Cell

Leading Edge
Review

Genomic Analysis in the Age of Human Genome Sequencing

Tuuli Lappalainen,[1,2,*] Alexandra J. Scott,[3,4,5] Margot Brandt,[1,2] and Ira M. Hall[3,4,5,*]
[1]New York Genome Center, New York, NY, USA
[2]Department of Systems Biology, Columbia University, New York, NY, USA
[3]McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO, USA
[4]Department of Medicine, Washington University School of Medicine, St. Louis, MO, USA
[5]Department of Genetics, Washington University School of Medicine, St. Louis, MO, USA
*Correspondence: tlappalainen@nygenome.org (T.L.), ihall@wustl.edu (I.M.H.)
https://doi.org/10.1016/j.cell.2019.02.032

Affordable genome sequencing technologies promise to revolutionize the field of human genetics by enabling comprehensive studies that interrogate all classes of genome variation, genome-wide, across the entire allele frequency spectrum. Ongoing projects worldwide are sequencing many thousands—and soon millions—of human genomes as part of various gene mapping studies, biobanking efforts, and clinical programs. However, while genome sequencing data production has become routine, genome analysis and interpretation remain challenging endeavors with many limitations and caveats. Here, we review the current state of technologies for genetic variant discovery, genotyping, and functional interpretation and discuss the prospects for future advances. We focus on germline variants discovered by whole-genome sequencing, genome-wide functional genomic approaches for predicting and measuring variant functional effects, and implications for studies of common and rare human disease.

Introduction

The development of high-throughput sequencing technologies has revolutionized human genetics and genomics. For the first time, widespread use of whole-genome sequencing (WGS) allows detection of a full range of common and rare genetic variants of different types across almost the entire genome, which facilitates rare disease research and clinical applications, and can improve common disease discovery and annotation of the causal variants. Now that hundreds of thousands of genomes have been sequenced worldwide, we are at the start of a new era where WGS will be a predominant technology for genetic analysis. This is a fundamental change compared to previous decades of human genetic studies that have relied on genetic markers that are indirect proxies of other genetic variants in the surrounding region or sequencing data only from the exonic regions of the genome.

Functional interpretation of variants discovered by WGS is an important component of human genetics studies and is essential for revealing the effects of variants on traits. Genome-wide functional genomics assays now allow for increasingly accurate detection, characterization, and prediction of the molecular effects of variants. However, since these effects reflect the full complexity of genome function, our understanding of which is incomplete, much remains to be discovered regarding variant molecular effects and their potential for impacting higher-level organismal phenotypes.

In this Review, we discuss approaches, advances, and future prospects for genetic variant discovery, genotyping, and functional interpretation. We focus on germline variants discovered by WGS and genome-wide functional genomic approaches for analysis of functional effects of these variants. These are foundational building blocks for the discovery and interpretation of genetic effects on rare and common human diseases and traits (Figure 1).

Human Genome Sequencing

WGS Technologies

The first aim of a typical WGS study is to create a high-quality map of genome variation for the samples of interest. This crucial step lays the foundation for all downstream analyses aimed at genome interpretation and genetic discovery because variants that are not accurately discovered and genotyped will not be directly assessed in trait-focused analyses. The methods used to map genome variation depend heavily on the sequencing technology and depth of coverage obtained.

There are currently three general WGS strategies (Figure 2): (1) short-read WGS using the Illumina technology, which currently yields paired-end ~150 bp reads with low error rates in the range of ~0.1%–0.5%; (2) long-read WGS using single-molecule technologies from Pacific Biosciences (PacBio) or Oxford Nanopore Technologies (ONT), which yield 10–100 kb reads—and occasionally much longer—with high error rates in the range of ~10%–15%; and (3) linked-read WGS using the technology from 10X Genomics, which generates barcoded Illumina short-reads from longer molecules (e.g., ~50 kb). Due to considerations of cost, ease of use, and accuracy, the overwhelming majority of human genetics studies employ short-read WGS using the Illumina HiSeq or NovaSeq

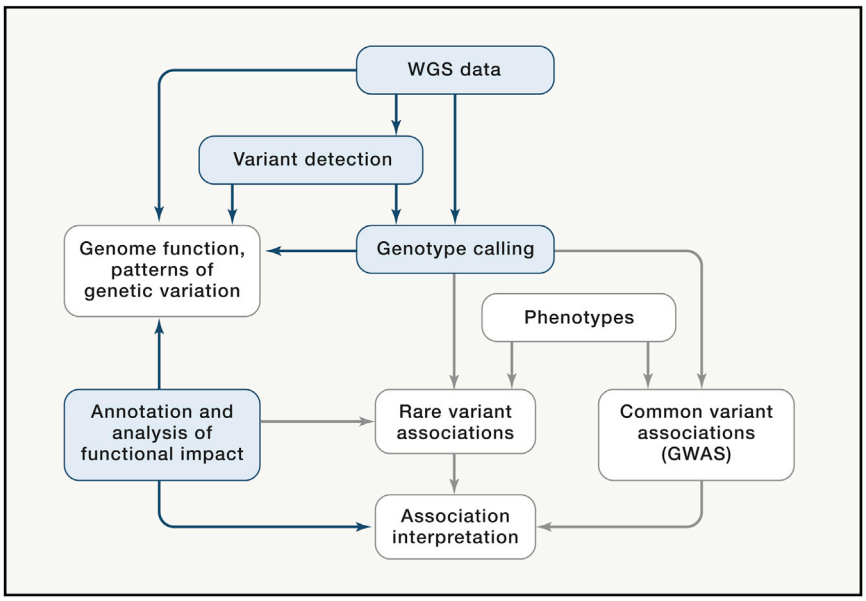

Figure 1. The General Framework of Genome Analysis in Studies of Human Phenotypes
Areas discussed in this Review highlighted in blue.

platform, and we therefore focus primarily on analysis of this data type.

An important consideration in the design of WGS studies is the desired level of coverage. To distinguish variants from errors, each base in the genome must be sequenced multiple times from randomly sampled DNA molecules. Deeper coverage improves variant detection sensitivity and also improves accuracy by allowing for more sophisticated filtering schemes. In general, family-based or n = 1 rare disease studies target deeper coverage (>30x) to ensure robust detection of rare or *de novo* heterozygous variants. Larger-scale complex trait studies may target somewhat lower coverage (e.g., >20x) to increase sample size for a given budget, while still allowing for sensitive rare variant detection. Early groundbreaking studies (e.g., Auton et al., 2015) employed low-coverage (<10x) WGS to reduce sequencing costs, but this approach fails to detect many rare variants and is no longer common. However, for complex trait mapping studies focused primarily on common variants, a powerful strategy is to maximize sample size by pursuing ultra-low-coverage (<1x) sequencing combined with variant imputation to infer missing genotypes (Pasaniuc et al., 2012). Although the optimal coverage model depends on the goals of the study, in practice, most current WGS studies are employing deep WGS (>20x).

