Annual Review of Biomedical Data Science

Analytic and Translational Genetics

Konrad J. Karczewski[1,2] and Alicia R. Martin[1,2]

[1]Program in Medical and Population Genetics and Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA; email: konradk@broadinstitute.org, armartin@broadinstitute.org

[2]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts 02114, USA

Annu. Rev. Biomed. Data Sci. 2020. 3:217–41

First published as a Review in Advance on April 29, 2020

The *Annual Review of Biomedical Data Science* is online at biodatasci.annualreviews.org

https://doi.org/10.1146/annurev-biodatasci-072018-021148

Keywords

population genetics, genome-wide association studies, medical genomics, variant deleteriousness, constraint, genetically guided drug discovery

Abstract

Understanding the influence of genetics on human disease is among the primary goals for biology and medicine. To this end, the direct study of natural human genetic variation has provided valuable insights into human physiology and disease as well as into the origins and migrations of humans. In this review, we discuss the foundations of population genetics, which provide a crucial context to the study of human genes and traits. In particular, genome-wide association studies and similar methods have revealed thousands of genetic loci associated with diseases and traits, providing invaluable information into the biology of these traits. Simultaneously, as the study of rare genetic variation has expanded, so-called human knockouts have elucidated the function of human genes and the therapeutic potential of targeting them.

INTRODUCTION

Human traits and diseases are the result of individual or combinatorial factors of genetics and the environment, the precise breakdown of which varies by disease. Some diseases such as lung cancer have primarily environmental contributions such as smoking. Some diseases such as cystic fibrosis are genetically mediated, although severity can vary with the environment and other genetic effects. Most diseases are intermediate, with some important genetic and environmental contributions (e.g., heart disease). Thus, understanding the influence of genetics on human disease is among the primary goals for biology and medicine.

Model organisms are often an ideal system for testing biological hypotheses because their environments can be highly regulated and genetic interventions are relatively simple. Trading off physiological complexity for similarity in genetic background, cell lines derived from human tissue can also be manipulated in relatively controlled environments, but in vitro systems may be insufficient to represent the biological complexity involved in disease. Given uncertainty in how precisely either model organisms or human cell lines reflect human biology, especially for higher-order functions, datasets of natural human variation can provide a direct lens into understanding human phenotypes when these approaches fall short.

These advantages notwithstanding, analyzing natural human variation datasets presents unique challenges spanning rare to common complex diseases that must be addressed in order to extract useful biological information. Because the human genome is shaped by human history, understanding evolutionary concepts from population genetics is critical for appropriate interpretation of genetic studies. Additionally, the number of individuals with a rare disease is inherently low, and thus a phenotype-first approach requires extensive domain knowledge to confidently associate genes with diseases when statistical power is low. Conversely, common diseases enable large-scale studies, but they are typically complex and variable, with many genes and pathways contributing weak effects alongside the environment that result in the disease state. This complexity means that large sample sizes are required to achieve statistical power for robust findings, leading to more expensive experiments. However, the success of genome-wide association studies (GWAS) in robustly finding and replicating thousands of associations for thousands of phenotypes has demonstrated the utility of these approaches and ushered in a new era of human genetics. These studies can home in on causal pathways and relevant cell types in an unbiased fashion, with the ultimate goals of understanding disease states and, ideally, of positive therapeutic outcomes.

POST–HUMAN GENOME PROJECT

Nearly 20 years ago, the Human Genome Project completed the first draft sequence of the human genome (1). This first genome provided a backbone for future studies of human genomes, including the location and base composition of genes. This gene map has since enabled numerous studies of gene functions and their associations to disease. Perhaps most importantly, this genome continues to serve as a reference genome that enables geneticists to speak a common language when analyzing any additional genomes, of which millions have been genotyped or sequenced over the ensuing decades (**Figure 1**). In this way, any variation can be characterized in a uniform manner.

Soon after the Human Genome Project, the dense sequencing of ten 500-kb (kilobase) segments in hundreds of individuals, performed as part of the International HapMap Consortium (2), identified a number of locations (loci) in the genome where a single base pair (bp) commonly varies among individuals, known as single-nucleotide polymorphisms (SNPs). Most variant sites are rare and typically private to an individual or population, but there are many common variants that are shared across populations. Nearby variants are typically inherited together and are

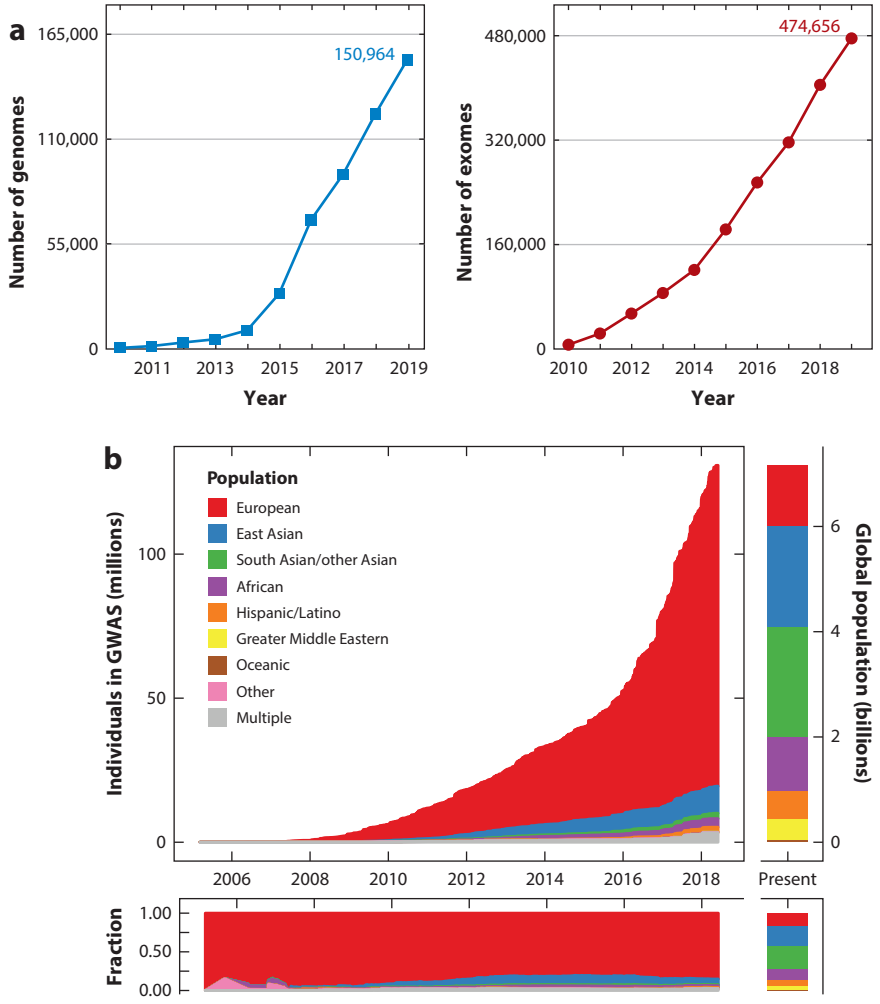

Figure 1

Growth in size of modern genetics datasets. (*a*) The number of samples whose whole genomes or exomes were sequenced over time through September 2019 by the Broad Institute. Panel *a* courtesy of Niall Lennon, Broad Genomics. (*b*) The increasing number (*top*) and fraction (*bottom*) of samples in genome-wide association studies as a function of ancestry compared to the global population. Panel *b* adapted with permission from Reference 58; copyright 2019 Springer Nature.

therefore often found in a correlated state known as linkage disequilibrium (LD), which leads to a block-like structure of haplotypes. Further, this sequencing has revealed the now well-established fact that any pair of individuals have remarkably similar genomes, with approximately 99.9% of sites in the genome being identical.

Later, the development of microarray technologies enabled the genotyping, or ascertainment of the genotypic state (homozygous reference, heterozygous, or homozygous alternate) of a known location in the genome, of hundreds of thousands to millions of genetic variants in an individual (3). Genotyping of hundreds of individuals spanning multiple populations provided insight into the level of haplotype sharing across individuals. Shared haplotypes paved the way for imputation,

the process of filling in genotypes, which enabled cost-effective GWAS. However, these genotyping technologies are limited to known loci and therefore are biased toward those that are variable in previously studied populations.

In the last decade, whole-genome sequencing (WGS) technologies have dramatically improved and become much cheaper, enabling the identification of nearly all the genetic variation specific to a given individual. In these technologies, genomic DNA from an individual is sheared and small fragments of DNA (currently about 150 bp on each end of the fragment) known as short reads are sequenced (4). Typically for a high-coverage genome, nearly one billion of these reads are sequenced so that the coverage, or the average number of times a base at each position in the genome is observed, of a sequencing run is about 30-fold. These fragments are aligned (mapped) to a reference genome, and differences from this reference are identified in a process known as variant calling. Over the years, methods for accurate variant calling have improved considerably, including those for single-nucleotide variants (SNVs), short insertions or deletions (indels), and larger variants including copy number variants and structural variants (SVs).

With these technologies, large sequencing projects have identified the scale of common and rare genetic variation in human populations. The 1000 Genomes Project performed low-coverage (4–8×) sequencing of 2,504 individuals, providing a high-resolution map of genetic variation. This catalog of variants spans the allele frequency spectrum, from those that are very common globally or within a population to those that are very rare, sometimes even newly arising in a single generation (de novo mutations). Assessments of these de novo mutations and phylogenetic comparisons with primates, our nearest ancestors, have enabled estimation of the human mutation rate (μ) to be approximately 10^{-8} per base per generation. Further, the project characterized detailed patterns of variation around genes, including depletions of variation in and near exons and transcription start sites. Data from the 1000 Genomes Project as well as from the Human Genome Diversity Panel have now been updated to include high-coverage (30×) sequence data, which are freely available to researchers (5, 6).

A targeted version of this technology known as whole-exome sequencing (WES), which enriches for protein-coding regions of the genome, is several-fold cheaper. This technology and cost reductions of WGS have enabled still larger projects (**Figure 1*a***), including 6,503 individuals from the Exome Sequencing Project (7) and 60,706 individuals from the Exome Aggregation Consortium (ExAC) (8). This cohort and the Genome Aggregation Database (gnomAD), which includes 141,456 individuals with exome and genome sequence data, are metacohorts that include individuals from several common disease cohorts and population studies (9).

POPULATION GENETICS

Evolutionary Forces Governing Genetic Variation

In each generation, the human genome accumulates genetic variation in the form of de novo mutations, which are new mutations that arise in offspring and are not present in the parents. On average, de novo mutations are found at a rate of approximately 65 SNVs and 5 indels (10), as well as 0.35 SVs (11), per individual. Given these mutation rates [ranging from 10^{-9} to 10^{-7} across genomic contexts (9)] and a population size of 7.5×10^9, it is likely that every possible SNV that is compatible with life exists in at least one individual. Most variants are noncoding and are benign to the development and life of the individual. As such, these variants can be passed on to further generations, and may stochastically rise or fall in frequency over multiple generations in a process known as genetic drift. The likelihood that a de novo variant remains in the population is inversely proportional to the effective population size (N_e): the number of individuals in a theoretically ideal population with the same level of genetic drift as the actual population (12).

