechanisms that underlie a disease (Figure 2). This is due to two major gaps in our knowledge: functional annotation of human genes is very incomplete, and understanding of cellular programs and regulatory networks that tie individual genes into broader cellular behaviors is even more incomplete. Furthermore, genes and variants can have pleiotropic effects across cells and tissues, making it difficult to distinguish disease-causing effects.

Thus, our advancing interpretation of *cis*-regulatory mechanisms must be coupled with vigorous efforts to link variants and genes to cellular programs and further to physiological mechanisms that underlie traits and diseases. There are numerous approaches to pursue this goal, typically extending the concepts outlined in Figure 3 to cellular phenotyping that is informative of functional effects beyond the *cis*-regulatory space. Well-powered GWAS with a large number of loci has allowed enrichment analyses of the implicated genes in annotations of cellular networks and pathways, pinpointing likely trait-relevant functions (e.g., Hsu et al.[148] and Morris et al.[149]). Furthermore, GWAS variants can be directly associated to molecular traits across the genome, in particular via *trans*-eQTL mapping. This requires very large sample sizes of thousands of individuals and careful analysis to avoid confounding factors.[101] Future single-cell analyses have the potential to further increase the informativeness of this approach. GWAS for measured or inferred cellular traits such as transcription factor activity or cell morphology can further link molecular changes to cellular programs, but phenotyping in adequately large sample sizes has been a challenge (Figure 2). GWAS for large-scale measurements of tissue-level phenotypes provide interesting examples of mechanistically more interpretable traits that lie between molecular traits measured directly from cells and highly complex physiological phenotypes captured by classical GWAS. An emerging approach for characterization of genetic effects for cellular traits is "cell villages" where cells from multiple donors are grown together and phenotyped by, e.g., cell sorting, and enrichment of genetic variants in cellular phenotype groups indicates an association to that trait.

Table 1. Outstanding challenges for human genetics research to tackle within the next 5–10 years, with a focus on population genetics, common complex traits, and basic research

Challenge	Goal	Ways forward
Completing genetic variation maps	comprehensive characterization of all types of genomic variation across the global population	long-read sequencing of tens of thousands of individuals across the global population
Mechanisms of human adaptation	identification of genetic variants and genes underlying human adaptation	data from diverse populations; statistical models; functional follow-up
Map of selective constraint	annotation of genomic elements and molecular processes under selective constraint, a key metric of functional relevance	massive genetic datasets
Gene-environment interactions and correlations	quantifying and controlling for environmental heterogeneity in biobank datasets, identifying important environmental confounders	harmonized metadata across biobanks with longitudinal follow-up and geographic mapping; diverse family-based study designs
Causal GWAS genes in *cis*	a robust and reasonably accurate toolkit for *in silico* annotation of likely causal driver genes for any GWAS locus	integration of different tools (molQTL also from single cells, enhancer maps, CRISPR) with gold-standard annotations
Regulatory code of the genome	prediction of context-specific *cis*-regulatory effects of genetic variants	functional genomics data and validation datasets combined with deep learning and artificial intelligence methods
Cellular programs underlying complex disease	identifying cellular processes that mediate GWAS associations and the cell states where they take place	integration of human genetics with large-scale *in vitro* experiments and molecular cell biology
Organ and physiological processes underlying complex disease	identifying changes in tissue and organ functions and other physiological phenotypes that mediate GWAS associations	measurement or inference of these lower-level traits for GWAS; organoids and model organism research
pheWAS (phenotype-wide association study) interpretation	inference of interpretable, causal relationships between traits from pheWAS data	advanced statistical models; comprehensive and interpretable phenotypes and metadata
Translatable and interpretable polygenic scores	genetic predictors that incorporate common and rare variants and environmental risk factors and are translatable between different groups	advanced statistical models; more diverse GWAS datasets; better fine-mapping

CONCLUSIONS AND FUTURE PROSPECTS

Human genetics continues to thrive as a very dynamic field. As discussed above, the expanding and diversifying datasets that include not only genetic variation but also phenotype data, environmental factors, and family relationships provide ample opportunity for understanding the population genetic processes that have given rise to the current spectrum of genetic variation in humans. Genetic association studies are finally starting to cover the full spectrum of different types of variants across the frequency spectrum. The integration of population genetics and statistical genetics provides a rich opportunity for improved mapping of genetic architecture of complex traits. As the number of robustly identified genetic associations has exploded, the challenge of their functional interpretation has become a central question in the field, now tackled with a combination of tools expanding beyond genetics to molecular computational biology.

These advances have also shed light on persistent challenges and open questions in the field, some of which are highlighted in Table 1. Beyond generating more data, advances are likely to come from the synthesis of insights across the disciplines of quantitative, molecular, population, and epidemiological genetics. Understanding the causes and consequences of disease architecture (widespread polygenicity and pleiotropy, in particular) will require advances in quantitative genetics to incorporate parameters of natural selection coupled with advances in genetic epidemiology to understand the relevant environmental contexts and risk factors shared across traits. The latter, in turn, will likely benefit from the partitioning of environmental and genetic variance enabled by advances in family-based study designs, which are also beginning to capitalize on fundamental theories from causal inference and counterfactual reasoning to aid interpretation.[150] Understanding the language and grammar of regulatory variation will require integration of population-scale quantitative genetics, which can quantify the effects of standing variation *in vivo* and in a disease context, with experimental molecular genetics, which can probe the effects of unobserved perturbations and validate novel predictions *in vitro* and in a non-disease or synthetic disease context. Both approaches can benefit from emerging tools in statistical genetics and machine learning for prediction, prioritization, and feature interpretation to more efficiently identify the most relevant disease genes and their broader disease network effects. Finally, all of these inquiries can benefit from diverse, multi-ancestry cohorts and advances in population genetics to

understand the complex genetic genealogy of contemporary populations using modern day and ancient genomic data.