Alignment and Data Processing

Since high-quality *de novo* assembly is not possible from short reads, standard WGS analysis pipelines align reads to the reference genome and map variants relative to the reference (Figure 2). Most modern pipelines use BWA-MEM (Li, 2013) for alignment and a combination of tools for subsequent processing. Although these methods are now well established, there are still several areas of innovation. Performance improvements and reductions in the alignment file size (Hsi-Yang Fritz et al., 2011; Regier et al., 2018) are important from the standpoint of efficiency and cost, especially in population-scale WGS studies involving thousands to millions of genomes. Additionally, differences in tools, parameters, or reference genome versions between different datasets affect variant calls and genotypes and introduce batch effects in downstream analyses. This issue is especially troublesome for large-scale trait association studies where subtle genotyping biases can grossly inflate false positives and where reprocessing of large datasets to achieve harmonization would require much time and expense. Data compatibility is also extremely important for small-scale studies that aim to accurately compare variant calls with public databases such as gnomAD (Karczewski et al., 2019). A recent multi-center effort established a model for implementing "functionally equivalent" pipelines, which are now in use at many genome centers worldwide, that alleviate batch effects and enable data sharing (Regier et al., 2018).

Genetic Variant Classes

Single-Nucleotide Variants and Small Insertion/Deletion Variants

Single-nucleotide variants (SNVs) and small insertion/deletion variants (indels) (<50 bp) comprise the vast majority of variants in the human population (Table 1). There are ~3–4 million SNVs and ~0.4–0.5 million indels apparent in a typical comparison of one human versus the reference, and the dbSNP catalog (build 151) has over 660 million SNVs and indels from diverse sequencing studies. While the vast majority of this huge number of variants have no functional impact at the molecular or phenotypic level, every genome has >100 protein truncating variants (PTVs) that introduce a premature stop codon, >20 of which are rare in the human population and potentially deleterious (Lek et al., 2016). Nonsynonymous or missense SNVs or in-frame indels lead to amino acid changes, which can be entirely benign or cause a severe disease. Finally, these variants can affect gene regulation by affecting transcriptional and posttranscriptional regulatory elements. Fundamentally, for an SNP or small indel to have an effect on gene regulation, a sequence-specific regulator whose activity is differentially affected by the two alleles is needed—at least at some point during development. These include (for example) transcription and splicing factors that bind to specific DNA motifs, as well as noncoding regulatory RNAs such as miRNAs.

These small variants are the easiest class of variants to detect from short-read data. In general, SNV/indel detection algorithms scan the reference genome in search of collections of aligned reads that exhibit mismatches, insertions, or deletions in a

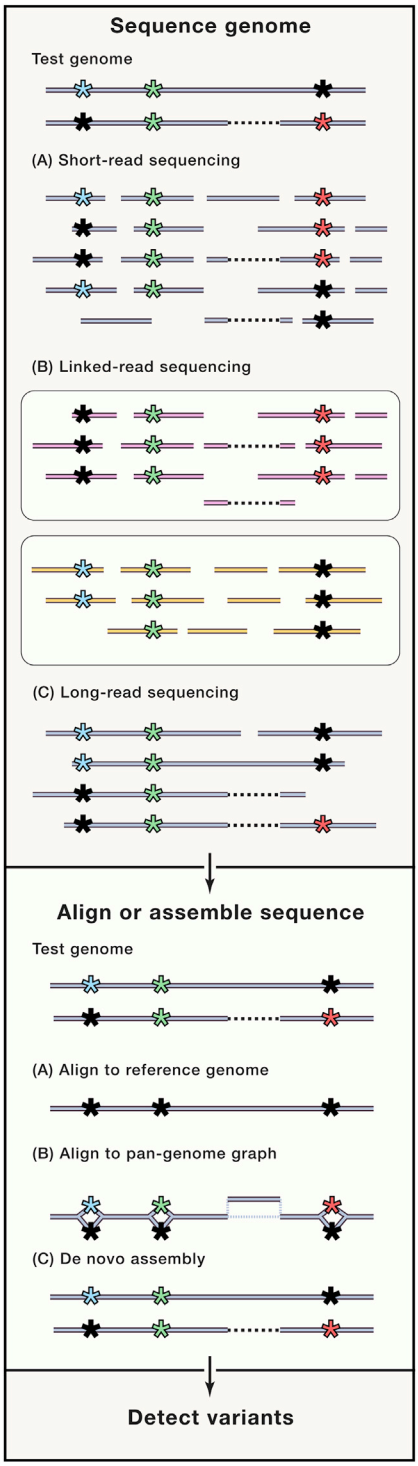

Figure 2. Overview of Genome Sequencing and Variant Detection Approaches

The experimentally sequenced "test" genome contains two heterozygous SNVs, each located on a different chromosome (blue and red stars), one homozygous SNV (green stars), and a heterozygous deletion (dashed line). Reference alleles are represented by solid lines and black stars. The pan-genome graph representation at right requires prior knowledge of all shown variants.

manner that suggests germline variation rather than sequencing or alignment error. Existing widely used tools (e.g., DePristo et al., 2011; Garrison and Marth, 2012) are highly effective in the ~72% of the genome that is unique and allows for accurate read alignment, with levels of sensitivity and specificity that exceed 99.5% for SNVs and 95% for indels (Regier et al., 2018; Zook et al., 2014). However, there is much room for improvement at larger and repeat-containing indels that confound read alignment. Some tools combine reference-based variant detection with local haplotype assembly to improve indel calling and phase nearby variants (DePristo et al., 2011; Garrison and Marth, 2012). About 8.5% of the genome is considered extremely difficult for SNV/indel calling due to the presence of segmental duplications and/or high-copy repeats that cause short-read misalignment (Regier et al., 2018). These regions include some clinically relevant multi-copy genes, and the poor detection of variants in these genes is a key weakness of short-read WGS. This is difficult to overcome algorithmically but will improve substantially as read lengths increase, allowing for more accurate alignment to the reference.

Structural Variation

Structural variation (SV) is a diverse form of genome variation ≥50 bp in size that includes copy number variants (CNVs), rearrangements, and mobile element insertions (MEIs). SVs are few in number compared to SNVs and indels (Table 1) but have more severe consequences on average due to their size. SVs can exert functional effects by changing gene dosage, disrupting gene function (similar to PTVs), or rearranging regulatory elements and/or genes to alter genomic context. Unsurprisingly, extremely large variants that delete or duplicate many genes or even entire chromosomes typically have drastic phenotypic effects and are not observed in most individuals. Smaller and more prevalent forms of SV typically affect only one or a few genes or lie within noncoding regions. Although SVs account for merely ~0.2% of total variants, recent WGS-based studies have estimated that they account for 3%–7% of common variants with *cis*-acting effects on gene expression, a much larger fraction of rare expression-altering variants, and 4%–12% of high-impact coding alleles (Abel et al., 2018; Chiang et al., 2017).