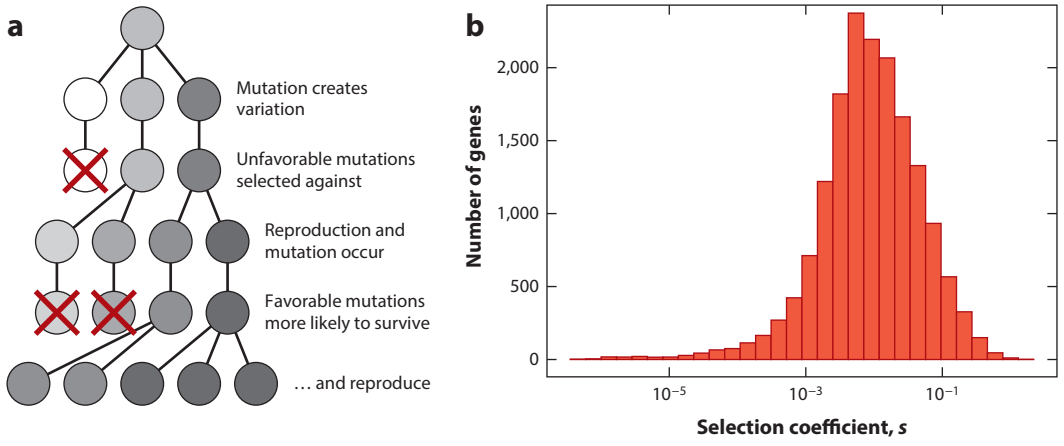

Figure 2

Natural selection shapes genome-wide variation. (*a*) A schematic of negative and positive selection acting over generations. Panel *a* adapted with permission from Wikipedia (**https://en.wikipedia.org/wiki/Evolution**) (CC BY-SA 3.0). (*b*) A histogram of the natural selection coefficient (*s*), here measuring only negative selection against deleterious (predicted loss-of-function) variation, for all protein-coding genes.

While most newly arising variants are evolutionarily neutral, natural selection also governs the longevity of variants in a population (**Figure 2*a***). In extremely rare cases, de novo mutations may be beneficial in the population and be positively selected, increasing in frequency by conferring a survival advantage. Conversely, some variants may result in severe disease or even embryonic lethality and thus may not be transmitted to future generations. Between strongly deleterious and neutral variants, those that slightly decrease fitness result in an average of fewer progeny per generation. These fitness effects lead to a decrease in the frequency of the mutation through a mechanism called negative selection. A selection coefficient measures the difference in relative fitness of individuals with a particular genotype (**Figure 2*b***), including positive and negative selection. Robust examples of positive selection in humans are less common but have occurred in traits like lactase persistence, skin pigmentation, hypoxia adaptation, diet, and immune function.

The evolutionary forces of mutation, drift, and natural selection govern the landscape of genetic variation and how it varies in frequency throughout the genome. Summaries of variant counts by frequency are commonly called the allele or site frequency spectrum (SFS; **Figure 3*a***). The SFS can be polarized with respect to the derived allele frequency (the human-specific allele, compared to the allele in the closest diverged species, the chimpanzee) or simply refer to the minor allele frequency (the less common allele in the population). The vast majority of standing variations across humans are very rare. Meanwhile, the increase in frequency of some variants has resulted in a seemingly paradoxical phenomenon: Most variants in a population are rare, but most variants in an individual are common (**Figure 3*b***). Further, due to natural selection removing deleterious variants from the population, variants that have risen in frequency tend to be less deleterious than those that are rare (**Figure 3*c***).

The infinite sites model is a widely used assumption in population genetics in which mutations occur only once per site with no recurrence or back mutations. As allele frequencies are ascertained in increasingly large genetic datasets, we have recently observed and quantified a phenomenon in which the same genetic variant originates multiple times, invalidating the infinite sites model for these datasets (8) (**Figure 3*d***). This has the effect of saturating observations of certain classes of variation, particularly for CpG transitions, which have a ~100-fold higher mutation rate than

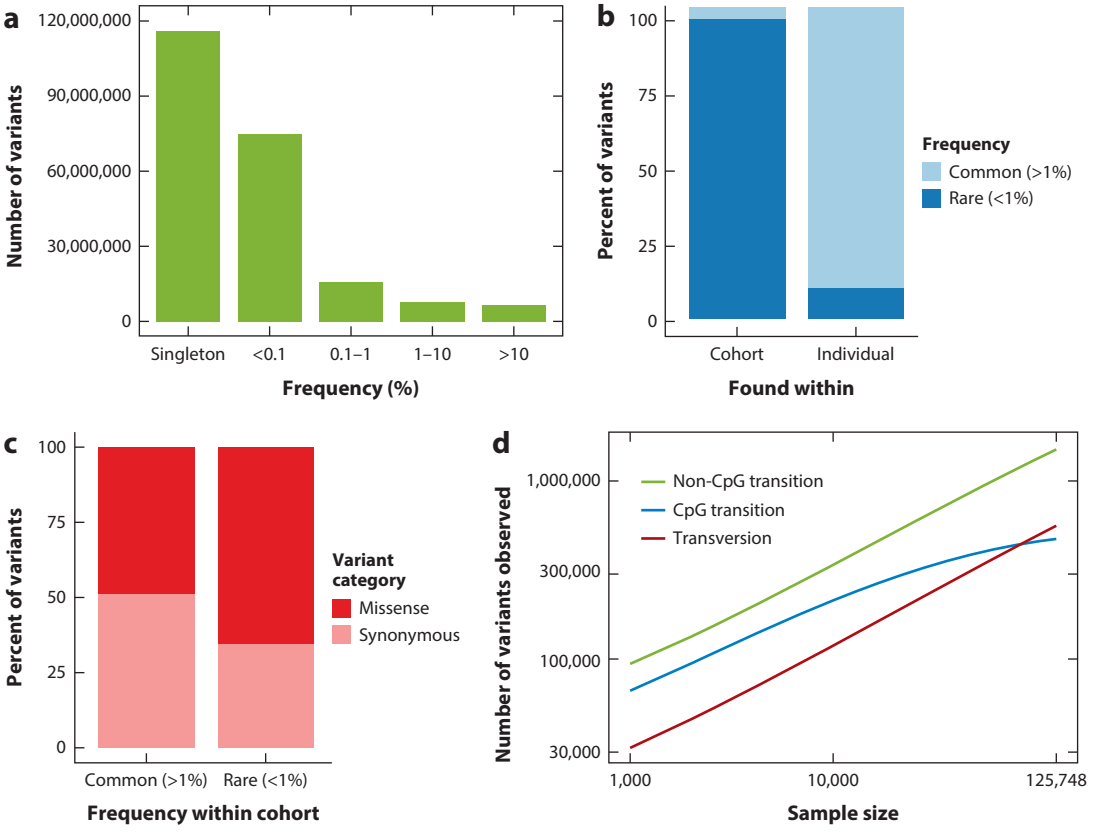

Figure 3

Allele frequency properties across the genome. (*a*) The number of variants for each frequency bin, showing that most variants across the genome are rare. (*b*) The percent of variants that are common or rare for all variants within a cohort of 125,748 individuals or within a single individual. Most variants in the population are rare but most of an individual's variants are common in the population. (*c*) The percent of variants that are missense and synonymous for rare and common variants in a cohort. Rare variants are more likely to be deleterious, while common variants are more likely to be neutral. (*d*) The number of variants observed for each mutational class as a function of increasing sample size. Mutational recurrence, the phenomenon that invalidates the infinite sites model according to which the same mutation originates multiple times, becomes clear at large sample sizes among CpG variants (8). Data from the Genome Aggregation Database (9).

other transitions or transversions, leading to a skew in the SFS. Specifically, these variants are less likely to be singletons (found only once) in large datasets due to multiple mutational origins, which can affect estimates of negative selection against these variants. For instance, the proportion of singletons among stop-lost variants is higher than that of stop-gained variants, despite the expectation that the latter have a larger deleterious effect; however, this is solely due to the inability of a CpG variant to create a stop-lost variant. This effect (the decrease in the proportion of singletons due to recurrence) can be adjusted using known mutation rates, estimated from background mutations in whole-genome data, in a metric termed the mutability-adjusted proportion of singletons (MAPS) (8); this metric has been used in other large cohorts to assess the level of negative selection against variant classes (13).

Recombination further shapes the inheritance of genetic variation by influencing how nearby variants are inherited. Offspring typically inherit one complete chromosome from their mother

and one complete chromosome from their father. Maternal and paternal chromosomes that are passed on have undergone meiosis, such that approximately one chromosomal crossover has taken place to physically swap chromosomal segments, resulting in a new mosaic of variation along the full chromosome that has not previously been seen. The segments that were directly copied from each parent are called haplotypes. Haplotype lengths decay exponentially as a function of generations due to recombination (2) (**Figure 4*a***). Intuitively, people inherit decreasing amounts of variation from their parents to their grandparents to their more distant ancestors. In this way, variants that are far away from each other on the same chromosome tend to not be inherited together, whereas variants that are close to each other are more correlated (**Figure 4*b–c***). This phenomenon (LD) is typically measured as the squared correlation between the genotypes of a pair of SNPs, r^2, which varies as a consequence of demographic history (**Figure 4*c***). Recombination events typically take place near hot spots: regions of the genome that are far more likely to undergo crossover events. These hot spots tend to be shared throughout a population and are regulated by *PRDM9* (14).

Consequences of Human Demography

A population's demographic history influences the landscape of genetic variation. Humans originated in Africa, and those humans who migrated to the Middle East and then to other parts of the world took sequential subsets of genetic diversity with them (**Figure 5*a***), a concept called a serial founder effect (15). Throughout human history, there have been population expansions and bottlenecks, which consequently increased or decreased the amount of genetic diversity in the population, respectively (**Figure 5*b***) (16–19). This history is important to understand for interpreting the function of variants across the allele frequency spectrum. For example, the mutation-selection balance that governs the efficiency of natural selection to purge deleterious variation from a population depends on effective population size (20). Relatedly, the classical concept of mutation load describes the burden of deleterious variants carried by a population (21, 22), which has been explored in depth more recently with a range of models and assumptions in sequencing data (23).

Natural variation keeps a direct ledger of population history, providing a useful tool to infer past demographic events. A typical first step for summarizing genetic variation in a population and subsequently correcting for this structure in association studies is using principal components analysis (PCA) on genotype data (24, 25). PCA applied to high-dimensional genetic data summarizes ancestry differences by providing continuous axes of variation that reflect genetic differences. These continuous principal components based on genetic data mirror geographical distance (**Figure 6*a*,*c***) (24–26). More complex insights into population history from modern genetic data, such as how effective population sizes changed, how populations mixed (**Figure 6*b***), and how populations migrated over time can be inferred by comparing the allele frequency spectrum across populations and the patterns of haplotypic variation across populations, or through more complex coalescent models (7, 27–29) (for a recent review, see Reference 30).

When genetic variation is assessed, we typically know its genotype state, but not which parent passed on each allele. Determining which alleles are inherited on a chromosome together, or the so-called haplotype structure, can be important for interpreting variant function and a helpful step prior to running a GWAS. Because nearby regions of the genome tend to be inherited together, creating LD, and because large numbers of community-wide reference genomes have been sequenced, an entire genome does not need to be deeply sequenced to estimate the genotype state at all variants (31).