Given the rapid progress during this millennium, human genetics is now well poised to provide a deeper understanding of human biology—and improve human health. Successful description of genetic variation, mapping of genetic associations, and identification of their functional effects provides the foundation for mechanistic understanding of these processes. This opens the door to successful prediction in diagnostic settings and identification of interventions to affect processes that contribute to disease. The ultimate goal is the integration of rare and common genetic risk factors with environmental risks, as well as pharmacogenetic advancements in tailoring treatment selection.[5] Equally important is to map the limits of genetics and understand how the broad patterns of heritability rise via complex interrelated processes of genetics and diverse environmental factors throughout an individual's life. Here, the interactions between human genetics and neighboring fields, such as epidemiology and molecular biology, are critical.

Beyond these challenges and ways forward, continued success and global justification of human genetics necessitate scrutiny of the field as a professional community, as well as its relationship with the surrounding society. Like other scientific domains, human genetics has a problematic history and ongoing issues linked to exploitation and exclusion of Indigenous communities and minorities both within the professional community and as research participants. Confronting and addressing these issues is essential to pave the way for a more inclusive and responsible future.[151,152] Genetics research is increasingly incorporated into the social sciences, and effective communications across these disciplines is needed to ensure the limits of genetic inference are fully understood.

Yet, human genetics provides some of the most compelling empirical evidence of our collective origins and the intertwined biological makeup of all humans, as well as the complexity and nondeterministic nature of human traits and diseases. This narrative could be disseminated more extensively. Genetics already has a major role in public understanding of our personal family history and ancestry, and it is assuming an increasingly prominent role in healthcare. Thus, the imperative to foster a society that is well-versed in the appropriate use and limitations of genetic data. This requires moving away from the reductive and deterministic language often employed in public communication of genetics. Embracing a more inclusive, transparent, and ethically aware approach is not just a moral imperative but also crucial for the sustained progress and credibility of the field.

ACKNOWLEDGMENTS

T.L. is funded by NIH grants R01AG057422, R01MH106842, and U24HG012090; ERC grant 101043238; and the Göran Gustafsson Foundation. Y.I.L. is funded by NIH grants R01GM130738 and R01HG011067, and a GREGoR Consortium Grant, and by the W.M. Keck Foundation. S.R. is funded by NIH grant R35 GM139628. A.G. is supported by NIH grants R01HG006399, R01HG012133, R01MH125252, and R01CA262577.

DECLARATION OF INTERESTS

T.L. is an advisor to Variant Bio with equity in Variant Bio. T.L. is a member of the advisory board of *Cell*.

REFERENCES

1. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.
2. Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. (2001). The sequence of the human genome. Science *291*, 1304–1351.
3. 100,000 Genomes Project Pilot Investigators, Smedley, D., Smith, K.R., Martin, A., Thomas, E.A., McDonagh, E.M., Cipriani, V., Ellingford, J.M., Arno, G., Tucci, A., et al. (2021). 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care – Preliminary Report. N. Engl. J. Med. *385*, 1868–1880.
4. Wright, C.F., Campbell, P., Eberhardt, R.Y., Aitken, S., Perrett, D., Brent, S., Danecek, P., Gardner, E.J., Chundru, V.K., Lindsay, S.J., et al. (2023). Genomic Diagnosis of Rare Pediatric Disease in the United Kingdom and Ireland. N. Engl. J. Med. *388*, 1559–1571.
5. Linder, J.E., Allworth, A., Bland, H.T., Caraballo, P.J., Chisholm, R.L., Clayton, E.W., Crosslin, D.R., Dikilitas, O., DiVietro, A., Esplin, E.D., et al. (2023). Returning integrated genomic risk and clinical recommendations: The eMERGE study. Genet. Med. *25*, 100006.
6. Trajanoska, K., Bhérer, C., Taliun, D., Zhou, S., Richards, J.B., and Mooser, V. (2023). From target discovery to clinical drug development with human genetics. Nature *620*, 737–745.
7. Minikel, E.V., Painter, J.L., Dong, C.C., and Nelson, M.R. (2023). Refining the Impact of Genetic Evidence on Clinical Success (Pharmacology and Therapeutics) https://doi.org/10.1101/2023.06.23.23291765.
8. Nelson, M.R., Tipney, H., Painter, J.L., Shen, J., Nicoletti, P., Shen, Y., Floratos, A., Sham, P.C., Li, M.J., Wang, J., et al. (2015). The support of human genetic evidence for approved drug indications. Nat. Genet. *47*, 856–860.
9. Lewontin, R.C. (1972). The Apportionment of Human Diversity. In Evolutionary Biology, T. Dobzhansky, M.K. Hecht, and W.C. Steere, eds. (Springer), pp. 381–398.
10. Coop, G. (2022). Genetic similarity versus genetic ancestry groups as sample descriptors in human genetics https://doi.org/10.48550/arXiv.2207.11595.
11. Lewis, A.C.F., Molina, S.J., Appelbaum, P.S., Dauda, B., Di Rienzo, A., Fuentes, A., Fullerton, S.M., Garrison, N.A., Ghosh, N., Hammonds, E.M., et al. (2022). Getting genetic ancestry right for science and society. Science *376*, 250–252.
12. Committee on the Use of Race; Ethnicity, and Ancestry as Population Descriptors in Genomics Research; Board on Health Sciences Policy; Committee on Population; Health and Medicine Division; Division of Behavioral and Social Sciences and Education; National Academies of Sciences, Engineering, and Medicine (2023). Using Population Descriptors in Genetics and Genomics Research: A New Framework for an Evolving Field (National Academies Press).
13. Green, E.D., Gunter, C., Biesecker, L.G., Di Francesco, V., Easter, C.L., Feingold, E.A., Felsenfeld, A.L., Kaufman, D.J., Ostrander, E.A., Pavan, W.J., et al. (2020). Strategic vision for improving human health at The Forefront of Genomics. Nature *586*, 683–692.
14. Hubby, J.L., and Lewontin, R.C. (1966). A molecular approach to the study of genic heterozygosity in natural populations. I. The number of alleles at different loci in Drosophila pseudoobscura. Genetics *54*, 577–594.
15. Bergström, A., McCarthy, S.A., Hui, R., Almarri, M.A., Ayub, Q., Danecek, P., Chen, Y., Felkel, S., Hallast, P., Kamm, J., et al. (2020). Insights into

human genetic variation and population history from 929 diverse genomes. Science *367*, eaay5012.