SV is recognized to be the most difficult form of variation to detect reliably from short-read data. Different variant classes require distinct algorithmic approaches (reviewed in (Alkan et al., 2011)), and SVs are enriched in repetitive elements that confound short-read mapping. A typical human genome has ~10,000 SVs that are detectable from short-read WGS data (Table 1), and >20,000 are detectable by long-read WGS (Audano et al., 2019; Chaisson et al., 2018), where the difference is primarily due to small and repetitive variants. There is a plethora of SV mapping tools but only a few general approaches. Overall, the most accurate and high-resolution approach is "breakpoint mapping." This method relies on direct detection of novel sequence junctions that are not present in the reference genome using a combination of read-pair alignments, split-read alignments, and/or local assembly. Popular tools combine multiple signals across populations of samples (e.g., Chen et al., 2016; Handsaker et al., 2011; Layer et al., 2014; Rausch et al., 2012). In theory, this approach can detect any SV whose breakpoints are relatively unique, which includes ~75% of the SVs

Table 1. The Landscape of Human Genome Variation

Variant Class		Subclass, Other Term(s)	Size	Num. / genome (Illumina WGS)	Arrays	Short-Read WGS	Long-Read WGS
Single Nucleotide Variation (SNV)		point mutation; substitution	1 bp	3.5×10^6	XX	XXX	XX
Small Insertion/Deletion Variation (indel)		insertion; deletion; complex indel	1-49 bp	4.5×10^5	XX	XX	X
Structural Variation (SV)	copy number variation (CNV)	deletion	≥50 bp	5,000	X	XX	XXX
		duplication (tandem, interspersed)		1,000	X	XX	XXX
		multi-allelic CNV; tandem repeats		450	X	XX	XX
	insertion	novel, templated or repeat insertion		1,500	-	X	XXX
	balanced rearrangement	inversion		40	-	XX	XXX
		reciprocal translocation	inter-chrom	0.001	-	XX	XXX
	complex genomic rearrangement	complex SV; chromothripsis	>1 mb	0.01	-	XX	XXX
	extremely large copy number variant	aneuploidy; chrom. abnormality	>1 mb	0.01	XXX	XXX	XXX
	retrogene insertion	retroduplication; retrocopy	gene coding length	10	-	XX	XXX
	mobile element insertion (MEI)	SINE; LINE; SVA	0.3-7 kb	2,000	-	X	XXX
Tandem Repeat Variation	short tandem repeat (STR)	microsatellite; simple sequence repeat	1-6 bp (repeat unit)	1×10^5	-	X	XXX
	variable number tandem repeat (VNTR)	minisatellite	7-49 bp (repeat unit)	unknown	-	X	XX
	centromeric & heterochromatic repeats	satellite DNA (α, β, 1-3)	various	unknown	-	-	XX

The major variant classes are shown at left, with subclasses and other terms shown in adjacent columns. Note that the precise terms and size definitions used for different variant classes varies in the literature, especially for the tandem repeat classes shown at bottom – we used common yet non-overlapping definitions. Shown at right is the relative utility of microarray, short-read WGS and long-read WGS technologies for detecting each class of variation, on a scale of 0-3, where "–" indicates a near-complete inability to detect that variant class, and 3 indicates that the technology is highly effective. Note that microarray technologies are generally only able to detect SNVs and indels known from prior sequencing studies. For each variant class we include a rough estimate of the number of variants detectable per human genome using Illumina short-read WGS. Note that these numbers will vary based on ancestry and methods. In particular, the numbers shown for SVs and STRs are highly approximate and depend heavily on the tools, sequencing depth and filtering methods employed. To derive the numbers for various SV classes, we assumed that 10,000 total SVs were detectable by Illumina WGS, and derived the relative contribution of each variant class based on the combined knowledge from studies cited in this review, as well as our own unpublished observations.

detectable by Illumina WGS. A key strength of this approach is that SV breakpoints are mapped to high resolution—usually <100 bp and often to a single base—which greatly facilitates downstream functional interpretation.

However, some SV breakpoints cannot be captured directly from short-read alignments because they are embedded within (or composed of) repeats that confound read mapping. Notable examples are recurrent CNVs formed by non-allelic homologous recombination that underlie many human disorders. However, larger such CNVs (>1 kb) can be detected by read-depth analysis. This method yields similar information as array-based approaches and has similar challenges, including poor sensitivity and high false discovery rate (FDR) at smaller CNVs (<10 kb), low-resolution breakpoint prediction, and artifactual fragmentation of large CNVs into many smaller CNV calls, which complicates functional interpretation. Only a modest number (5%–10%) of CNVs are detectable solely by read-depth analysis. However, these also tend to be among the largest CNVs, are enriched in genes, and are often not well tagged by SNVs. Read-depth analysis is therefore crucial for human genetics studies that aim to be comprehensive (Handsaker et al., 2015; Sudmant et al., 2015a).

Given the diverse and complementary approaches, it is not surprising that many studies employ compendium strategies that combine the results of multiple tools (e.g., as in Mills et al., 2011). Multi-algorithm approaches inevitably outperform single tools in terms of sensitivity, but weaknesses include increased FDR, increased compute time and cost, and the complexity of merging and adjudicating conflicting variant calls to create a consensus set. Thus, although compendium approaches are clearly superior, there are significant practical obstacles to their efficient and effective use in human genetics studies.

The architectural diversity of SV also poses challenges. For example, ~5% of SVs are "complex" variants with multiple adjacent or intertwined breakpoints, the structure and consequences of which are often difficult to infer (reviewed in Quinlan and Hall, 2012). Most complex germline SVs are small, but extreme forms can involve multiple chromosomes or distant loci. Other non-canonical SV classes requiring specialized methods include "retrogene insertions" derived from the action of retroelement machinery on processed transcripts, leading to insertion of coding sequences lacking introns (Sudmant et al., 2015b), and insertions of novel sequence not found in the reference genome (Kidd et al., 2008; Sherman et al., 2019).

Repetitive Variant Classes

The detection of variants involving high-copy repeats is difficult due to challenges in accurate alignment and requires even more specialized approaches than SVs in general. Mobile element insertions (MEIs) caused by retrotransposition are a relatively common form of variation, with >2,000 detectable MEIs in a typical genome (Sudmant et al., 2015b). Although MEIs do not generally appear to be a major source of causal disease variants, there are notable examples (reviewed in Kazazian and Moran, 2017), and MEIs have the potential to disrupt genes and regulatory elements at insertion sites and can serve as alternative promoters. Current MEI detection algorithms extract candidate MEI-containing reads based on reference genome alignments and then re-align them to a library of consensus mobile element sequences to obtain higher quality evidence.

Short tandem repeats (STRs, also known as microsatellites) are typically defined as repetitive arrays with repeat unit of 1–6 bp. STRs are extremely abundant (Willems et al., 2014) and highly polymorphic due to a high mutation rate. Coding STRs have been linked to >40 monogenic disorders typically due to amino acid repeat expansions (Mirkin, 2007), and non-coding STRs have been reported to account for 10%–15% of common variant *cis* heritability of gene expression (Gymrek et al., 2016). Various tools have been developed to detect STRs from short-read data, traditionally for shorter STRs (Gymrek et al., 2012) but recently also for longer STRs that include most pathogenic loci (Dashnow et al., 2018; Dolzhenko et al., 2017; Mousavi et al., 2018).

Variable number tandem repeats (VNTRs, also known as minisatellites) are repetitive arrays with repeat unit 7–49 bp. Most VNTRs are noncoding with some having strong regulatory effects on neighboring genes (Pugliese et al., 1997), and a few coding VNTRs are known to cause Mendelian diseases (Kirby et al., 2013). Despite noteworthy examples, VNTR detection has received scant attention, and few specialized algorithms exist (Bakhtiari et al., 2018). Larger satellite repeat arrays at centromeres and heterochromatic regions are even less well studied. For this reason, the prevalence and functional importance of VNTRs and satellite repeat variation remain unclear.