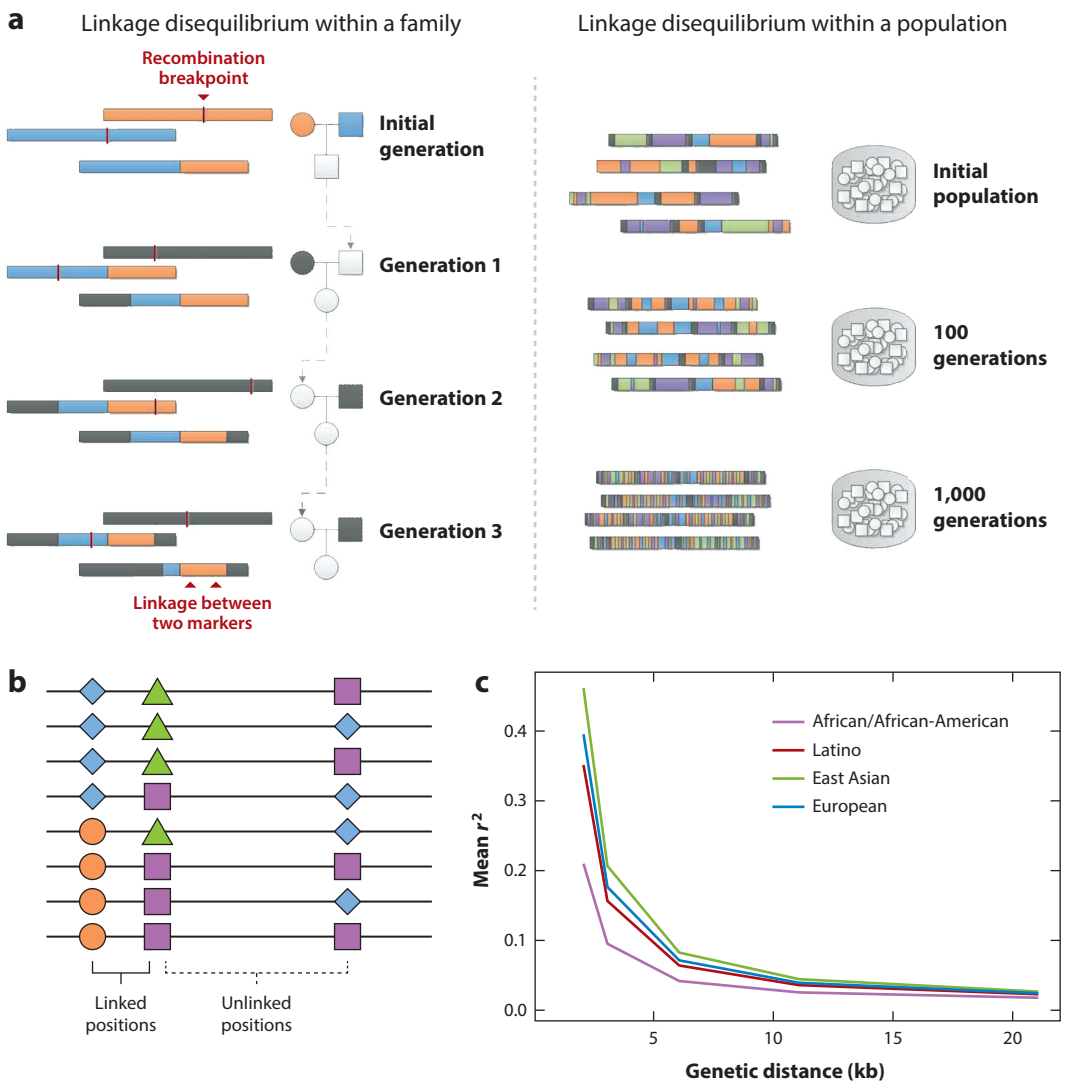

Figure 4

Variants that are physically close tend to be transmitted together. (*a*) A schematic of linkage disequilibrium (LD), the phenomenon in which variants are linked in the population. Nearby variants are inherited together, and recombination begins to break down these blocks over varying timespans within families (*left*) and in populations (*right*). In both cases, the local haplotype structure remains apparent. Panel *a* adapted with permission from Reference 98 (CC BY). (*b*) Recombination over time results in higher correlation between nearby variants, and lower or no correlation for distal pairs of variants. (*c*) Mean-squared correlation (r^2, a measure of LD) across pairs of variants from individuals in the Genome Aggregation Database (9) as a function of distance between the variants. Notably, patterns of LD vary as a function of human demographic history.

This principle is commonly used in GWAS to balance cost considerations, as GWAS arrays are more cost effective than sequencing a whole genome to high coverage. Variants are first statistically phased with a reference panel of haplotypes to obtain estimates of genome-wide genotype states in a cohort. Phasing approximately decouples which haplotype segments were inherited together in each individual. Next, using these phased haplotypes, researchers fill in gaps between variants on

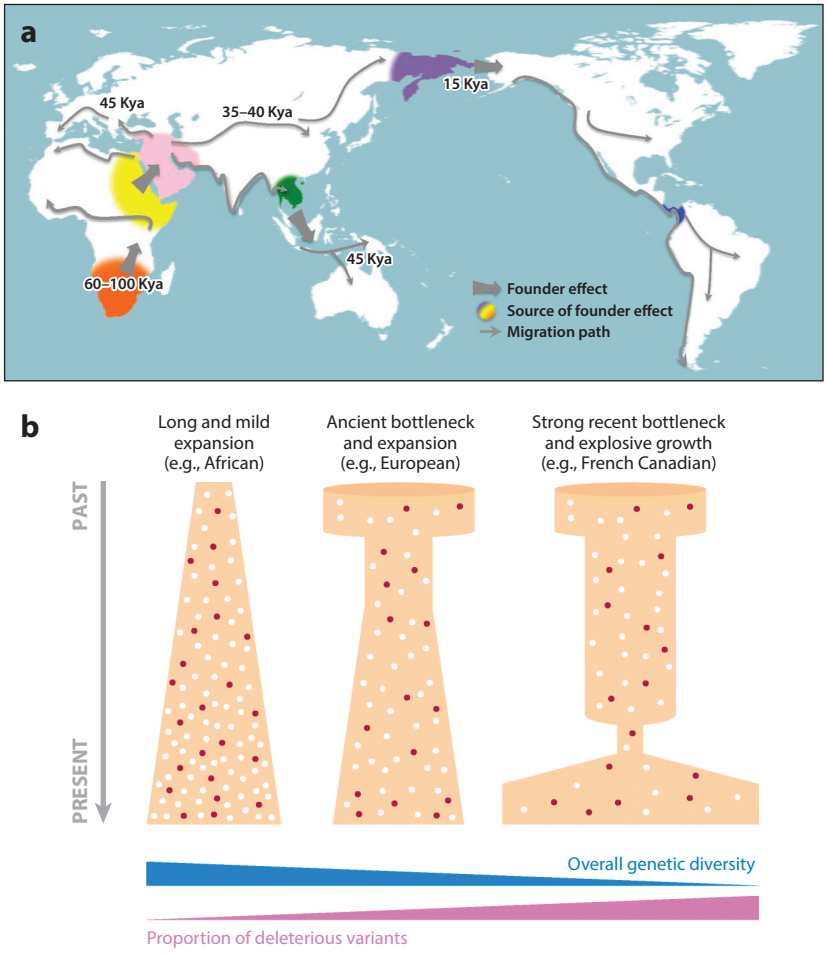

Figure 5

Human history shapes patterns of genetic diversity. (*a*) A map of global human migration history over time, including spatiotemporal founder effects (bottlenecks) as humans populated the globe. Panel *a* adapted from Reference 16. (*b*) Population expansion and bottleneck models modify the extent of genetic diversity and the spectrum of variant deleteriousness. Panel *b* adapted with permission from Reference 19 (CC BY 4.0).

the GWAS array that are captured in these reference haplotypes through imputation. This process is important for harmonizing large-scale datasets, especially when they have been genotyped on multiple GWAS arrays. Because SNP ascertainment on GWAS arrays is often biased (often toward variants most common in European ancestry populations), low-coverage sequencing has been proposed and used as an unbiased strategy with a similar cost to GWAS arrays (32–34).

ASSOCIATING COMMON VARIANTS WITH PHENOTYPES

A mainstay of human genetics is determining the extent to which genetic versus environmental factors contribute to a phenotype, i.e., its heritability (35). After inferring a trait's heritability, a typical next step involves identifying genetic variants responsible for diseases or trait differences. However, natural selection decreases the frequency of high-impact deleterious variants, altering

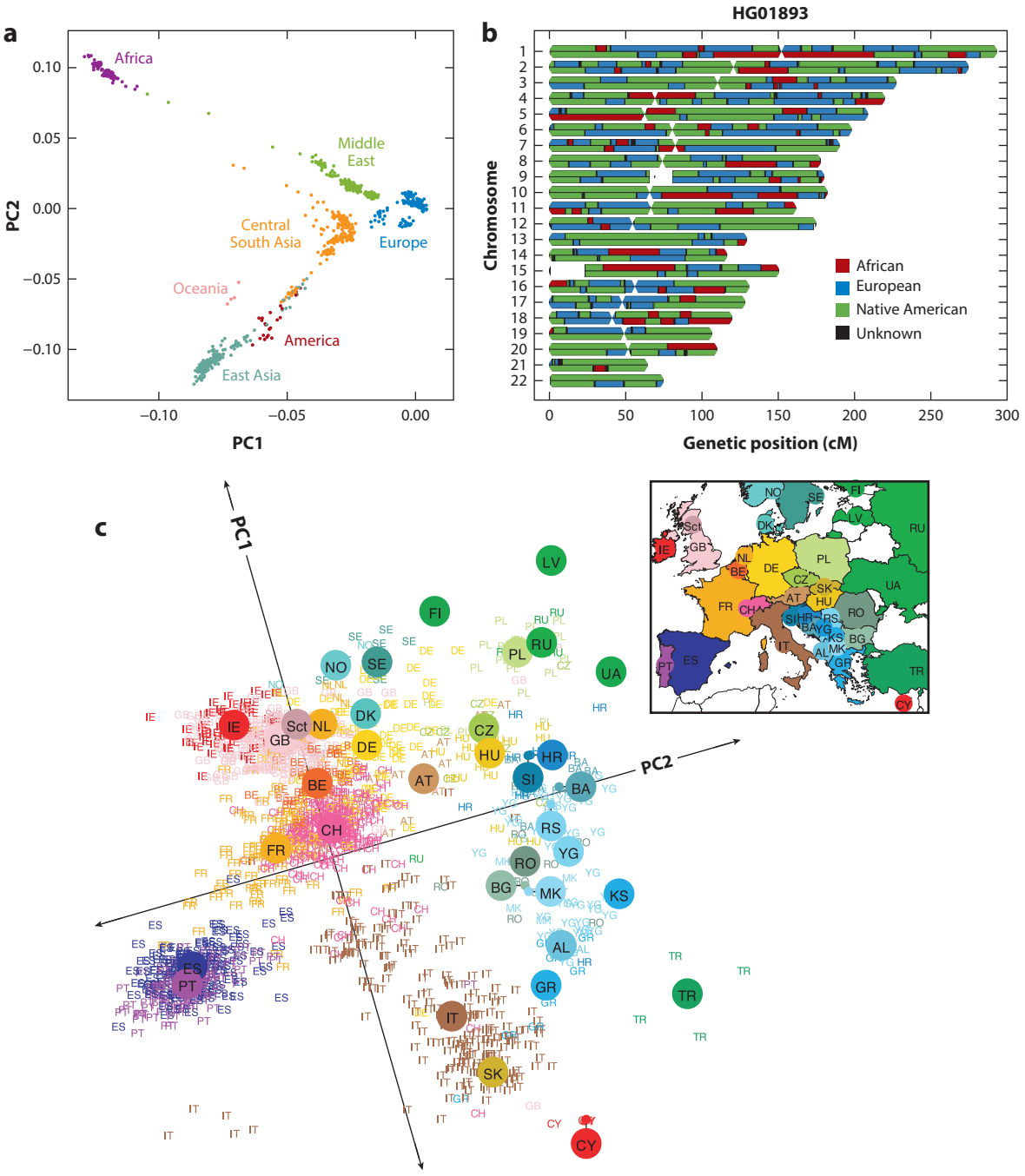

Figure 6

Population structure can be inferred from genetic data. (*a*) Principal components analysis (PCA) across the Human Genome Diversity Panel reflects global genetic diversity (99). The first two principal components are shown. (*b*) Ancestry karyogram illustrates ancestry mosaic in chromosome 1 in a recently admixed Peruvian individual sequenced in the 1000 Genomes Project (62, 63). (*c*) PCA across PopRes (Population Reference Sample) reflects genetic diversity within Europe. Panel *c* adapted with permission from Reference 26; copyright 2008 Springer Nature.