16. International; HapMap Consortium, Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., et al. (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851–861.
17. Greely, H.T. (2001). Human genome diversity: what about the other human genome project? Nat. Rev. Genet. *2*, 222–227.
18. 1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., et al. (2015). A global reference for human genetic variation. Nature *526*, 68–74.
19. Byrska-Bishop, M., Evani, U.S., Zhao, X., Basile, A.O., Abel, H.J., Regier, A.A., Corvelo, A., Clarke, W.E., Musunuri, R., Nagulapalli, K., et al. (2022). High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. Cell *185*, 3426–3440.e19.
20. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209.
21. Kurki, M.I., Karjalainen, J., Palta, P., Sipilä, T.P., Kristiansson, K., Donner, K.M., Reeve, M.P., Laivuori, H., Aavikko, M., Kaunisto, M.A., et al. (2023). FinnGen provides genetic insights from a well-phenotyped isolated population. Nature *613*, 508–518.
22. Martin, A.R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B.M., and Daly, M.J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. Nat. Genet. *51*, 584–591.
23. Henn, B.M., Gignoux, C.R., Jobin, M., Granka, J.M., Macpherson, J.M., Kidd, J.M., Rodríguez-Botigué, L., Ramachandran, S., Hon, L., Brisbin, A., et al. (2011). Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. Proc. Natl. Acad. Sci. USA *108*, 5154–5162.
24. SenGupta, D., Choudhury, A., Fortes-Lima, C., Aron, S., Whitelaw, G., Bostoen, K., Gunnink, H., Chousou-Polydouri, N., Delius, P., Tollman, S., et al. (2021). Genetic substructure and complex demographic history of South African Bantu speakers. Nat. Commun. *12*, 2080.
25. Fan, S., Spence, J.P., Feng, Y., Hansen, M.E.B., Terhorst, J., Beltrame, M.H., Ranciaro, A., Hirbo, J., Beggs, W., Thomas, N., et al. (2023). Whole-genome sequencing reveals a complex African population demographic history and signatures of local adaptation. Cell *186*, 923–939.e14.
26. Atkinson, E.G., Dalvie, S., Pichkar, Y., Kalungi, A., Majara, L., Stevenson, A., Abebe, T., Akena, D., Alemayehu, M., Ashaba, F.K., et al. (2022). Genetic structure correlates with ethnolinguistic diversity in eastern and southern Africa. Am. J. Hum. Genet. *109*, 1667–1679.
27. Ben-Eghan, C., Sun, R., Hleap, J.S., Diaz-Papkovich, A., Munter, H.M., Grant, A.V., Dupras, C., and Gravel, S. (2020). Don't ignore genetic data from minority populations. Nature *585*, 184–186.
28. Smith, S.P., Shahamatdar, S., Cheng, W., Zhang, S., Paik, J., Graff, M., Haiman, C., Matise, T.C., North, K.E., Peters, U., et al. (2022). Enrichment analyses identify shared associations for 25 quantitative traits in over 600,000 individuals from seven diverse ancestries. Am. J. Hum. Genet. *109*, 871–884.
29. Ragsdale, A.P., Weaver, T.D., Atkinson, E.G., Hoal, E.G., Möller, M., Henn, B.M., and Gravel, S. (2023). A weakly structured stem for human origins in Africa. Nature *617*, 755–763.
30. Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A.R., Auton, A., Indap, A., King, K.S., Bergmann, S., Nelson, M.R., et al. (2008). Genes mirror geography within Europe. Nature *456*, 98–101.
31. Wang, C., Zöllner, S., and Rosenberg, N.A. (2012). A quantitative comparison of the similarity between genes and geography in worldwide human populations. PLoS Genet. *8*, e1002886.
32. Moreno-Estrada, A., Gravel, S., Zakharia, F., McCauley, J.L., Byrnes, J.K., Gignoux, C.R., Ortiz-Tello, P.A., Martínez, R.J., Hedges, D.J., Morris, R.W., et al. (2013). Reconstructing the population genetic history of the Caribbean. PLoS Genet. *9*, e1003925.
33. Goldberg, A., Rastogi, A., and Rosenberg, N.A. (2020). Assortative mating by population of origin in a mechanistic model of admixture. Theor. Popul. Biol. *134*, 129–146.
34. Baharian, S., Barakatt, M., Gignoux, C.R., Shringarpure, S., Errington, J., Blot, W.J., Bustamante, C.D., Kenny, E.E., Williams, S.M., Aldrich, M.C., et al. (2016). The Great Migration and African-American Genomic Diversity. PLoS Genet. *12*, e1006059.
35. Biddanda, A., Rice, D.P., and Novembre, J. (2020). A variant-centric perspective on geographic patterns of human allele frequency variation. eLife *9*, e60107.
36. Koenig, Z., Yohannes, M.T., Nkambule, L.L., Goodrich, J.K., Kim, H.A., Zhao, X., Wilson, M.W., Tiao, G., Hao, S.P., Sahakian, N., et al. (2023). A harmonized public resource of deeply sequenced diverse human genomes https://doi.org/10.1101/2023.01.23.525248.
37. Duncan, L.E., Ostacher, M., and Ballon, J. (2019). How genome-wide association studies (GWAS) made traditional candidate gene studies obsolete. Neuropsychopharmacology *44*, 1518–1523.
38. Eichler, E.E., Flint, J., Gibson, G., Kong, A., Leal, S.M., Moore, J.H., and Nadeau, J.H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. Nat. Rev. Genet. *11*, 446–450.
39. Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. Am. J. Hum. Genet. *101*, 5–22.
40. McClellan, J., and King, M.-C. (2010). Genetic heterogeneity in human disease. Cell *141*, 210–217.
41. Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. *42*, 565–569.
42. Young, A.I., Benonisdottir, S., Przeworski, M., and Kong, A. (2019). Deconstructing the sources of genotype-phenotype associations in humans. Science *365*, 1396–1400.
43. Howe, L.J., Nivard, M.G., Morris, T.T., Hansen, A.F., Rasheed, H., Cho, Y., Chittoor, G., Ahlskog, R., Lind, P.A., Palviainen, T., et al. (2022). Within-sibship genome-wide association analyses decrease bias in estimates of direct genetic effects. Nat. Genet. *54*, 581–592.
44. Van Hout, C.V., Tachmazidou, I., Backman, J.D., Hoffman, J.D., Liu, D., Pandey, A.K., Gonzaga-Jauregui, C., Khalid, S., Ye, B., Banerjee, N., et al. (2020). Exome sequencing and characterization of 49,960 individuals in the UK Biobank. Nature *586*, 749–756.
45. Backman, J.D., Li, A.H., Marcketta, A., Sun, D., Mbatchou, J., Kessler, M.D., Benner, C., Liu, D., Locke, A.E., Balasubramanian, S., et al. (2021). Exome sequencing and analysis of 454,787 UK Biobank participants. Nature *599*, 628–634.
46. Purcell, S. (2002). Variance components models for gene-environment interaction in twin analysis. Twin Res. *5*, 554–571.
47. Tenesa, A., and Haley, C.S. (2013). The heritability of human disease: estimation, uses and abuses. Nat. Rev. Genet. *14*, 139–149.
48. Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. PLoS Genet. *9*, e1003348.
49. Bulik-Sullivan, B.K., Loh, P.-R., Finucane, H.K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric; Genomics Consortium, Patterson, N., Daly, M.J., Price, A.L., and Neale, B.M. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet. *47*, 291–295.
50. Ge, T., Chen, C.-Y., Neale, B.M., Sabuncu, M.R., and Smoller, J.W. (2017). Phenome-wide heritability analysis of the UK Biobank. PLoS Genet. *13*, e1006711.