It is worth noting that the distinction between repetitive variant classes are often arbitrary, and definitions are not consistently applied in the literature nor in variant catalogs. We expect this to be clarified as sequencing and variant calling methods improve.

Joint Variant Detection and Genotyping in Population Studies

The above discussion has largely focused on the task of *discovering* genetic variants on a per-sample basis. However, this is not sufficient for studies that involve multiple individuals in a family, cohort, biobank, or case-control design. The desired output of variant calling is a "squared-off" file (typically in VCF format) that describes all the variants discovered in the study, with an accurate genotype for all individuals at all variable sites. To achieve this, we cannot simply paste together the results of per-sample variant calling. There is a consensus in the field that *joint* variant calling methods that co-analyze raw WGS data across the full collection of sites, and individuals in a single probabilistic model yields higher quality variant sites and genotypes. Systematic quality control (QC) of the resulting VCF can improve genotyping accuracy at common variants, reduce FDR at rare variants, and alleviate batch effects among data from different centers, instruments, and sample collection sites. For these reasons, virtually all current human genetics studies employ joint variant calling.

However, joint variant calling poses significant technical challenges for large-scale studies. Current approaches employ complex workflows involving (1) parallelized per-sample calling, (2) merging of variant calls across samples, (3) re-assessment of raw (or nearly raw) data to produce maximally sensitive genotypes for all sites in all samples, and finally, (4) filtering, tuning, and QC on the entire dataset. Even the most efficient of current

pipelines do not easily scale to >10,000 samples. Technical challenges are especially pronounced for SVs and other recalcitrant variant types, where cross-sample merging is more difficult (due to spatial imprecision) and where WGS-based genotyping methods are far less accurate, scalable, and affordable. Thus, although several prior large-scale studies have managed to create high-quality maps of diverse variant types (as noted in prior sections), these have been heroic efforts that required customized algorithms and clever *ad hoc* methods. Creating comprehensive and accurate variant maps for large-scale WGS studies remains a complex, laborious, and expensive process that is accessible to few groups, and significant advances will be required before it is routine. In particular, this remains a key barrier for efforts to understand the contribution of SVs and repetitive variant classes to complex traits.

It is important to recognize that the value of population-scale joint variant calling extends beyond the immediate benefits for trait mapping studies because harmonized variant callsets are the basis for databases such as gnomAD (Karczewski et al., 2019) that enable community-wide variant interpretation efforts and are likely to be the foundational resource for next-generation data sharing platforms implemented at public and private data repositories.

Future Improvements in Algorithms and Data for Genome Analysis

Pan-genome References and Analysis Tools

A limitation of the current WGS analysis paradigm is the core reliance on the reference genome. Although the GRCh38 reference is extremely high quality in most respects, it still has gaps and errors at repetitive and structurally diverse regions. An even bigger problem for genetic studies is that the reference human genome is inherently not able to represent the diversity our species' collective "human genome," i.e., 12 billion haploid genomes in the entire population. The reference is a mosaic haploid representation of multiple individuals, such that a collection of haplotypes has been "smashed" together in an unpredictable manner. There have been laudable efforts to append "alternative loci" to GRCh38 in highly diverse regions (e.g., MHC, KIR, CYP2D6) (Church et al., 2015). However, the impact of this effort has been modest because the alternate loci in GRCh38 are fairly limited in number and scope, and most tools and pipelines do not make use of them.

One unfortunate consequence of the mosaic haploid reference is that reference-based variant calling methods are slightly more accurate for individuals whose "local ancestry" at a given locus is more closely related to that of the reference genome. A second consequence is that poor alignment severely affects detection of many relevant alleles in regions of high genomic diversity; prominent and clinically important examples are the MHC locus, KIR genes, CYP2D6, olfactory gene clusters, and the ancient inversion at 17q21.31 (Stefansson et al., 2005). Finally, it is impossible to find variants in "novel sequences" that are simply not present in the reference genome (Kidd et al., 2008; Sherman et al., 2019). Although most such regions are small, non-genic, and/or highly repetitive, it would nonetheless be preferable to assess them in WGS studies.

Due to these limitations, momentum has built around the idea of creating a next-generation reference "pan-genome" resource that represents all relatively common DNA sequences and alleles in the human population. Within a few years, we expect high-quality haplotype-resolved diploid assemblies to be available for at least 500 ancestrally diverse humans. This resource will also provide a set of highly accurate genomes that can be used as a benchmarking dataset to improve short-read analysis tools. Even more importantly, these genomes allow completely new designs for more effective short-read analysis strategies that overcome many of the limitations described above.

Transitioning to a pan-genome reference will require development of new algorithms and pipelines for analyzing short-read WGS data. Pan-genome tools typically employ a graph-based data structure, referred to as "genome graphs" or "variant graphs," to represent allelic diversity (Figure 2) (Paten et al., 2017). This is a major change relative to traditional tools. There has been substantial progress in this area during recent years, and new tools have been developed to allow for efficient genome graph representation, short-read alignment, and variant genotyping at specific loci such as HLA (Dilthey et al., 2015) and CYP2D6 (Numanagić et al., 2015), and more recently genome-wide (Garrison et al., 2018). Remaining challenges include encoding of linkage disequilibrium (LD) information, repetitive elements and complex variants, and providing user-friendly visualization tools. Although much more work is needed before graph-based WGS analysis pipelines are used routinely for human genetics research—let alone clinical applications—these methods have taken an important first step.

Affordable and High-Quality Long-Read WGS

Most of the technical difficulties discussed in the above sections are caused by the inherent difficulty of interpreting short-read alignments to a complex and repetitive reference. These issues would largely disappear if we were able to affordably and accurately sequence human genomes with long-read technologies, enabling *de novo* haplotype-resolved genome assembly without reliance on prior reference data. It is still unclear when this day will come. However, there have been notable advances in the past year driven by PacBio and ONT platforms, with further improvements expected in the near future. ONT recently made significant advances in pushing read-lengths out to unprecedented length, making high-quality assemblies possible (when combined with Illumina data) (Jain et al., 2018). PacBio has developed a new circular consensus sequencing approach that promises highly accurate 10–15 kb synthetic reads. Both platforms have recently achieved and are forecasting steep cost decreases. However, both still have major challenges that need to be overcome, first and foremost the extremely high error rate (> 10%) of raw data and resulting high cost to achieve accurate genome-wide variant detection. At present, long reads must be complemented with Illumina short-read data, primarily because the small indel error rate is still too high. In contrast, long-read variation maps excel for SV (Audano et al., 2019; Chaisson et al., 2018; Jain et al., 2018; Sedlazeck et al., 2018).