the spectrum of disease-associated variation. In particular, common variants typically have small effects on disease risk and thus large sample sizes are needed to robustly identify associations. Despite small effects identified in common variant studies, the effect of drugs that target genes with significant associations need not be small. For instance, *PCSK9* inhibitors can dramatically decrease LDL (low-density lipoprotein) cholesterol in families with familial hypercholesterolemia, even though common variant associations with *PCSK9* identified in GWAS modulate LDL by a modest amount (**Table 1**).

In contrast to common variant association studies, large effect sizes are typically difficult to detect due to their low frequency (**Figure 7*a***). The relationship between effect size and allele frequency has been explored across many traits and diseases and follows a predictable relationship in which common variants tend to have smaller effect sizes than rare variants. Combining concepts of effect size and allele frequency, common variants together tend to explain more of the heritable variation in a population, with few exceptions (36) (**Figure 7*b***), while rare variants can be more impactful especially to smaller numbers of individuals. For example, the genetic basis underlying type II diabetes risk for the vast majority of the population arises primarily from common variants. However, for some individuals, high-impact rare variants cause subtypes or other types of the disease, such as maturity onset diabetes of the young.

GWAS have been highly successful in identifying an abundance of genotype-to-phenotype associations. In GWAS, a large number of individuals with quantitative, binary (i.e., case/control), or other phenotype information are genotyped, phased, and imputed. Typically, millions of statistical association tests are performed between each variant and phenotype using an appropriate statistical test. Because variants in the genome are linked (in LD), many of the variants analyzed in GWAS are not independent. Previous works assessing the number of independent tests have shown that there are approximately one million independent tests in the genome. Thus, variants in GWAS that pass multiple test corrections and thus meet a p-value threshold less than 5×10^{-8} ($0.05/10^6$) are considered genome-wide significant. Following laws of statistical power, GWAS are better powered for common variants, as many individuals with a given variant are observed.

Methods and Applications of Genome-Wide Association Studies

Depending on the nature of the trait and sample set, different statistical tests are warranted when conducting a GWAS. Linear regression is commonly used for quantitative traits, and logistic regression is used for binary traits; in both scenarios, covariates that may confound or stratify results should be included, such as sex, age, principal components for ancestry, and potentially combinations of or interactions between these factors. Across all GWAS models, the modeled outcome is the phenotype, while the covariates and SNPs are predictors of the phenotype, each with an effect size coefficient. Each SNP is tested separately. Mixed models, which incorporate a genetic relatedness matrix between all pairs of individuals, can provide additional power, especially when related samples and heterogeneous populations are included, although at an additional computational efficiency cost. Fortunately, novel approaches have recently been developed that improve the efficiency of these methods and enable their use for biobank-scale GWAS (37, 38).

GWAS for thousands of traits and diseases have now yielded hundreds of thousands of associations with variants across the genome (39). The larger the sample size, the better powered the study (**Figure 7*c***), and thus many of the most successful GWAS efforts have resulted from consortia that have aggregated very large quantities of data. These efforts have uncovered new biology important for understanding the molecular basis of disease, including schizophrenia (40), type II diabetes (41–43), inflammatory bowel disease (44, 45), and coronary artery disease (46), along with its associated traits such as triglyceride and cholesterol levels (47, 48). Additionally, large-scale biobanks

Table 1 Genetic studies inform therapeutic discoveries and development

Gene target	Disease/indication	Genetic model and phenotype	Drug/candidate	References
PCSK9	Familial hypercholesterolemia, primary or secondary prevention of ASCVD	GoF mutations cause autosomal-dominant familial hypercholesterolemia; LoF mutations cause low LDL, reduced ASCVD	Praluent (alirocumab) Repatha (evolocumab)	80, 82
CFTR	Cystic fibrosis	In-frame deletions cause cystic fibrosis	Kalydeco (ivacaftor) Trikafta (elexacaftor/ivacaftor/tezacaftor)	102–104
HBB	Sickle cell	GoF mutations (β6Glu→Val) cause red blood cells to form a sickle shape and clump together	Adakveo (crizanlizumab)	105, 106
CCR5	HIV/AIDS	Naturally occurring genetic variants protect against HIV infection	Selzentry (maraviroc)	107, 108
LRRK2	Parkinson's disease	GoF mutations cause Parkinson's disease	DNL201 and DNL151 (phase I)	109, 110
APOC3	Familial chylomicronemia syndrome; familial partial lipodystrophy	LoF mutations cause reduced TG, higher HDL, reduced CHD risk	Waylivra (volanesorsen)	111, 112
SGLT2	Type II diabetes	LoF mutations cause familial renal glucosuria, decreased glucose	Invokana (canagliflozin) Farxiga (dapagliflozin) Jardiance (empagliflozin) Steglatro (ertugliflozin)	113–116
SMN1	Spinal muscular atrophy	LoF mutations cause spinal muscular atrophy	Zolgensma (onasemnogene abeparvovec-xioi); uses AAV to deliver new *SMN1* Spinraza (nusinersen); restores a functional copy of *SMN1* by regulating *SMN2*	68
NLRP3	Cryopyrin-associated periodic syndromes	GoF mutations cause increased IL-1β, autoimmune disorders	Arcalyst (rilonacept) Ilaris (canakinumab)	117, 118
TYK2	Immunologically mediated diseases (psoriasis, lupus, IBD)	Variants are associated with immune diseases via GWAS (reviewed in Reference 119)	BMS-986165 (phase II)	120, 121

In this noncomprehensive list, there is an increasingly well-characterized set of genetic factors that lead to the disease state. The genetic model provides a mechanism to modulate or perturb via therapeutic avenues, many of which are currently approved or under development. Abbreviations: AAV, adeno-associated virus; ASCVD, atherosclerotic cardiovascular disease; CHD, congenital heart disease; GoF, gain of function; GWAS, genome-wide association study; HDL, high-density lipoprotein; HIV/AIDS, human immunodeficiency virus/acquired immunodeficiency syndrome; IBD, inflammatory bowel disease; LDL, low-density lipoprotein; LoF, loss of function; TG, triglyceride.

increasingly share GWAS summary statistics for a breadth of phenotype measures across tens to hundreds of thousands of individuals, including electronic health record information, which in turn enables novel genetic discoveries in an untargeted manner (49). Among seemingly disparate phenotypes, many of the associations that have been identified appear to overlap across multiple traits and diseases. This phenomenon highlights the pervasive nature of pleiotropy, in which genetic variants influence two or more seemingly unrelated traits (50).

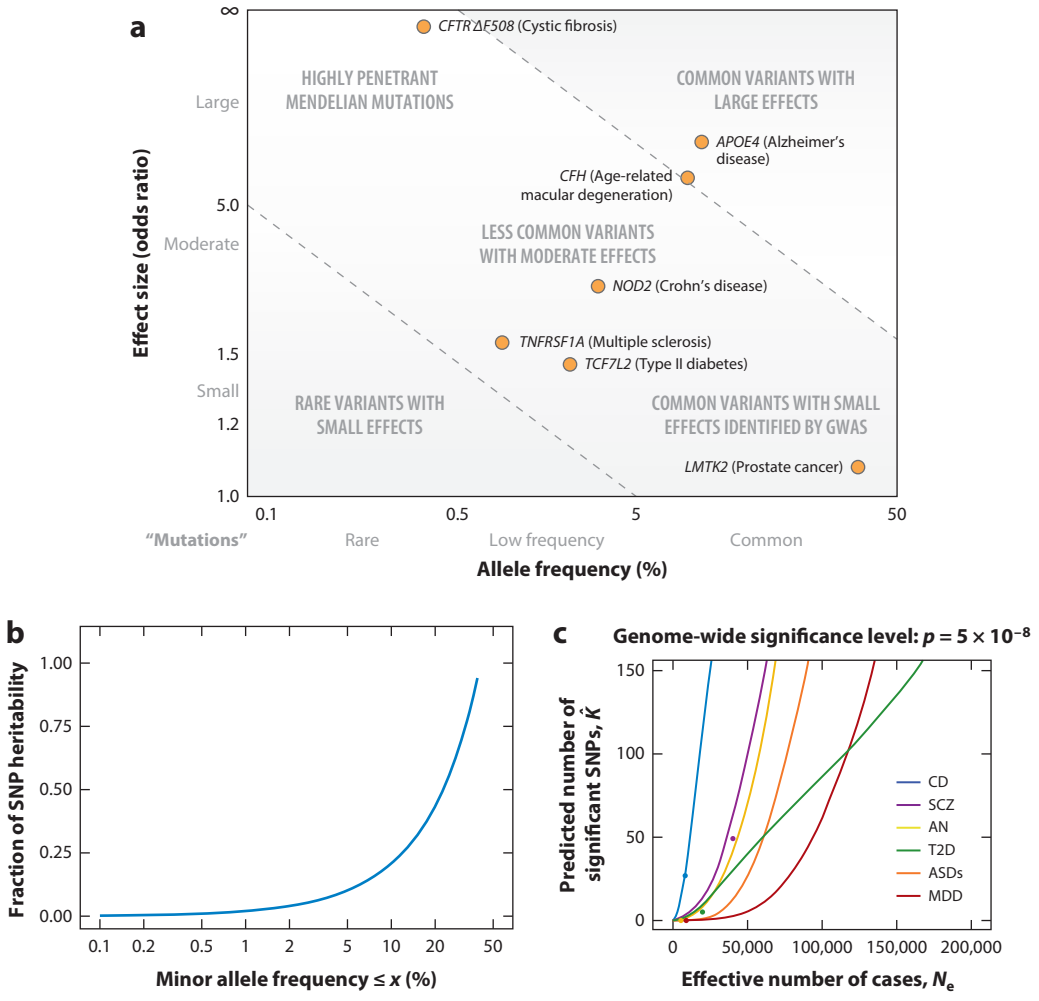

Figure 7

Power in GWAS is informed by allele frequency, effect size, and sample size. (*a*) The relationship between frequency and effect size in genetic studies. Panel *a* adapted with permission from Reference 98 (CC BY). (*b*) A cumulative plot of SNP heritability as a function of allele frequency. Common variants tend to explain most of SNP heritability (36). (*c*) The number of independent genome-wide significant SNPs from GWAS increases with sample size across many complex diseases. The discovery curve varies with the genetic architecture of the disease. Panel *c* adapted with permission from Reference 100 (CC BY). Abbreviations: AN, anorexia nervosa; ASD, autism spectrum disorder; CD, Crohn's disease; GWAS, genome-wide association studies; MDD, major depressive disorder; SCZ, schizophrenia; SNP, single-nucleotide polymorphism; T2D, type II diabetes.