51. Kanai, M., Akiyama, M., Takahashi, A., Matoba, N., Momozawa, Y., Ikeda, M., Iwata, N., Ikegawa, S., Hirata, M., Matsuda, K., et al. (2018). Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. Nat. Genet. *50*, 390–400.

52. Zhang, Y., Qi, G., Park, J.-H., and Chatterjee, N. (2018). Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. Nat. Genet. *50*, 1318–1326.

53. O'Connor, L.J., Schoech, A.P., Hormozdiari, F., Gazal, S., Patterson, N., and Price, A.L. (2019). Extreme Polygenicity of Complex Traits Is Explained by Negative Selection. Am. J. Hum. Genet. *105*, 456–476.

54. Weissbrod, O., Hormozdiari, F., Benner, C., Cui, R., Ulirsch, J., Gazal, S., Schoech, A.P., van de Geijn, B., Reshef, Y., Márquez-Luna, C., et al. (2020). Functionally informed fine-mapping and polygenic localization of complex trait heritability. Nat. Genet. *52*, 1355–1363.

55. Simons, Y.B., Bullaughey, K., Hudson, R.R., and Sella, G. (2018). A population genetic interpretation of GWAS findings for human quantitative traits. PLoS Biol. *16*, e2002985.

56. Schoech, A.P., Jordan, D.M., Loh, P.-R., Gazal, S., O'Connor, L.J., Balick, D.J., Palamara, P.F., Finucane, H.K., Sunyaev, S.R., and Price, A.L. (2019). Quantification of frequency-dependent genetic architectures in 25 UK Biobank traits reveals action of negative selection. Nat. Commun. *10*, 790.

57. Yengo, L., Vedantam, S., Marouli, E., Sidorenko, J., Bartell, E., Sakaue, S., Graff, M., Eliasen, A.U., Jiang, Y., Raghavan, S., et al. (2022). A saturated map of common genetic variants associated with human height. Nature *610*, 704–712.

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

59. Wray, N.R., Wijmenga, C., Sullivan, P.F., Yang, J., and Visscher, P.M. (2018). Common Disease Is More Complex Than Implied by the Core Gene Omnigenic Model. Cell *173*, 1573–1580.

60. Liu, X., Li, Y.I., and Pritchard, J.K. (2019). Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. Cell *177*, 1022–1034.e6.

61. Gazal, S., Loh, P.R., Finucane, H.K., Ganna, A., Schoech, A., Sunyaev, S., and Price, A.L. (2018). Functional architecture of low-frequency variants highlights strength of negative selection across coding and non-coding annotations. Nat. Genet. *50*, 1600–1607.

62. Wainschtein, P., Jain, D., Zheng, Z., Cupples, L.A., Shadyab, A.H., McKnight, B., Shoemaker, B.M., Mitchell, B.D., et al.; TOPMed Anthropometry Working Group; NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium (2022). Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. Nat. Genet. *54*, 263–273.

63. Weiner, D.J., Nadig, A., Jagadeesh, K.A., Dey, K.K., Neale, B.M., Robinson, E.B., Karczewski, K.J., and O'Connor, L.J. (2023). Polygenic architecture of rare coding variation across 394,783 exomes. Nature *614*, 492–499.