In the interim, the long-molecule linked-read technology from 10X Genomics offers an affordable alternative; however, thus far, the performance improvements for reference-based variant detection are rather modest (Marks et al., 2018) and are limited to specific regions and variant types (i.e., complex rearrangements) (Spies et al., 2017). Approaches for variant detection

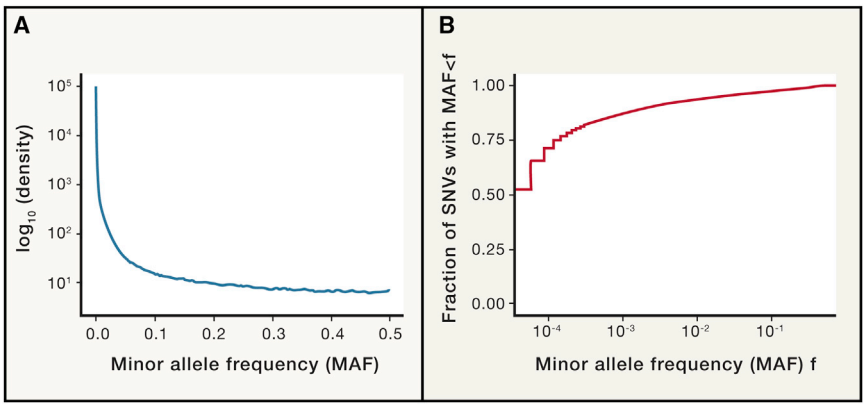

Figure 3. Allele Frequency of SNVs from the gnomAD Database

The gnomAD database can be found at https://gnomad.broadinstitute.org.

(A) Density plot showing the minor allele frequency (MAF) distribution, known as the "site frequency spectrum."

(B) Cumulative distribution function of the site frequency spectrum, showing the fraction of variants (y axis) with a frequency smaller than a given MAF (x axis). Note that the leftmost data point represents "singleton" variants present in only one person. These plots are based on a randomly sampled subset of ~19 million SNVs from gnomAD version 2.0, which in total includes ~188 million SNVs from 15,496 genomes.

via *de novo* haplotype-resolved assembly of linked-reads show promise (Weisenfeld et al., 2017) but will require further work to match the performance of reference-based approaches. One area where linked reads may prove useful in the short term is haplotype phasing, which is a key aspect of genome analysis. Statistical population-based phasing methods are inaccurate for rare and ultra-rare variants, which can be phased only when sequencing reads link these variants to other nearby variants. This is much more effective with synthetic or true long reads (Figure 2). Fully resolved haplotypes are crucial for understanding compound effects of multiple variants affecting the same gene, and for diagnosing compound heterozygotes for gene disrupting variants.

Sooner or later, long-read technologies will improve substantially. With significant improvements to the error rate, it will be possible to identify the overwhelming majority of genetic variants via long-read alignment to a pan-genome graph. With significant improvements to error rate *and* read length, it should be possible to routinely create clinical-grade diploid genome assemblies.

The Spectrum of Genetic Variation in Human Populations

Variant allele frequency is a key factor to consider in the analysis of genome variation. As a consequence of human population history, which includes ancient bottlenecks and recent expansion, the vast majority of variants in the human population are rare (Figure 3). Traditionally, "rare" variants are defined as those with minor allele frequency (MAF) < 1%, "common" variants have MAF > 5%, and "low-frequency" variants are those in between. The definition of "ultra-rare" varies and is often used to denote "singleton" variants identified in only one person from a large study; here, we define ultra-rare variants as MAF <0.01% (<1 in 10,000 chromosomes).

Although most variants in the *population* are ultra-rare, most variants identified in an *individual* are common (>95%). This is explained by the fact that most inter-individual variation is due to ancient polymorphisms that arose early during human history when the effective population size was small and that are now present in all major ancestry groups (albeit often at different frequencies). However, 50–100 new mutations occur each generation, and during the course of many recent generations of population growth, a very large number of ultra-rare variants have accumulated in the population. Each individual sequenced in a large study contributes more such variants. In general, rarer variants are more difficult to analyze because there are fewer observations to rely upon during population-level variant detection and trait association.

The vast majority of genetic variants are non-functional and neutral and have no discernible phenotypic effects at the individual level. Most variants with strong phenotypic effects are deleterious, and most deleterious variants are rare due to the effects of purifying selection (Karczewski et al., 2019). Indeed, the allele frequency of a genetic variant—reflecting the allele's age and selective forces acting on it—is probably the single most powerful proxy for its potential phenotypic effects. For example, studies of Mendelian or early-onset diseases with strong effects on reproductive fitness typically use allele frequency as the primary variant prioritization criterion and focus solely on rare and *de novo* variants. Early groundbreaking resources for allele frequency estimation were based on low-coverage WGS (Auton et al., 2015) or exome data (Lek et al., 2016). At present, the most accurate estimates come from the gnomAD (Karczewski et al., 2019) and Bravo (from NHLBI TOPMed) (https://bravo.sph.umich.edu/freeze5/hg38) databases, which together are based on deep WGS data from ~75,000 individuals. These databases are invaluable but have limitations related to the diversity of ancestry groups and variant classes that they represent. Neither contains structural variants or repetitive variants, and larger indels are poorly represented. A recent effort from the NHGRI CCDG program created an SV map from >17,000 genomes that will help in this regard (Abel et al., 2018). We expect that future efforts from these and other projects will improve representation of all variant classes, across a more diverse set of ancestry groups, in ever-larger WGS datasets.

Functional Interpretation in Genetic Study Designs

Interpretation of molecular effects of genetic variants is an essential part of genetic analysis because the discovery of disease genes often requires prioritization of variants that are predicted to have a strong impact on gene function (Figure 1). In Mendelian disease studies as well as in genetic diagnosis, variant impact prediction and prioritization is used to define the most likely one or two disease-causing mutation(s) from a much larger set of candidates. This is necessary in both research