Post–Genome-Wide Association Study Analysis

After GWAS have been conducted, analyses are often undertaken to make use of summary statistics, along with LD or functional information to better understand biology. A major consideration for all of these analyses is that while GWAS can robustly identify associations, the most significant associations are not necessarily causal variants. Thus, some examples of post-GWAS analyses include pinpointing causal variants through fine-mapping (51, 52), understanding the causal relationships between genetic instruments for a given exposure and disease via Mendelian

randomization (53), identifying functional signatures in the genome enriched in cell types and pathways that explain more or less of the heritable variation (54), and assessing how predictive these genetic studies are in new cohorts via polygenic prediction (55–58). Relatedly, thousands of phenotypes have now been analyzed in large-scale biobanks, but these traits and diseases are not independent. Thus, genetic correlation enables a data-driven assessment of how genetically similar two phenotypes are (59, 60). Additionally, these approaches can take the conjecture or specialty out of phenotype precision and certainty, while potentially informing new epidemiological considerations.

Caveats

An ongoing challenge with the interpretation of GWAS and downstream analyses is the effect of study bias and stratification (**Figure 8*a***). GWAS are currently vastly Eurocentric, limiting many insights into non-European ancestry populations. For example, a recent genetic study in Africans focusing on Ugandan populations recently conducted a GWAS (61); despite the fact that sample sizes in this study were smaller than several European studies, they found that more than half of the fine-mapped genome-wide significant associations in their study had previously been missed due to their low or invariant frequencies in European ancestry populations. Similarly, because GWAS are best powered to discover variants most common in the population studied, the frequencies of variants in the GWAS catalog tend to be most common in European ancestry populations and less common elsewhere (62, 63). When GWAS are underpowered, effect sizes can be overestimated due to winner's curse.

A pragmatic approach to the data privacy concerns in the genetics consortium model, which has allowed genetic studies to become increasingly large, is for each contributing researcher to share GWAS summary statistics after correcting for age, sex, and ancestry covariates. These summary statistics are then meta-analyzed to aggregate results, thereby increasing power. With this approach, however, residual stratification is likely to remain because principal components from smaller cohorts cannot sufficiently control for population structure (**Figure 8*b***). This has resulted in overestimated signals of polygenic differences in height, cardiovascular disease, body mass index, waist–hip ratio, and other traits and diseases (63–66). While currently there are not clear approaches to removing stratification effects that arise from aggregating small cohorts when investigating polygenic differences geographically, vigilance about how these effects may impact interpretations is critical.

RARE VARIANT–DISEASE ASSOCIATIONS

In addition to GWAS, which are best powered for identifying common variant associations with disease (often with small effect sizes), different conceptual approaches must be taken for rare variants. Due to their low frequency, it is more difficult to ascribe function to rare variants; however, they may have very strong influences on severe diseases. In particular, analyses of exome or genome data from patients with severe Mendelian diseases have identified hundreds of causal gene effects. For more common diseases, rare variants can be aggregated using various association methods, but very large sample sizes are required to achieve statistical power.

Mendelian Disease

Mendelian diseases, such as sickle cell disease and cystic fibrosis, are single-gene disorders that are inherited in a dominant or recessive manner. These are typically more severe than complex

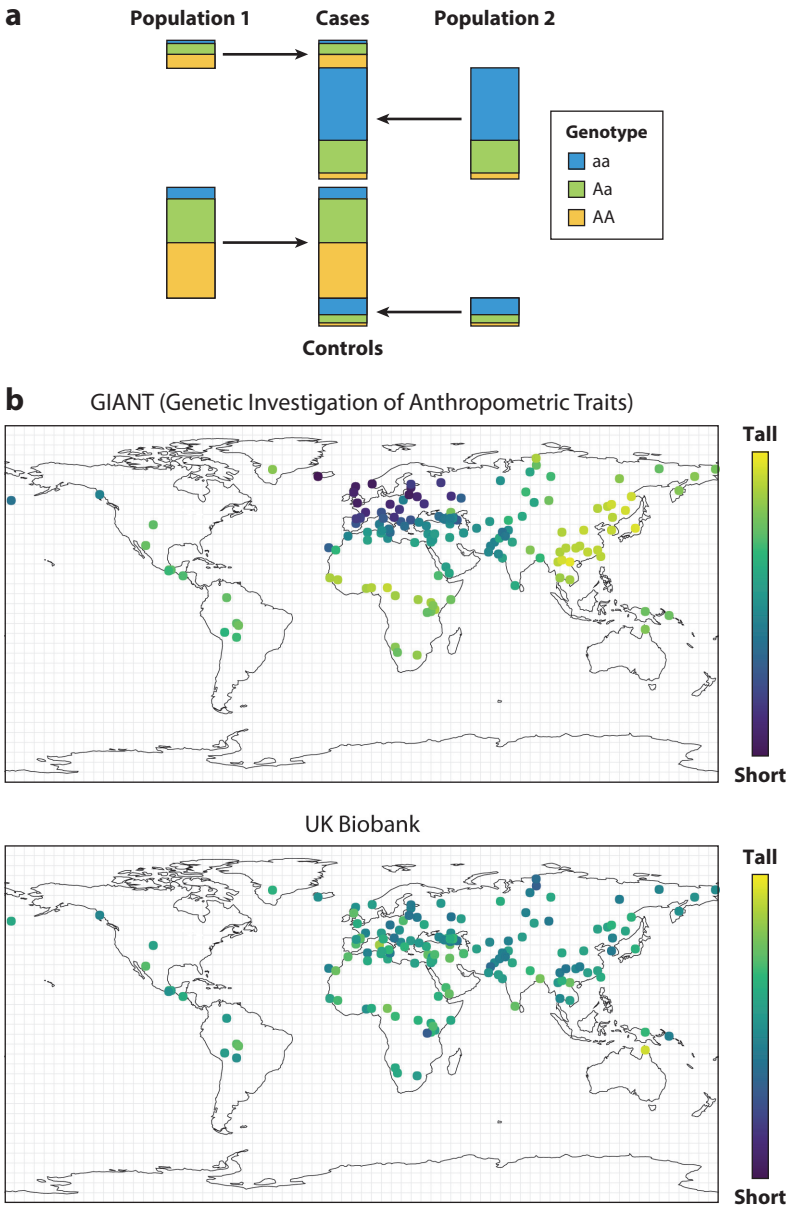

Figure 8

Population stratification is an important consideration in genome-wide association studies (GWAS). (*a*) Failing to account for population structure can result in spurious associations when population and case/control sampling are imbalanced. Here, ascertaining more cases from Population 2 and controls from Population 1 would result in a spurious association of the "a" allele with case status. Panel *a* adapted with permission from Reference 101; copyright 2004 Springer Nature. (*b*) Due to differences in GWAS design and control, inferred polygenic differences in height applied to various populations (shown as a color gradient) are markedly different based on the GWAS discovery cohort from which the scores are derived (64). Panel *b* courtesy of Robert Maier.

diseases and are more likely to elicit effects early in life. Within a family, these diseases can be mediated by a single variant, and identifying the causal variants among the background set of variants found in every individual remains a significant challenge.

Decades of linkage studies have successfully identified genetic factors for recessive and dominant diseases (reviewed in Reference 67). For example, spinal muscular atrophy (SMA) is a group of neuromuscular diseases that are genetically caused by mutations in *SMN1* (68). Recently, two therapeutic approaches have been developed to restore the function of *SMN1* in SMA patients (**Table 1**). More recently, approaches using exome and genome sequencing (69) have been applied to affected individuals and unaffected family members to accelerate Mendelian gene discovery (70, 71).

As patients continue to have their exomes and genomes sequenced, knowledge bases of causal genes and variants such as ClinVar (72) are invaluable for the diagnosis of patients; these diagnoses are beneficial for ending time-consuming and challenging diagnostic odysseys, and potentially for informing therapeutic avenues. Large reference panels of genetic data that are depleted for rare disease patients, such as ExAC and gnomAD, can also aid in variant interpretation. Specifically, they provide accurate estimates of the allele frequency of each variant, which, as a result of the population genetics forces discussed above, is one of the most informative features of a variant's function or deleteriousness. Thus, these databases provide a mechanism by which to filter variants in patients (8), as variants that are at a significantly higher frequency in the general population than the disease prevalence cannot be causal for the disease (73). For syndromes where multiple genes have been previously implicated, filtering to these genes can reduce the search space; however, for previously undescribed phenotypes, further functional experiments such as enzymatic assays on biologically plausible genes can be informative. Overall, exome sequencing typically yields diagnostic rates of 30–50% depending on the specific ascertainment strategy and phenotype, which can be further improved with functional information (74).

A Continuum Between De Novo and Rare Variants

Similarly, in disorders where rare high-impact variation is expected to have a major influence, scans for de novo damaging missense or predicted loss-of-function (pLoF) variants have been successful in identifying causal genetic factors, including *ARID1B* and *PPM1D* for developmental disorders (75) and *SCN2A* and *CHD8* for autism spectrum disorder (ASD) (76). A crucial observation from these studies is that even in the absence of genome-wide significant signals, there is an excess burden of de novo variation in cases (**Figure 9**), particularly in constrained genes (see below) (77). The genes and pathways that are damaged by these de novo signals can inform disease etiology and warrant additional sample sizes to pinpoint causal factors. For instance, a recent study showed that the burden of de novo pLoF variation was similar in cases of ASD and attention-deficit/hyperactivity disorder, and that the patterns of genes harboring these variants were generally indistinguishable, suggesting a shared etiology in the genetic makeup of these disorders (78). Further, combining analyses of de novo variants with inherited variants, the same authors implicated over 100 risk genes for ASD and showed that these genes are involved in specific neuronal lineages (79). Preliminary analyses of de novo variation in regulatory regions have shown an overall enrichment in cases with developmental disorders, but larger sample sizes and improved annotation methods will be required to pinpoint individual causal elements (13).