64. Rajagopal, V.M., Watanabe, K., Mbatchou, J., Ayer, A., Quon, P., Sharma, D., Kessler, M.D., Praveen, K., Gelfman, S., Parikshak, N., et al. (2023). Rare coding variants in CHRNB2 reduce the likelihood of smoking. Nat. Genet. *55*, 1138–1148.

65. Martin, A.R., Gignoux, C.R., Walters, R.K., Wojcik, G.L., Neale, B.M., Gravel, S., Daly, M.J., Bustamante, C.D., and Kenny, E.E. (2017). Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. Am. J. Hum. Genet. *100*, 635–649.

66. Carlson, C.S., Matise, T.C., North, K.E., Haiman, C.A., Fesinmeyer, M.D., Buyske, S., Schumacher, F.R., Peters, U., Franceschini, N., Ritchie, M.D., et al. (2013). Generalization and dilution of association results from European GWAS in populations of non-European ancestry: the PAGE study. PLoS Biol. *11*, e1001661.

67. Ding, Y., Hou, K., Xu, Z., Pimplaskar, A., Petter, E., Boulier, K., Privé, F., Vilhjálmsson, B.J., Olde Loohuis, L.M., and Pasaniuc, B. (2023). Polygenic scoring accuracy varies across the genetic ancestry continuum. Nature *618*, 774–781.

68. Hou, K., Ding, Y., Xu, Z., Wu, Y., Bhattacharya, A., Mester, R., Belbin, G.M., Buyske, S., Conti, D.V., Darst, B.F., et al. (2023). Causal effects on complex traits are similar for common variants across segments of different continental ancestries within admixed individuals. Nat. Genet. *55*, 549–558.

69. Wang, Y., Guo, J., Ni, G., Yang, J., Visscher, P.M., and Yengo, L. (2020). Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. Nat. Commun. *11*, 3865.

70. Patel, R.A., Musharoff, S.A., Spence, J.P., Pimentel, H., Tcheandjieu, C., Mostafavi, H., Sinnott-Armstrong, N., Clarke, S.L., Smith, C.J., V.A., et al.. (2022). Genetic interactions drive heterogeneity in causal variant effect sizes for gene expression and complex traits. Am. J. Hum. Genet. *109*, 1286–1297.

71. Brown, B.C., Ye, C.J., Price, A.L., and Zaitlen, N.; Asian Genetic Epidemiology Network Type 2 Diabetes Consortium (2016). Transethnic Genetic-Correlation Estimates from Summary Statistics. Am. J. Hum. Genet. *99*, 76–88.

72. Wojcik, G.L., Graff, M., Nishimura, K.K., Tao, R., Haessler, J., Gignoux, C.R., Highland, H.M., Patel, Y.M., Sorokin, E.P., Avery, C.L., et al. (2019). Genetic analyses of diverse populations improves discovery for complex traits. Nature *570*, 514–518.

73. Johnson, R., Ding, Y., Venkateswaran, V., Bhattacharya, A., Boulier, K., Chiu, A., Knyazev, S., Schwarz, T., Freund, M., Zhan, L., et al. (2022). Leveraging genomic diversity for discovery in an electronic health record linked biobank: the UCLA ATLAS Community Health Initiative. Genome Med. *14*, 104.

74. Amariuta, T., Ishigaki, K., Sugishita, H., Ohta, T., Koido, M., Dey, K.K., Matsuda, K., Murakami, Y., Price, A.L., Kawakami, E., et al. (2020). Improving the trans-ancestry portability of polygenic risk scores by prioritizing variants in predicted cell-type-specific regulatory elements. Nat. Genet. *52*, 1346–1354.

75. Weissbrod, O., Kanai, M., Shi, H., Gazal, S., Peyrot, W.J., Khera, A.V., Okada, Y., Biobank; Japan Project, Martin, A.R., Finucane, H.K., et al. (2022). Leveraging fine-mapping and multipopulation training data to improve cross-population polygenic risk scores. Nat. Genet. *54*, 450–458.

76. Chen, F., Wang, X., Jang, S.K., Quach, B.C., Weissenkampen, J.D., Khunsriraksakul, C., Yang, L., Sauteraud, R., Albert, C.M., Allred, N.D.D., et al. (2023). Multi-ancestry transcriptome-wide association analyses yield insights into tobacco use biology and drug repurposing. Nat. Genet. *55*, 291–300.

77. Lu, Z., Gopalan, S., Yuan, D., Conti, D.V., Pasaniuc, B., Gusev, A., and Mancuso, N. (2022). Multi-ancestry fine-mapping improves precision to identify causal genes in transcriptome-wide association studies. Am. J. Hum. Genet. *109*, 1388–1404.

78. Bitarello, B.D., and Mathieson, I. (2020). Polygenic Scores for Height in Admixed Populations. G3 (Bethesda) *10*, 4027–4036.

79. Marnetto, D., Pärna, K., Läll, K., Molinaro, L., Montinaro, F., Haller, T., Metspalu, M., Mägi, R., Fischer, K., and Pagani, L. (2020). Ancestry deconvolution and partial polygenic score can improve susceptibility predictions in recently admixed individuals. Nat. Commun. *11*, 1628.

80. Mostafavi, H., Harpak, A., Agarwal, I., Conley, D., Pritchard, J.K., and Przeworski, M. (2020). Variable prediction accuracy of polygenic scores within an ancestry group. eLife *9*, e48376.

81. Border, R., Athanasiadis, G., Buil, A., Schork, A.J., Cai, N., Young, A.I., Werge, T., Flint, J., Kendler, K.S., Sankararaman, S., et al. (2022). Cross-trait assortative mating is widespread and inflates genetic correlation estimates. Science *378*, 754–761.