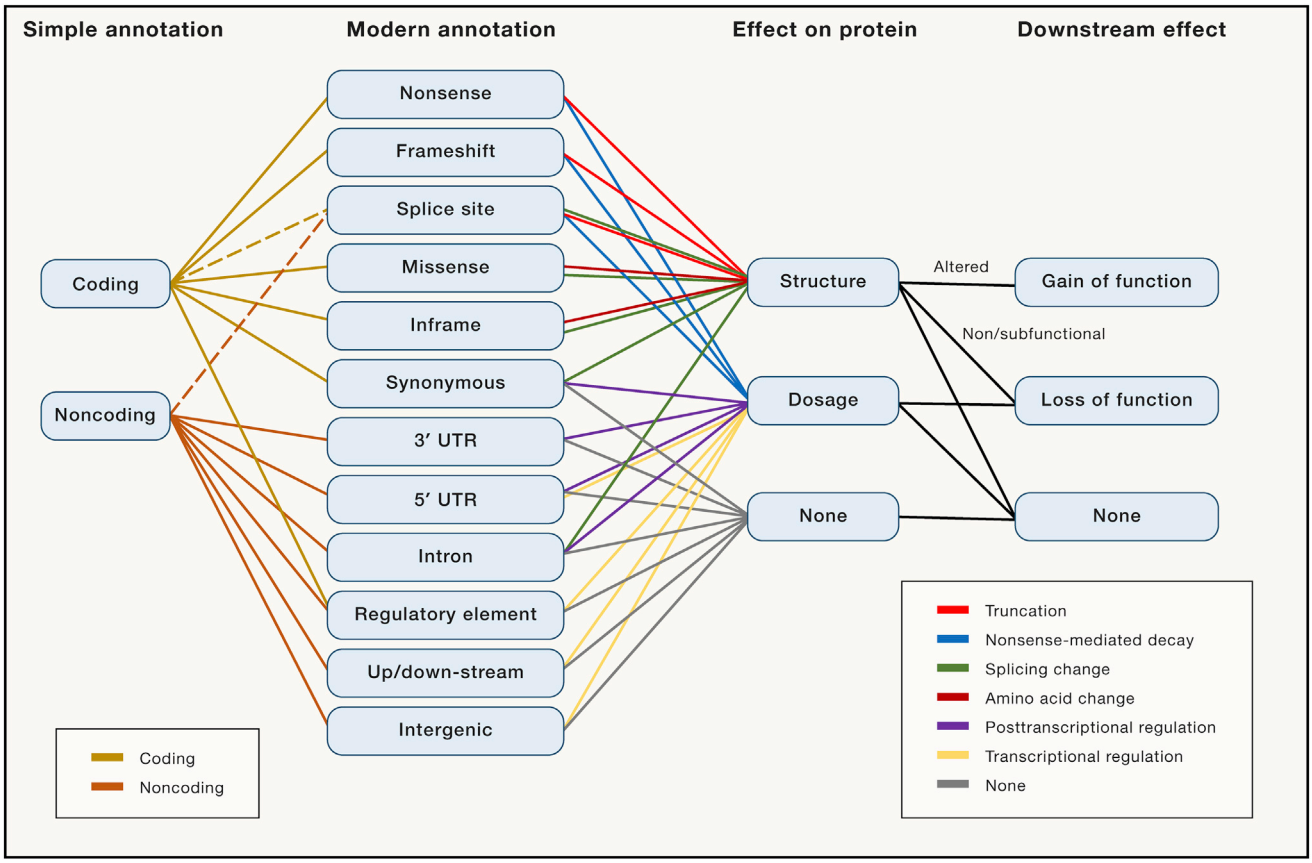

Figure 4. Functional Annotation and Downstream Consequences of SNVs and Small Indels

Annotation of genetic variants according to their type, position, and downstream effects for SNVs and indels. The annotations include the most commonly used ones, and the potential effects on protein are shown here in an approximate sense, asking the question "if a variant with a given annotation has any effect on gene function, what are the most likely processes." The downstream effect indicates the change on protein's function in the cell. This illustration highlights the complexity challenge of understanding even the proximal molecular effects of diverse types of genetic variants and building biologically and medically meaningful understanding of their downstream effects.

and diagnostic settings. In rare variant association studies of common disease or other complex traits, variant allele frequencies and impact predictions are used to prioritize or "weight" variants in the association tests and to group them by gene or another genomic unit that are then tested for phenotype association. This approach has been crucial for studying common diseases that are heavily influenced by rare and *de novo* variation, such as autism and schizophrenia. In contrast, the discovery of common variant associations to complex diseases by genome-wide association studies is typically performed without functional priors on genetic variants, although transcriptome-wide association studies rely on genetic predictors on gene expression (Gamazon et al., 2015; Gusev et al., 2016). However, even in classical GWAS, understanding the molecular effects of associated variants is an essential next step to understand cellular processes that underlie disease risk.

The small subset of variants that are functional exert their influence on organismal phenotype in two general ways: (1) by inducing *qualitative* changes in the composition of gene products by altering the sequence of proteins or noncoding RNAs or (2) by inducing *quantitative* changes in protein or RNA abundance (Figures 4 and 5). These proximal molecular changes may then exert downstream effects on cellular or physiological pathways that ultimately contribute to organismal phenotypes. Exhaustive exploration of these mechanisms or downstream pathway-level effects is beyond the scope of this Review. Here, we focus on practical interpretation of variants associated with human disease with large-scale functional genomic data that are used to (1) predict functional effects of variants based on a reference functional annotation of the genome and (2) associate genetic variation with empirical molecular measurements in a population sample of individuals (Figure 6).

Functional Annotation and Prediction of Genetic Variant Effects

Qualitative and Quantitative Effects

The most straightforward annotation of genetic variants (Figure 6A) is done based on their allele frequency and their position either in the coding or noncoding part of the genome. These have traditionally been analyzed by distinct research communities. The rare and Mendelian disease community has generally focused on rare, strong-effect gene-disrupting coding

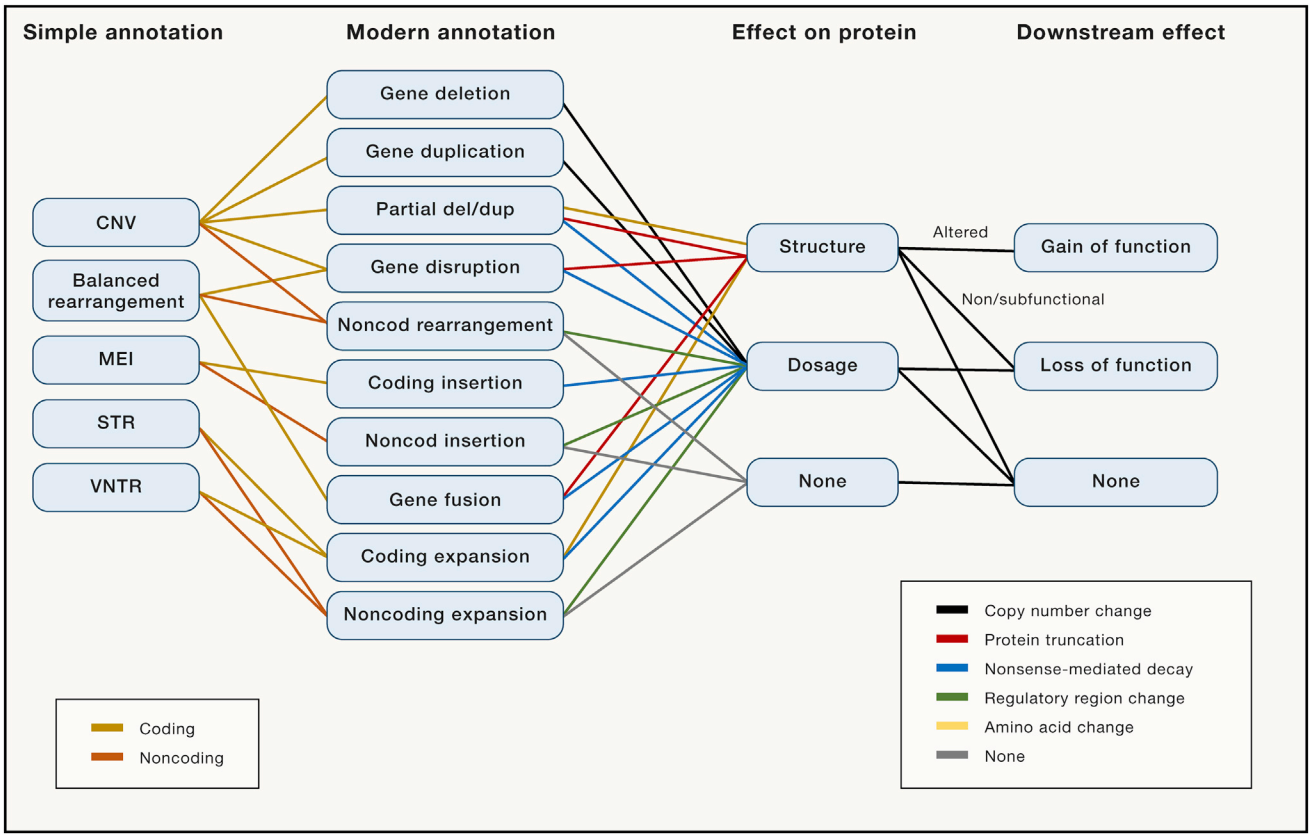

Figure 5. Functional Annotation and Downstream Consequences of Structural Variants
Annotation of genetic variants according to their type, position, and downstream effects for structural variants.

variants discovered by exome sequencing. In contrast, the common disease community has generally focused on common variants genotyped by SNP arrays and analysis of noncoding variants with likely regulatory effects driving GWAS associations.