Sequencing studies that identify rare pLoF variants in healthy individuals have also fundamentally informed therapeutic development (**Table 1**). In some cases, gain-of-function variants result in a disease phenotype; in these genes, pLoF variants may be protective for the same disease. For instance, gain-of-function variants in *PCSK9* are causal for hypercholesterolemia (80); accordingly,

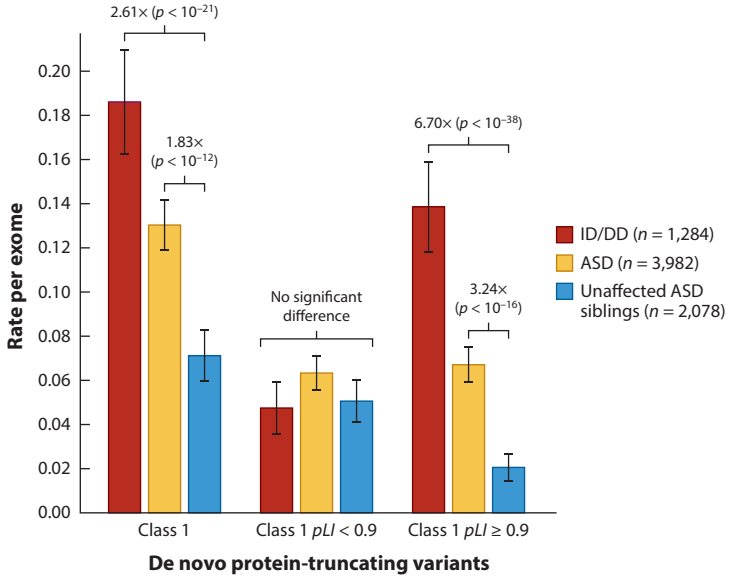

Figure 9

Increased burden of de novo variants in rare disease. Patients with ID/DD have a 2.61-fold higher burden of ultrarare pLoF variants than unaffected siblings (1.83-fold for ASD). This enrichment is higher for pLoF variants in highly constrained genes. Abbreviations: ASD, autism spectrum disorder; ID/DD, intellectual disability/developmental delay; *pLI*, probability of loss-of-function intolerance; pLoF, predicted loss of function. Figure adapted with permission from Reference 77; copyright 2017 Springer Nature.

individuals with pLoF variants in *PCSK9* have lower LDL cholesterol (81, 82). Because the loss of *PCSK9* is known to be tolerated and protective for coronary heart disease, pharmaceutical inhibition of PCSK9 is likely to be a safe and effective strategy, respectively, for controlling lipid levels. Similarly, individuals with *LRRK2* gain-of-function variants are at increased risk for Parkinson's disease (80, 81), and therefore, inhibition of LRRK2 may be a viable therapeutic strategy. Recently, we showed that pLoF variant carriers of *LRRK2* who have a lifelong 50% reduction of the gene do not appear to have any adverse phenotypes (83). This suggests that inhibition of LRRK2 is unlikely to result in any major on-target toxicity. These and other examples (**Table 1**) represent a new generation of drug targets with genetic support: Indeed, such drugs are more likely to be approved than those without (84). This has led to calls for a human knockout project to identify new druggable targets. Given the rare nature of these variants and phenotypes, large sample sizes are required to ascertain these phenotypic effects or lack thereof. In some cases, sequencing individuals from consanguineous matings or bottlenecked populations can enrich for homozygous pLoF variant carriers.

Gene-based tests that combine multiple variants, including burden tests and variance tests, can boost signal for intermediate frequency variants, where an individual variant may not have statistical power to be implicated with a given disease. In burden tests, we consider the aggregate burden of variants in a gene (e.g., the sum of minor alleles weighted by inverse allele frequency) in a statistical test against a phenotype. These tests are most effective under the assumption that all variants in the gene contribute signal in the same direction, but they are generally underpowered at current sample sizes. Indeed, a recent study of rare pLoF variants in highly constrained genes identified an aggregate burden of these variants in height, psychiatric disorders, and educational attainment, but it was not powered to detect signals at individual genes (85). Alternatively,

variance tests, including C-alpha (86) and SKAT (sequence kernel association test) (87), relax this assumption by modeling the shift in the phenotypic variance rather than the mean. For instance, the overdispersion of rare variation among individuals with extreme levels of triglycerides provided support for *APOB*'s influence on lipid levels (86). Recent optimizations of gene-based tests (38) have enabled their use in biobank-scale datasets, including the UK Biobank (88). For both GWAS and sequencing analyses, the high cost of sample recruitment and data generation makes the use of public reference datasets as control cohorts a tempting proposition. However, technical stratification from a variety of sources (e.g., differences in sequencing depth, variant calling, etc.) is a major confounder, limiting the robustness of these studies.

INFERENCE OF VARIANT AND GENE DELETERIOUSNESS

Large datasets of exome and genome sequence data can provide not only information about the levels of current standing variation but also insight into regions of the genome where there are significant depletions of variation. Different methods have been developed to detect depletions of deleterious variation, including the Residual Variation Intolerance Score (RVIS) and a method known as constraint. Both approaches model an expectation of deleterious variation based on neutral variation and compare this expectation to the observed number of variants. RVIS models the number of common deleterious (missense or truncating) variants as a function of the total number of variants in a given gene, and the residuals for each gene are transformed into the RVIS score (89). This score was applied to data from 6,503 individuals with exome sequence data from the Exome Sequencing Project to provide an unbiased predictor of haploinsufficient and developmental genes, which could thus prioritize new disease genes.

Constraint methods use the underlying mutation rate of each possible variant in a set of variants of interest (for instance, missense variants in a gene) to compute the expected number of variants in that set in a given cohort (90) (**Figure 10*a–c***). A significant depletion (i.e., constraint) of observed variation compared to this expectation indicates the presence of negative selection. This framework was also applied to pLoF variants in the ExAC dataset to identify 3,230 genes with a strong depletion of pLoF variants (~9% observed/expected ratio), which were defined as having a high probability of loss-of-function intolerance (*pLI*) (8). These genes are enriched for known haploinsufficient disease genes, indicating their strong impact on human phenotypes. However, approximately 70% of high-*pLI* genes have no known human disease association, suggesting that they may be undiscovered disease genes, or possibly incompatible with human life when inactivated.

Recently, we extended this framework, improving the mutational model and variant filtration (9). We applied these extensions to the gnomAD dataset; these improvements and the larger sample size enabled the creation of a continuous score for pLoF depletion we termed LOEUF (loss-of-function observed/expected upper bound fraction). This score is correlated with numerous metrics of gene function (**Figure 10*d***) and can be used to prioritize disease genes. Unlike the dichotomous nature of *pLI*, LOEUF provides more fine-grained resolution for assessing depletion of deleterious variation across the whole genome.

Outside the exome, several efforts have begun to model the depletion of variation in noncoding regions. However, the lack of a triplet code or boundaries on functional elements (as in exons for coding regions) and, thus, the lower accuracy of annotation methods complicate functional interpretation. Without accurate annotation methods, sample sizes that are orders of magnitude larger than currently available datasets will be required to identify regulatory elements undergoing natural selection (13). However, some global patterns are beginning to emerge, including the presence of redundancy in enhancers that regulate essential genes (91). Recently, we extended

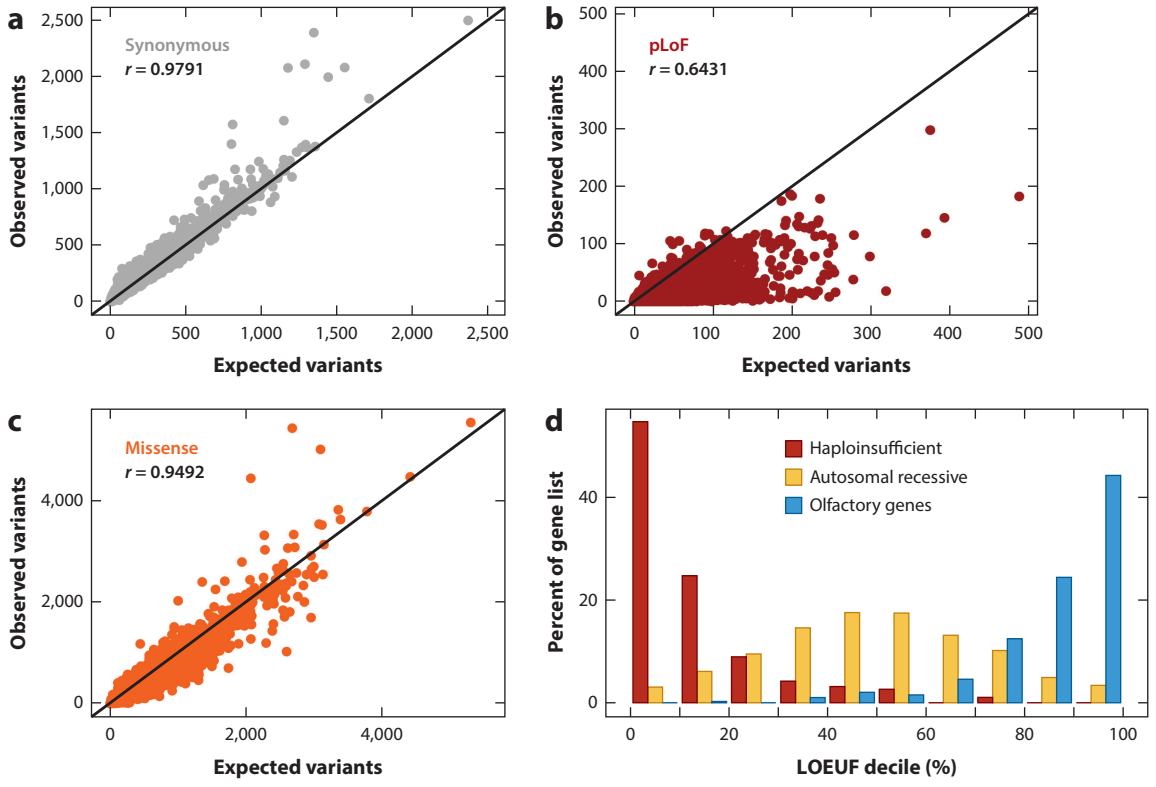

Figure 10

Identifying genes depleted of genetic variation. The observed and expected number of variants, as described by mutational models of constraint, for (*a*) synonymous, (*b*) predicted loss-of-function (pLoF), and (*c*) missense variation in the Genome Aggregation Database (gnomAD). (*d*) The depletion of pLoF variation is quantified using LOEUF (loss-of-function observed/expected upper bound fraction; here binned into deciles), which is correlated with established gene lists based on their phenotypic severity. Figure adapted with permission from Reference 9 (CC BY 4.0).

the MAPS approach described above to characterize variants that create upstream open reading frames (92), suggesting that they are at least as deleterious as an average missense variant. Such targeted approaches are likely to continue to be successful in identifying deleterious classes of variation. Additionally, as much larger sets of WGS data become available along with functional data, approaches that integrate these data types will provide valuable insight into the regulatory processes of the genome and permit interpretation of disease variation in noncoding regions.

CONCLUSION

In humans, analysis of large-scale natural variation data can provide valuable insights into human disease and gene function. Joint analysis of genotype and phenotype data is especially powerful, but even in the absence of phenotype data, patterns of variation data across the genome can inform population history, natural selection, and the biological function of genes.

Analyses of natural human variation data have numerous applications in translational settings. From a rare disease perspective, the creation of databases such as gnomAD aids in the interpretation of rare variation: Sequencing the exomes and genomes of rare disease patients and providing

a molecular diagnosis can have a profound impact by suggesting therapeutic avenues or ending diagnostic odysseys. From a common disease perspective, GWAS enable the creation of polygenic risk scores, which alongside other clinical factors can stratify patients into high- and low-risk bins for a given disease. Relatedly, GWAS can identify regions of the genome where biological signal is enriched and guide functional studies toward biological mechanisms.