82. Diaz-Papkovich, A., Zabad, S., Ben-Eghan, C., Anderson-Trocmé, L., Femerling, G., Nathan, V., Patel, J., and Gravel, S. (2023). Topological

stratification of continuous genetic variation in large biobanks (Genomics) https://doi.org/10.1101/2023.07.06.548007.

83. Gorla, A., Sankararaman, S., Burchard, E., Flint, J., Zaitlen, N., and Rahmani, E. (2023). Phenotypic subtyping via contrastive learning https://doi.org/10.1101/2023.01.05.522921.

84. Kong, A., Thorleifsson, G., Frigge, M.L., Vilhjalmsson, B.J., Young, A.I., Thorgeirsson, T.E., Benonisdottir, S., Oddsson, A., Halldorsson, B.V., Masson, G., et al. (2018). The nature of nurture: Effects of parental genotypes. Science *359*, 424–428.

85. Veller, C., and Coop, G. (2023). Interpreting population and family-based genome-wide association studies in the presence of confounding https://doi.org/10.1101/2023.02.26.530052.

86. Simons, Y.B., Mostafavi, H., Smith, C.J., Pritchard, J.K., and Sella, G. (2022). Simple scaling laws control the genetic architectures of human complex traits (Genetics) https://doi.org/10.1101/2022.10.04.509926.

87. Yair, S., and Coop, G. (2022). Population differentiation of polygenic score predictions under stabilizing selection. Philos. Trans. R. Soc. Lond. B Biol. Sci. *377*, 20200416.

88. Durvasula, A., and Lohmueller, K.E. (2021). Negative selection on complex traits limits phenotype prediction accuracy between populations. Am. J. Hum. Genet. *108*, 620–631.

89. Berg, J.J., Harpak, A., Sinnott-Armstrong, N., Joergensen, A.M., Mostafavi, H., Field, Y., Boyle, E.A., Zhang, X., Racimo, F., Pritchard, J.K., et al. (2019). Reduced signal for polygenic adaptation of height in UK Biobank. eLife *8*, e39725.

90. Sohail, M., Maier, R.M., Ganna, A., Bloemendal, A., Martin, A.R., Turchin, M.C., Chiang, C.W., Hirschhorn, J., Daly, M.J., Patterson, N., et al. (2019). Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. eLife *8*, e39702.

91. Zhang, B.C., Biddanda, A., Gunnarsson, Á.F., Cooper, F., and Palamara, P.F. (2023). Biobank-scale inference of ancestral recombination graphs enables genealogical analysis of complex traits. Nat. Genet. *55*, 768–776.

92. Hujoel, M.L.A., Sherman, M.A., Barton, A.R., Mukamel, R.E., Sankaran, V.G., Terao, C., and Loh, P.-R. (2022). Influences of rare copy-number variation on human complex traits. Cell *185*, 4233–4248.e27.

93. Popic, V., Rohlicek, C., Cunial, F., Hajirasouliha, I., Meleshko, D., Garimella, K., and Maheshwari, A. (2023). Cue: a deep-learning framework for structural variant discovery and genotyping. Nat. Methods *20*, 559–568.

94. Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A.V., Mikheenko, A., Vollger, M.R., Altemose, N., Uralsky, L., Gershman, A., et al. (2022). The complete sequence of a human genome. Science *376*, 44–53.

95. Heyne, H.O., Karjalainen, J., Karczewski, K.J., Lemmelä, S.M., Zhou, W., FinnGen, Havulinna, A.S., Kurki, M., Rehm, H.L., Palotie, A., et al. (2023). Mono- and biallelic variant effects on disease at biobank scale. Nature *613*, 519–525.

96. Albert, F.W., Bloom, J.S., Siegel, J., Day, L., and Kruglyak, L. (2018). Genetics of trans-regulatory variation in gene expression. eLife *7*, e35471.

97. Hemani, G., Shakhbazov, K., Westra, H.-J., Esko, T., Henders, A.K., McRae, A.F., Yang, J., Gibson, G., Martin, N.G., Metspalu, A., et al. (2021). Retraction Note: Detection and replication of epistasis influencing transcription in humans. Nature *596*, 306.

98. Smith, S.P., Darnell, G., Udwin, D., Harpak, A., Ramachandran, S., and Crawford, L. (2022). Accounting for statistical non-additive interactions enables the recovery of missing heritability from GWAS summary statistics https://doi.org/10.1101/2022.07.21.501001.

99. Norman, T.M., Horlbeck, M.A., Replogle, J.M., Ge, A.Y., Xu, A., Jost, M., Gilbert, L.A., and Weissman, J.S. (2019). Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. Science *365*, 786–793.

100. van der Wijst, M., de Vries, D.H., Groot, H.E., Trynka, G., Hon, C.C., Bonder, M.J., Stegle, O., Nawijn, M.C., Idaghdour, Y., van der Harst, P., et al. (2020). The single-cell eQTLGen consortium. eLife *9*, e52155.

101. Võsa, U., Claringbould, A., Westra, H.-J., Bonder, M.J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Yazar, S., et al. (2021). Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nat. Genet. *53*, 1300–1310.

102. Aguet, F., Alasoo, K., Li, Y.I., Battle, A., Im, H.K., Montgomery, S.B., and Lappalainen, T. (2023). Molecular quantitative trait loci. Nat. Rev. Methods Primers *3*, 4.

103. Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J.F., Darbandi, S.F., Knowles, D., Li, Y.I., Kosmicki, J.A., Arbelaez, J., Cui, W., Schwartz, G.B., et al. (2019). Predicting Splicing from Primary Sequence with Deep Learning. Cell *176*, 535–548.e24.

104. Frazer, J., Notin, P., Dias, M., Gomez, A., Min, J.K., Brock, K., Gal, Y., and Marks, D.S. (2021). Disease variant prediction with deep generative models of evolutionary data. Nature *599*, 91–95.