However, this dichotomy of variant annotation and genetic research is currently being challenged from multiple fronts. First, the transition to WGS as a universal technology will detect variants regardless of location and frequency, making the coding-noncoding and rare-common distinctions unnecessary from a technical perspective. Furthermore, several studies have demonstrated a more complex, mixed genetic architecture of both common and rare disease even though optimal study designs for traits of different genetic architectures remain a matter of debate (Castel et al., 2018; Freund et al., 2018; Niemi et al., 2018; Weiner et al., 2017). Finally, a more refined understanding of functional effects of genetic variants challenges the simple coding-noncoding classification that often carries implicit assumptions that coding variants cause gene knockouts or disrupt protein structure, whereas noncoding variants fine-tune transcription levels. In reality, both coding and noncoding variants can have qualitative and quantitative effects of varying magnitudes on both protein structure and dosage (Figures 4 and 5). Aiming to annotate variants by their predicted functional effects, rather than genomic position, will ultimately have better biological justification and downstream applicability. For example, noncoding variants with strong effects on gene expression should have similar loss-of-function consequences as coding variants triggering nonsense-mediated decay of the same gene.

Predicting Variant Effects

The aim to assess variants' qualitative and quantitative effects depends upon accurate prediction of molecular effects of diverse types of genetic variants, which is currently one of the most actively pursued challenges in human genomics research. For coding variants, the genetic code and the high-quality human gene annotation provide a relatively straightforward means to accurately predict amino acid changes and premature stop codons that lead to either truncated protein or transcript degradation via nonsense-mediated decay. However, predicting whether an amino acid change actually changes protein structure and function is extremely difficult. Diverse computational prediction models use both protein structure and conservation data (Glusman et al., 2017), but sparsity of experimental validation data remains a challenge (Raraigh et al., 2018). Splicing changes may dramatically alter protein structure or introduce a premature stop codon. Variants in the 2-basepair canonical splice sites lead to disrupted splicing. Putative splicing changes can be caused by variants in the surrounding splicing motif, novel splice sites introduced by exonic or intronic variants, and variants in splicing enhancer and repressor sequences (Barash et al., 2010; Rivas et al., 2015; Savisaar and Hurst, 2018;

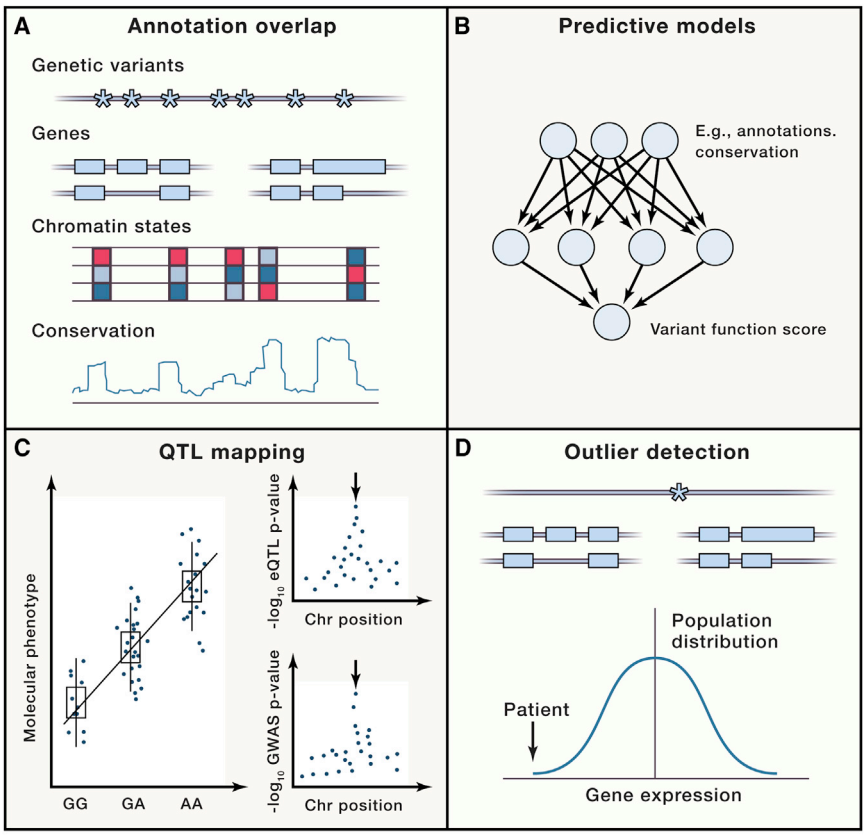

Figure 6. Illustration of the Approaches to Interpret Molecular Effects of Genetic Variants

(A) Straightforward overlap with tissue-specific annotations of genes and regulatory elements as well as genome constraint scores.

(B) Predictive machine learning models of variant function.

(C) Mapping of common variant associations to molecular phenotypes (left side) and their colocalization with GWAS associations.

(D) Interrogating if a rare variant carrier is also an outlier with respect to a proximal molecular phenotype.

Vaz-Drago et al., 2017). Several algorithms have been developed to predict variant effects on splicing, but their performance is far from perfect, especially further away from the canonical splice site (for example, Xiong et al., 2015; Yeo and Burge, 2004). However, a recent machine learning approach has promising performance for even more distant variants that alter splicing (Jaganathan et al., 2019).

Gene dosage can be affected in even more diverse ways than protein structure, with complex transcriptional and posttranscriptional regulation. For SNVs and small indels, a challenge is that assays for measuring sequence-specific regulators—such as transcription or splicing factors and RNA targeting or secondary structure—are not particularly scalable and robust, and binding motif predictions are not very accurate. Thus, many regulatory elements of the genome are annotated by measuring chromatin accessibility and histone modifications that reflect chromatin state. These assays have been applied at scale in multiple cell types by many major projects such as ENCODE and Epigenomics Roadmap (ENCODE Project Consortium, 2012; Kundaje et al., 2015). Enrichment of disease-associated variants in genes and noncoding annotations with differential activity across cell types has provided valuable insights into the cell types and states most relevant to disease (Farh et al., 2015; Finucane et al., 2015; Trynka et al., 2013). However, inference of genetic variant effects on regulatory element function and gene expression remains a major challenge. While this is a highly active area of research and many recent reports utilizing machine learning approaches report promising results (Figure 6B) (Alipanahi et al., 2015; Lee et al., 2015; Zhou et al., 2018), strong performance in practical applications is still lacking. Our inability to accurately read the genome's regulatory code is a persistent challenge for WGS studies.