Similarly, studies of naturally occurring pLoF variants can provide insights into gene function and therapeutics in a very concrete way: If individuals who lack a gene are protected from a particular disease and are otherwise healthy, a drug that reduces expression of that gene may be both safe and effective. Examples of genetically guided drug development are now available (**Table 1**), and further genetic studies can help identify new drug targets (93, 94) or provide evidence of safety (83, 95). Genetics already provides useful information about the organ systems impacted in retrospective analyses of clinical trial successes and failures (96). In this way, the genetic understanding of human biology has ushered in a new era of drug discovery that is likely to continue for decades to come (97).

DISCLOSURE STATEMENT

A.R.M. is a consultant for 23andMe and a member of the scientific advisory board at Precisely, Inc.

ACKNOWLEDGMENTS

The authors would like to thank Mark Daly for his exceptional leadership building the Analytic and Translational Genetics Unit. A.R.M. is supported by funding from the National Institutes of Health (K99MH117229).

LITERATURE CITED

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* 409(6822):860–921
2. Int. HapMap Consort. 2005. A haplotype map of the human genome. *Nature* 437(7063):1299–320
3. Int. HapMap 3 Consort. 2010. Integrating common and rare genetic variation in diverse human populations. *Nature* 467(7311):52–58
4. Mardis ER. 2008. Next-generation DNA sequencing methods. *Annu. Rev. Genom. Hum. Genet.* 9:387–402
5. Bergström A, McCarthy SA, Hui R, Almarri MA, Ayub Q, et al. 2019. Insights into human genetic variation and population history from 929 diverse genomes. bioRxiv 674986. **https://doi.org/10.1101/674986**
6. Fairley S, Lowy-Gallego E, Perry E, Flicek P. 2019. The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. *Nucleic Acids Res.* 48:D941–47
7. Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, et al. 2012. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 337(6090):64–69
8. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, et al. 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536(7616):285–91
9. Karczewski KJ, Francioli LC, Tiao G, Cummings BB. 2019. The mutational constraint spectrum quantified from variation in 141,456 humans. bioRxiv 531210. **https://doi.org/10.1101/531210**
10. Jónsson H, Sulem P, Kehr B, Kristmundsdottir S, Hardarson MT, et al. 2017. Parental influence on human germline *de novo* mutations in 1,548 trios from Iceland. *Nature* 549(7673):519–22
11. Collins RL, Brand H, Karczewski KJ, Zhao X, Alfoldi J, et al. 2019. An open resource of structural variation for medical and population genetics. bioRxiv 578674. **https://doi.org/10.1101/578674**
12. Hartl DL, Clark AG. 2006. *Principles of Population Genetics*. Sunderland, MA: Sinauer Assoc. 4th ed.

13. Short PJ, Gallone G, Geschwind DH, Barrett JC, Hurles ME. 2018. *De novo* mutations in regulatory elements in neurodevelopmental disorders. *Nature* 555:611–16
14. Baudat F, Buard J, Grey C, Fledel-Alon A, Ober C, et al. 2010. PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science* 327(5967):836–40
15. Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. 2005. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *PNAS* 102(44):15942–47
16. Henn BM, Cavalli-Sforza LL, Feldman MW. 2012. The great human expansion. *PNAS* 109(44):17758–64
17. Martin AR, Karczewski KJ, Kerminen S, Kurki MI, Sarin A-P, et al. 2018. Haplotype sharing provides insights into fine-scale population history and disease in Finland. *Am. J. Hum. Genet.* 102(5):760–75
18. Lim ET, Würtz P, Havulinna AS, Palta P, Tukiainen T, et al. 2014. Distribution and medical impact of loss-of-function variants in the Finnish founder population. *PLOS Genet.* 10(7):e1004494
19. Quintana-Murci L. 2016. Understanding rare and common diseases in the context of human evolution. *Genome Biol.* 17(1):225
20. Crow JF, Kimura M. 1970. *An Introduction to Population Genetics Theory*. Minneapolis, MN: Burgess. 1st ed.
21. Ohta T, Gillespie JH. 1996. Development of neutral and nearly neutral theories. *Theor. Popul. Biol.* 49(2):128–42
22. Kimura M, Maruyama T, Crow JF. 1963. The mutation load in small populations. *Genetics* 48:1303–12
23. Henn BM, Botigué LR, Bustamante CD, Clark AG, Gravel S. 2015. Estimating the mutation load in human genomes. *Nat. Rev. Genet.* 16(6):333–43
24. Menozzi P, Piazza A, Cavalli-Sforza L. 1978. Synthetic maps of human gene frequencies in Europeans. *Science* 201(4358):786–92
25. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38(8):904–9
26. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, et al. 2008. Genes mirror geography within Europe. *Nature* 456(7218):98–101
27. Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, et al. 2011. Demographic history and rare allele sharing among human populations. *PNAS* 108(29):11983–88
28. Bryc K, Durand EY, Macpherson JM, Reich D, Mountain JL. 2015. The genetic ancestry of African Americans, Latinos, and European Americans across the United States. *Am. J. Hum. Genet.* 96(1):37–53
29. Han E, Carbonetto P, Curtis RE, Wang Y, Granka JM, et al. 2017. Clustering of 770,000 genomes reveals post-colonial population structure of North America. *Nat. Commun.* 8:14238
30. Schraiber JG, Akey JM. 2015. Methods and models for unravelling human evolutionary history. *Nat. Rev. Genet.* 16(12):727–40
31. Marchini J, Howie B. 2010. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* 11(7):499–511
32. Pasaniuc B, Rohland N, McLaren PJ, Garimella K, Zaitlen N, et al. 2012. Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nat. Genet.* 44(6):631–35
33. Pickrell J. 2017. It is time to replace genotyping arrays with sequencing. *The Gencove Blog*, Aug 14. **https://medium.com/the-gencove-blog/it-is-time-to-replace-genotyping-arrays-with-sequencing-73535efa66ed**
34. Homburger JR, Neben CL, Mishne G, Zhou AY, Kathiresan S, Khera AV. 2019. Low coverage whole genome sequencing enables accurate assessment of common variants and calculation of genome-wide polygenic scores. *Genome Med.* 11:74
35. Visscher PM, Hill WG, Wray NR. 2008. Heritability in the genomics era—concepts and misconceptions. *Nat. Rev. Genet.* 9(4):255–66
36. Schoech AP, Jordan DM, Loh P-R, Gazal S, O'Connor LJ, et al. 2019. Quantification of frequency-dependent genetic architectures in 25 UK Biobank traits reveals action of negative selection. *Nat. Commun.* 10(1):790

37. Loh P-R, Tucker G, Bulik-Sullivan BK, Vilhjálmsson BJ, Finucane HK, et al. 2015. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* 47(3):284–90
38. Zhou W, Zhao Z, Nielsen JB, Fritsche LG, LeFaive J, et al. 2019. Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts. bioRxiv 583278. **https://doi.org/10.1101/583278**
39. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, et al. 2017. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* 45(D1):D896–901
40. Ripke S, Neale BM, Corvin A, Walters JTR, Farh K-H, et al. 2014. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* 511(7510):421–27
41. Xue A, Wu Y, Zhu Z, Zhang F, Kemper KE, et al. 2018. Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. *Nat. Commun.* 9(1):2941
42. Suzuki K, Akiyama M, Ishigaki K, Kanai M, Hosoe J, et al. 2019. Identification of 28 new susceptibility loci for type 2 diabetes in the Japanese population. *Nat. Genet.* 51(3):379–86
43. Mahajan A, Wessel J, Willems SM, Zhao W, Robertson NR, et al. 2018. Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes. *Nat. Genet.* 50(4):559–71
44. Ellinghaus D, Jostins L, Spain SL, Cortes A, Bethune J, et al. 2016. Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci. *Nat. Genet.* 48(5):510–18
45. Liu JZ, van Sommeren S, Huang H, Ng SC, Alberts R, et al. 2015. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* 47(9):979–86
46. van der Harst P, Verweij N. 2018. Identification of 64 novel genetic loci provides an expanded view on the genetic architecture of coronary artery disease. *Circ. Res.* 122(3):433–43
47. Klarin D, Damrauer SM, Cho K, Sun YV, Teslovich TM, et al. 2018. Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. *Nat. Genet.* 50(11):1514–23
48. Hoffmann TJ, Theusch E, Haldar T, Ranatunga DK, Jorgenson E, et al. 2018. A large electronic-health-record-based genome-wide study of serum lipids. *Nat. Genet.* 50(3):401–13
49. Abul-Husn NS, Kenny EE. 2019. Personalized medicine and the power of electronic health records. *Cell* 177(1):58–69
50. Verbanck M, Chen C-Y, Neale B, Do R. 2018. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat. Genet.* 50(5):693–98
51. Benner C, Spencer CCA, Havulinna AS, Salomaa V, Ripatti S, Pirinen M. 2016. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* 32(10):1493–501
52. Wang G, Sarkar A, Carbonetto P, Stephens M. 2018. A simple new approach to variable selection in regression, with application to genetic fine-mapping. bioRxiv 501114. **https://doi.org/10.1101/501114**
53. Holmes MV, Ala-Korpela M, Smith GD. 2017. Mendelian randomization in cardiometabolic disease: challenges in evaluating causality. *Nat. Rev. Cardiol.* 14(10):577–90
54. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, et al. 2015. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* 47(11):1228–35
55. Martin AR, Daly MJ, Robinson EB, Hyman SE, Neale BM. 2019. Predicting polygenic risk of psychiatric disorders. *Biol. Psychiatry* 86(2):97–109
56. Torkamani A, Wineinger NE, Topol EJ. 2018. The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* 19(9):581–90
57. Lambert SA, Abraham G, Inouye M. 2019. Towards clinical utility of polygenic risk scores. *Hum. Mol. Genet.* 28(R2):R133–42
58. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. 2019. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* 51(4):584–91
59. Bulik-Sullivan BK, Loh P-R, Finucane HK, Ripke S, Yang J, et al. 2015. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* 47(3):291–95