105. Zeng, T., and Li, Y.I. (2022). Predicting RNA splicing from DNA sequence using Pangolin. Genome Biol. *23*, 103.

106. Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J.R., Grabska-Barwinska, A., Taylor, K.R., Assael, Y., Jumper, J., Kohli, P., and Kelley, D.R. (2021). Effective gene expression prediction from sequence by integrating long-range interactions. Nat. Methods *18*, 1196–1203.

107. Sasse, A., Ng, B., Spiro, A.E., Tasaki, S., Bennett, D.A., Gaiteri, C., De Jager, P.L., Chikina, M., and Mostafavi, S. (2023). Benchmarking of deep neural networks for predicting personal gene expression from DNA sequence highlights shortcomings https://doi.org/10.1101/2023.03.16.532969.

108. Nicolae, D.L., Gamazon, E., Zhang, W., Duan, S., Dolan, M.E., and Cox, N.J. (2010). Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet. *6*, e1000888.

109. Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science *337*, 1190–1195.

110. Iotchkova, V., Ritchie, G.R.S., Geihs, M., Morganella, S., Min, J.L., Walter, K., Timpson, N.J., Dunham, I., Birney, E., et al.; UK10K Consortium (2019). GARFIELD classifies disease-relevant genomic features through integration of functional annotations with association signals. Nat. Genet. *51*, 343–353.

111. Finucane, H.K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H., Zang, C., Farh, K., et al. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet. *47*, 1228–1235.

112. Farh, K.K., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W.J., Beik, S., Shoresh, N., Whitton, H., Ryan, R.J., Shishkin, A.A., et al. (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature *518*, 337–343.

113. Gusev, A., Lee, S.H., Trynka, G., Finucane, H., Vilhjálmsson, B.J., Xu, H., Zang, C., Ripke, S., Bulik-Sullivan, B., Stahl, E., et al. (2014). Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. Am. J. Hum. Genet. *95*, 535–552.

114. Banovich, N.E., Lan, X., McVicker, G., van de Geijn, B., Degner, J.F., Blischak, J.D., Roux, J., Pritchard, J.K., and Gilad, Y. (2014). Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. PLoS Genet. *10*, e1004663.

115. Chen, L., Ge, B., Casale, F.P., Vasquez, L., Kwan, T., Garrido-Martín, D., Watt, S., Yan, Y., Kundu, K., Ecker, S., et al. (2016). Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. Cell *167*, 1398–1414.e24.

116. Kumasaka, N., Knights, A.J., and Gaffney, D.J. (2016). Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. Nat. Genet. *48*, 206–213.

117. Hormozdiari, F., Gazal, S., van de Geijn, B., Finucane, H.K., Ju, C.J.-T., Loh, P.-R., Schoech, A., Reshef, Y., Liu, X., O'Connor, L., et al. (2018). Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits. Nat. Genet. *50*, 1041–1047.

118. Mitchell, J.M., Nemesh, J., Ghosh, S., Handsaker, R.E., Mello, C.J., Meyer, D., Raghunathan, K., De Rivera, H., Tegtmeyer, M., Hawes, D., et al. (2020). Mapping genetic effects on cellular phenotypes with "Cell Villages" https://doi.org/10.1101/2020.06.29.174383.

119. Lappalainen, T., Sammeth, M., Friedländer, M.R., 't Hoen, P.A., Monlong, J., Rivas, M.A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501*, 506–511.

120. Li, Y.I., van de Geijn, B., Raj, A., Knowles, D.A., Petti, A.A., Golan, D., Gilad, Y., and Pritchard, J.K. (2016). RNA splicing is a primary link between genetic variation and disease. Science *352*, 600–604.

121. Alasoo, K., Rodrigues, J., Danesh, J., Freitag, D.F., Paul, D.S., and Gaffney, D.J. (2019). Genetic effects on promoter usage are highly context-specific and contribute to complex traits. eLife *8*, e41673.

122. Li, Q., Gloudemans, M.J., Geisinger, J.M., Fan, B., Aguet, F., Sun, T., Ramaswami, G., Li, Y.I., Ma, J.-B., Pritchard, J.K., et al. (2022). RNA editing underlies genetic risk of common inflammatory diseases. Nature *608*, 569–577.

123. Qi, T., Wu, Y., Fang, H., Zhang, F., Liu, S., Zeng, J., and Yang, J. (2022). Genetic control of RNA splicing and its distinct role in complex trait variation. Nat. Genet. *54*, 1355–1363.

124. Finucane, H.K., Reshef, Y.A., Anttila, V., Slowikowski, K., Gusev, A., Byrnes, A., Gazal, S., Loh, P.-R., Lareau, C., Shoresh, N., et al. (2018). Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. Nat. Genet. *50*, 621–629.

125. Albert, F.W., and Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. Nat. Rev. Genet. *16*, 197–212.

126. GTEx Consortium, Laboratory, Data Analysis & Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, NIH/NIMH, NIH/NIDA, Biospecimen Collection Source Site—NDRI, et al. (2017) Genetic effects on gene expression across human tissues. Nature, 550, 204–213.

127. Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace, C., and Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. *10*, e1004383.

128. Wallace, C. (2021). A more accurate method for colocalisation analysis allowing for multiple causal variants. PLoS Genet. *17*, e1009440.

129. Gamazon, E.R., Wheeler, H.E., Shah, K.P., Mozaffari, S.V., Aquino-Michaels, K., Carroll, R.J., Eyler, A.E., Denny, J.C., GTEx Consortium, and Nicolae, D.L., et al. (2015). A gene-based association method for mapping traits using reference transcriptome data. Nat. Genet. *47*, 1091–1098.

130. Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B.W.J.H., Jansen, R., de Geus, E.J.C., Boomsma, D.I., Wright, F.A., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. Nat. Genet. *48*, 245–252.

131. Barbeira, A.N., Dickinson, S.P., Bonazzola, R., Zheng, J., Wheeler, H.E., Torres, J.M., Torstenson, E.S., Shah, K.P., Garcia, T., Edwards, T.L., et al. (2018). Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nat. Commun. *9*, 1825.

132. Hukku, A., Sampson, M.G., Luca, F., Pique-Regi, R., and Wen, X. (2022). Analyzing and reconciling colocalization and transcriptome-wide association studies from the perspective of inferential reproducibility. Am. J. Hum. Genet. *109*, 825–837.

133. Wainberg, M., Sinnott-Armstrong, N., Mancuso, N., Barbeira, A.N., Knowles, D.A., Golan, D., Ermel, R., Ruusalepp, A., Quertermous, T., Hao, K., et al. (2019). Opportunities and challenges for transcriptome-wide association studies. Nat. Genet. *51*, 592–599.

134. Umans, B.D., Battle, A., and Gilad, Y. (2021). Where Are the Disease-Associated eQTLs? Trends Genet. *37*, 109–124.

135. Yao, D.W., O'Connor, L.J., Price, A.L., and Gusev, A. (2020). Quantifying genetic effects on disease mediated by assayed gene expression levels. Nat. Genet. *52*, 626–633.

136. Chun, S., Casparino, A., Patsopoulos, N.A., Croteau-Chonka, D.C., Raby, B.A., De Jager, P.L., Sunyaev, S.R., and Cotsapas, C. (2017). Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. Nat. Genet. *49*, 600–605.

137. Mu, Z., Wei, W., Fair, B., Miao, J., Zhu, P., and Li, Y.I. (2021). The impact of cell type and context-dependent regulatory variants on human immune traits. Genome Biol. *22*, 122.

138. Nasser, J., Bergman, D.T., Fulco, C.P., Guckelberger, P., Doughty, B.R., Patwardhan, T.A., Jones, T.R., Nguyen, T.H., Ulirsch, J.C., Lekschas, F., et al. (2021). Genome-wide enhancer maps link risk variants to disease genes. Nature *593*, 238–243.

139. Baca, S.C., Singler, C., Zacharia, S., Seo, J.-H., Morova, T., Hach, F., Ding, Y., Schwarz, T., Huang, C.-C.F., Anderson, J., et al. (2022). Genetic determinants of chromatin reveal prostate cancer risk mediated by context-dependent gene regulation. Nat. Genet. *54*, 1364–1375.

140. Aracena, K.A., Lin, Y.-L., Luo, K., Pacis, A., Gona, S., Mu, Z., Yotova, V., Sindeaux, R., Pramatarova, A., Simon, M.-M., et al. (2022). Epigenetic variation impacts ancestry-associated differences in the transcriptional response to influenza infection https://doi.org/10.1101/2022.05.10.491413.

141. Gusev, A., Mancuso, N., Won, H., Kousi, M., Finucane, H.K., Reshef, Y., Song, L., Safi, A., Schizophrenia Working Group of the Psychiatric; Genomics Consortium, and McCarroll, S., et al. (2018). Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. Nat. Genet. *50*, 538–548.

142. Banovich, N.E., Li, Y.I., Raj, A., Ward, M.C., Greenside, P., Calderon, D., Tung, P.Y., Burnett, J.E., Myrthil, M., Thomas, S.M., et al. (2018). Impact of regulatory variation across human iPSCs and differentiated cells. Genome Res. *28*, 122–131.

143. Alasoo, K., Rodrigues, J., Mukhopadhyay, S., Knights, A.J., Mann, A.L., Kundu, K., HIPSCI Consortium, Hale, C., Dougan, G., and Gaffney, D.J. (2018). Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. Nat. Genet. *50*, 424–431.

144. Wang, X., and Goldstein, D.B. (2020). Enhancer Domains Predict Gene Pathogenicity and Inform Gene Discovery in Complex Disease. Am. J. Hum. Genet. *106*, 215–233.

145. Mostafavi, H., Spence, J.P., Naqvi, S., and Pritchard, J.K. (2022). Limited overlap of eQTLs and GWAS hits due to systematic differences in discovery https://doi.org/10.1101/2022.05.07.491045.

146. Strober, B.J., Elorbany, R., Rhodes, K., Krishnan, N., Tayeb, K., Battle, A., and Gilad, Y. (2019). Dynamic genetic regulation of gene expression during cellular differentiation. Science *364*, 1287–1290.

147. Mountjoy, E., Schmidt, E.M., Carmona, M., Schwartzentruber, J., Peat, G., Miranda, A., Fumis, L., Hayhurst, J., Buniello, A., Karim, M.A., et al. (2021). An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. Nat. Genet. *53*, 1527–1533.

148. Hsu, Y.-H.H., Pintacuda, G., Liu, R., Nacu, E., Kim, A., Tsafou, K., Petrossian, N., Crotty, W., Suh, J.M., Riseman, J., et al. (2023). Using brain cell-type-specific protein interactomes to interpret neurodevelopmental genetic signals in schizophrenia. iScience *26*, 106701.

149. Morris, J.A., Caragine, C., Daniloski, Z., Domingo, J., Barry, T., Lu, L., Davis, K., Ziosi, M., Glinos, D.A., Hao, S., et al. (2023). Discovery of target genes and pathways at GWAS loci by pooled single-cell CRISPR screens. Science *380*, eadh7699.

150. Veller, C., Przeworski, M., and Coop, G. (2023). Causal interpretations of family GWAS in the presence of heterogeneous effects https://doi.org/10.1101/2023.11.13.566950.

151. Jackson, C.S., Turner, D., June, M., and Miller, M.V. (2023). Facing our History—Building an Equitable Future. Am. J. Hum. Genet. *110*, 377–395.

152. Carlson, J., Henn, B.M., Al-Hindi, D.R., and Ramachandran, S. (2022). Counter the weaponization of genetics research by extremists. Nature *610*, 444–447.