60. Brown BC, Asian Genet. Epidemiol. Netw. Type 2 Diabetes Consort., Ye CJ, Price AL, Zaitlen N. 2016. Transethnic genetic-correlation estimates from summary statistics. *Am. J. Hum. Genet.* 99(1):76–88
61. Gurdasani D, Carstensen T, Fatumo S, Chen G, Franklin CS, et al. 2019. Uganda genome resource enables insights into population history and genomic discovery in Africa. *Cell* 179(4):984–1002.e36
62. 1000 Genomes Proj. Consort. 2015. A global reference for human genetic variation. *Nature* 526(7571):68–74
63. Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, et al. 2017. Human demographic history impacts genetic risk prediction across diverse populations. *Am. J. Hum. Genet.* 100(4):635–49
64. Sohail M, Maier RM, Ganna A, Bloemendal A, Martin AR, et al. 2019. Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. *eLife* 8:e39702
65. Berg JJ, Harpak A, Sinnott-Armstrong N, Joergensen AM, Mostafavi H, et al. 2019. Reduced signal for polygenic adaptation of height in UK Biobank. *eLife* 8:e39725
66. Kerminen S, Martin AR, Koskela J, Ruotsalainen SE, Havulinna AS, et al. 2019. Geographic variation and bias in the polygenic scores of complex diseases and traits in Finland. *Am. J. Hum. Genet.* 104(6):1169–81
67. Ott J, Wang J, Leal SM. 2015. Genetic linkage analysis in the age of whole-genome sequencing. *Nat. Rev. Genet.* 16(5):275–84
68. Lefebvre S, Bürglen L, Reboullet S, Clermont O, Burlet P, et al. 1995. Identification and characterization of a spinal muscular atrophy-determining gene. *Cell* 80(1):155–65
69. Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, et al. 2010. Exome sequencing identifies *MLL2* mutations as a cause of Kabuki syndrome. *Nat. Genet.* 42(9):790–93
70. Bamshad MJ, Nickerson DA, Chong JX. 2019. Mendelian gene discovery: fast and furious with no end in sight. *Am. J. Hum. Genet.* 105(3):448–55
71. Chong JX, Buckingham KJ, Jhangiani SN, Boehm C, Sobreira N, et al. 2015. The genetic basis of Mendelian phenotypes: discoveries, challenges, and opportunities. *Am. J. Hum. Genet.* 97(2):199–215
72. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, et al. 2018. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* 46(D1):D1062–67
73. Whiffin N, Walsh R, Ing AY, Barton PJR, Funke B, Cook SA. 2017. Using high-resolution variant frequencies to empower clinical genome interpretation. *Genet. Med.* 19(10):1151–58
74. Cummings BB, Marshall JL, Tukiainen T, Lek M, Donkervoort S, et al. 2017. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Sci. Transl. Med.* 9(386):eaal5209
75. Deciphering Dev. Disord. Study. 2017. Prevalence and architecture of de novo mutations in developmental disorders. *Nature* 542(7642):433–38
76. Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, et al. 2012. Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* 485(7397):242–45
77. Kosmicki JA, Samocha KE, Howrigan DP, Sanders SJ, Slowikowski K, et al. 2017. Refining the role of de novo protein-truncating variants in neurodevelopmental disorders by using population reference samples. *Nat. Genet.* 49(4):504–10
78. Satterstrom FK, Walters RK, Singh T, Wigdor EM, Lescai F, et al. 2018. ASD and ADHD have a similar burden of rare protein-truncating variants. bioRxiv 277707. **https://doi.org/10.1101/277707**
79. Satterstrom FK, Kosmicki JA, Wang J, Breen MS, De Rubeis S, et al. 2018. Novel genes for autism implicate both excitatory and inhibitory cell lineages in risk. bioRxiv 484113. **https://doi.org/10.1101/484113**
80. Abifadel M, Varret M, Rabès J-P, Allard D, Ouguerram K, et al. 2003. Mutations in *PCSK9* cause autosomal dominant hypercholesterolemia. *Nat. Genet.* 34(2):154–56
81. Cohen JC, Boerwinkle E, Mosley TH Jr., Hobbs HH. 2006. Sequence variations in *PCSK9*, low LDL, and protection against coronary heart disease. *N. Engl. J. Med.* 354(12):1264–72
82. Cohen J, Pertsemlidis A, Kotowski IK, Graham R, Garcia CK, Hobbs HH. 2005. Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in *PCSK9*. *Nat. Genet.* 37(2):161–65
83. Whiffin N, Armean IM, Kleinman A, Marshall JL, Minikel EV, et al. 2019. Human loss-of-function variants suggest that partial LRRK2 inhibition is a safe therapeutic strategy for Parkinson's disease. bioRxiv 561472. **https://doi.org/10.1101/561472**

84. King EA, Davis JW, Degner JF. 2019. Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. *PLOS Genet.* 15(12):e1008489
85. Ganna A, Satterstrom FK, Zekavat SM, Das I, Churchhouse C, et al. 2018. Quantifying the impact of rare and ultra-rare coding variation across the phenotypic spectrum. *Am. J. Hum. Genet.* 102(6):1204–11
86. Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, et al. 2011. Testing for an unusual distribution of rare variants. *PLOS Genet.* 7(3):e1001322
87. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. 2011. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* 89(1):82–93
88. Cirulli ET, White S, Read RW, Elhanan G, Metcalf WJ, et al. 2019. Genome-wide rare variant analysis for thousands of phenotypes in 54,000 exomes. bioRxiv 692368. **https://doi.org/10.1101/692368**
89. Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. 2013. Genic intolerance to functional variation and the interpretation of personal genomes. *PLOS Genet.* 9(8):e1003709
90. Samocha KE, Robinson EB, Sanders SJ, Stevens C, Sabo A, et al. 2014. A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* 46(9):944–50
91. Wang X, Goldstein DB. 2018. Enhancer redundancy predicts gene pathogenicity and informs complex disease gene discovery. bioRxiv 459123. **https://doi.org/10.1101/459123**
92. Whiffin N, Karczewski KJ, Zhang X, Chothani S, Smith MJ, et al. 2019. Characterising the loss-of-function impact of 5′ untranslated region variants in whole genome sequence data from 15,708 individuals. bioRxiv 543504. **https://doi.org/10.1101/543504**
93. Plenge RM. 2019. Priority index for human genetics and drug discovery. *Nat. Genet.* 51(7):1073–75
94. Fang H, ULTRA-DD Consort., De Wolf H, Knezevic B, Burnham KL, et al. 2019. A genetics-led approach defines the drug target landscape of 30 immune-related traits. *Nat. Genet.* 51(7):1082–91
95. Harper AR, Nayee S, Topol EJ. 2015. Protective alleles and modifier variants in human health and disease. *Nat. Rev. Genet.* 16(12):689–701
96. Nguyen PA, Born DA, Deaton AM, Nioi P, Ward LD. 2019. Phenotypes associated with genes encoding drug targets are predictive of clinical trial side effects. *Nat. Commun.* 10(1):1579
97. Plenge RM, Scolnick EM, Altshuler D. 2013. Validating therapeutic targets through human genetics. *Nat. Rev. Drug Discov.* 12(8):581–94
98. Bush WS, Moore JH. 2012. Genome-wide association studies. *PLOS Comput. Biol.* 8(12):e1002822
99. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, et al. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319(5866):1100–4
100. Nishino J, Ochi H, Kochi Y, Tsunoda T, Matsui S. 2018. Sample size for successful genome-wide association study of major depressive disorder. *Front. Genet.* 9:227
101. Marchini J, Cardon LR, Phillips MS, Donnelly P. 2004. The effects of human population structure on large genetic association studies. *Nat. Genet.* 36(5):512–17
102. Rommens JM, Iannuzzi MC, Kerem B, Drumm ML, Melmer G, et al. 1989. Identification of the cystic fibrosis gene: chromosome walking and jumping. *Science* 245(4922):1059–65
103. Riordan JR, Rommens JM, Kerem B, Alon N, Rozmahel R, et al. 1989. Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science* 245(4922):1066–73
104. Du K, Sharma M, Lukacs GL. 2005. The ΔF508 cystic fibrosis mutation impairs domain-domain interactions and arrests post-translational folding of CFTR. *Nat. Struct. Mol. Biol.* 12(1):17–25
105. Scriver JB, Waugh TR. 1930. Studies on a case of sickle-cell anaemia. *Can. Med. Assoc. J.* 23(3):375–80
106. Saiki RK, Scharf S, Faloona F, Mullis KB, Horn GT, et al. 1985. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* 230(4732):1350–54
107. Samson M, Libert F, Doranz BJ, Rucker J, Liesnard C, et al. 1996. Resistance to HIV-1 infection in Caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. *Nature* 382(6593):722–25
108. Gulick RM, Lalezari J, Goodrich J, Clumeck N, DeJesus E, et al. 2008. Maraviroc for previously treated patients with R5 HIV-1 infection. *N. Engl. J. Med.* 359(14):1429–41

109. Greggio E, Jain S, Kingsbury A, Bandopadhyay R, Lewis P, et al. 2006. Kinase activity is required for the toxic effects of mutant *LRRK2*/dardarin. *Neurobiol. Dis.* 23(2):329–41
110. West AB, Moore DJ, Biskup S, Bugayenko A, Smith WW, et al. 2005. Parkinson's disease-associated mutations in leucine-rich repeat kinase 2 augment kinase activity. *PNAS* 102(46):16842–47
111. TG HDL Work. Group Exome Seq. Proj., Natl. Heart Lung Blood Inst., Crosby J, Peloso GM, Auer PL, Crosslin DR, et al. 2014. Loss-of-function mutations in *APOC3*, triglycerides, and coronary disease. *N. Engl. J. Med.* 371(1):22–31
112. Jørgensen AB, Frikke-Schmidt R, Nordestgaard BG, Tybjærg-Hansen A. 2014. Loss-of-function mutations in *APOC3* and risk of ischemic vascular disease. *N. Engl. J. Med.* 371(1):32–41
113. Kanai Y, Lee WS, You G, Brown D, Hediger MA. 1994. The human kidney low affinity Na+/glucose cotransporter SGLT2. Delineation of the major renal reabsorptive mechanism for D-glucose. *J. Clin. Investig.* 93(1):397–404
114. Santer R, Kinner M, Schneppenheim R, Hillebrand G, Kemper M, et al. 2000. The molecular basis of renal glucosuria: mutations in the gene for a renal glucose transporter (SGLT2). *J. Inherit. Metab. Dis.* 23(Suppl. 1):178
115. Santer R, Calado J. 2010. Familial renal glucosuria and SGLT2: from a Mendelian trait to a therapeutic target. *Clin. J. Am. Soc. Nephrol.* 5(1):133–41
116. Hsia DS, Grove O, Cefalu WT. 2017. An update on sodium-glucose co-transporter-2 inhibitors for the treatment of diabetes mellitus. *Curr. Opin. Endocrinol. Diabetes Obes.* 24(1):73–79
117. Verma D, Särndahl E, Andersson H, Eriksson P, Fredrikson M, et al. 2012. The Q705K polymorphism in NLRP3 is a gain-of-function alteration leading to excessive interleukin-1β and IL-18 production. *PLOS ONE* 7(4):e34977
118. Verma D, Lerm M, Blomgran Julinder R, Eriksson P, Söderkvist P, Särndahl E. 2008. Gene polymorphisms in the NALP3 inflammasome are associated with interleukin-1 production and severe inflammation: relation to common inflammatory diseases? *Arthritis Rheum.* 58(3):888–94
119. Dendrou CA, Cortes A, Shipman L, Evans HG, Attfield KE, et al. 2016. Resolving *TYK2* locus genotype-to-phenotype differences in autoimmunity. *Sci. Transl. Med.* 8:363ra149
120. Burke JR, Cheng L, Gillooly KM, Strnad J, Zupa-Fernandez A, et al. 2019. Autoimmune pathways in mice and humans are blocked by pharmacological stabilization of the TYK2 pseudokinase domain. *Sci. Transl. Med.* 11(502):eaaw1736
121. Diogo D, Bastarache L, Liao KP, Graham RR, Fulton RS, et al. 2015. *TYK2* protein-coding variants protect against rheumatoid arthritis and autoimmunity, with no evidence of major pleiotropic effects on non-autoimmune complex traits. *PLOS ONE* 10(4):e0122271

Annual Review of Biomedical Data Science

Volume 3, 2020

Contents

Errata

An online log of corrections to *Annual Review of Biomedical Data Science* articles may be found at http://www.annualreviews.org/errata/biodatasci