Efforts to predict variant effects on molecular function are often complemented with different measures of evolutionary conservation or selective constraint in the human population. These are very powerful proxies for overall fitness effects of genetic changes, providing data as to whether the potential molecular change actually affects organismal function. These metrics are essential components of composite methods such as CADD and fitCons (Gulko et al., 2015; Kircher et al., 2014) (among others) that integrate annotation and constraint data to prioritize variants. Furthermore, estimates of constraint at the level of genes or their parts (Petrovski et al., 2015; Samocha et al., 2014) have been essential for prioritizing genes that are particularly sensitive to functional genetic perturbations for rare variant association studies.

There are few tools designed to predict the effects of structural variants, and those that exist use fairly rudimentary strategies (Ganel et al., 2017; McLaren et al., 2016). Although interpretation of whole-gene deletions and duplications is straightforward, predicting the effects of balanced rearrangements and smaller exonic deletions and duplications is not. As yet, there is not a rigorous statistical framework for predicting the impact of noncoding SVs, although simple approaches that summarize per-base impact scores (e.g., from CADD) at SV breakpoints and within affected genomic segments have proven effective (Ganel et al., 2017).

Molecular Phenotypes to Characterize Functional Effects of Variants

Given that our ability to *predict* molecular effects of genetic variants is imperfect, especially for regulatory variants, the natural complementary approach is to empirically *measure* molecular effects of genetic variants in individuals that carry them. This has been made possible by scalable and affordable assays to

measure gene expression, splicing, chromatin state, and other molecular traits genome wide in hundreds to thousands of samples. Integrated with WGS data, these measurements enable direct inference into how genetic variants affect gene regulation.

Common variants and eQTL mapping

Of genetic risk loci discovered by GWAS, up to 90% are in noncoding regions, raising the challenge of functional interpretation of their proximal regulatory mechanisms, target genes, and relevant cell types of activity (Visscher et al., 2017). This challenge was a major motivator for expression quantitative trait loci (eQTL) studies, which are genetic association analyses where the phenotype is expression level of a gene (Figure 6C). This has been pursued by consortium projects such as GTEx (GTEx Consortium et al., 2017), and common *cis*-regulatory genetic associations have now been discovered for nearly every gene in the human genome (GTEx Consortium et al., 2017; Võsa et al., 2018). These studies have also provided insights into the functional importance of SVs and STRs (Chiang et al., 2017; Gymrek et al., 2016). *Trans*-eQTLs between distant genes and variants have been substantially more challenging to find, but they provide particularly valuable insights into variant and gene effects on regulatory networks (GTEx Consortium et al., 2017; Võsa et al., 2018). Transcriptome-wide association studies have provided a new approach to use genetically predicted gene expression or splicing changes from large eQTL datasets to find new associations to traits and diseases (Gamazon et al., 2015; Gusev et al., 2016). These general approaches can also be applied to other molecular traits such as splicing, epigenomic features, and protein levels (Li et al., 2016; Sun et al., 2018; Waszak et al., 2015).

The enrichment of molecular QTL associations among GWAS loci indicates their ability to inform on regulatory mechanisms of complex traits (Gamazon et al., 2018; Nicolae et al., 2010; Ongen et al., 2017). However, similar to other association studies, QTL analysis does not provide direct distinction of the causal variant versus variants in linkage disequilibrium with the causal variant. This complicates the interpretation and applications of eQTLs. When integrated with GWAS data, spurious overlap must be excluded by statistical colocalization analysis to estimate whether the causal variant of GWAS and eQTL signals is shared (Giambartolomei et al., 2014; Hormozdiari et al., 2016; Wen et al., 2017), providing a stronger hypothesis that the gene expression change indicated by the eQTL is causally related to the GWAS trait. The complexity of LD patterns and the wealth of eQTL effects on different genes in different tissues and cell types may make it difficult to identify specific functional hypotheses from eQTL data (Chun et al., 2017). However, molecular QTL data have provided likely causal mechanisms for thousands of GWAS loci by implicating specific epigenomic features, proximal gene expression or splicing, and/or downstream network effects (Gamazon et al., 2018; GTEx Consortium et al., 2017; Hormozdiari et al., 2018; Ongen et al., 2017; Võsa et al., 2018). Importantly, eQTL evidence indicates that the relevant disease gene is usually not the nearest gene to the genetic locus (GTEx Consortium et al., 2017), highlighting the importance of eQTL and other functional follow-up to interpret GWAS loci and to provide valuable information about regulatory effects of common genetic variants.

Rare Variant Analysis via Molecular Trait Outliers

While the eQTL approach is powerful for characterizing common regulatory variants or loci, association approaches cannot be used for very rare variants. In WGS studies of rare variant effects, as discussed above, variants need to be prioritized and grouped according to their predicted functional impact, but these predictions are often inaccurate especially for variants affecting transcriptional regulation or splicing. A complementary approach is to see if rare variant appears to have a molecular effect that is unusual compared to the general population (Li et al., 2017) (Figure 6D). Indeed, it has been shown that diagnostic rate in Mendelian disease can be improved by transcriptome sequencing that facilitates detection of ultra-rare variants with expression or splicing effects that are difficult to predict from WGS data alone (Cummings et al., 2017; Fresard et al., 2018; Kremer et al., 2017). Further approaches for WGS and RNA-seq integration in rare variant interpretation are under development. Molecular phenotype data can therefore be an important complement to WGS in interpretation and prioritization of rare genetic variants.

Future Challenges in Variant Interpretation

One of the fundamental challenges in both functional prediction of genetic variants and population-scale molecular analyses is the tissue, cell type, and cell-state specificity of molecular function of the genome. Functional annotation does not cover all these contexts—including developmental stage and environmental conditions—and thus has incomplete information of cell type and state specific transcripts and transcriptional and postranscriptional regulatory elements. The current eQTL catalogs are vast but mostly based on bulk tissue data often from whole blood. Emerging analyses of specific cell types and cell states—for example, in the form of stimulated immune cells—has highlighted the importance of extending these analyses to understand context-specific effects of genetic variants and their contribution to disease (Fairfax et al., 2014; Zhernakova et al., 2017). This is also relevant for understanding how applicable the information obtained from *in vitro* cell line assays is for modeling genetic effects in the complex *in vivo* cellular environment of the human body. With the development of single-cell technology, we anticipate significant progress in understanding context-specific genetic effects during the next few years.

Additionally, the current toolkit of genomic assays that are robust and scalable do not cover all aspects of genome function. In particular, the difficulty of measuring transcription and splicing factor binding and thus improving their motif predictions contributes to our inability to read the genome's regulatory code and to accurately predict genetic variant effects. One technology that will aid in solving this problem is large-scale experimental testing of genetic effects on gene expression and splicing (Soemedi et al., 2017; Tewhey et al., 2016; van Arensbergen et al., 2018), which will also create essential testing and training data for machine-learning approaches.

Finally, most eQTL studies are still limited to relatively small sample sizes appropriate mainly for cis-eQTL mapping, with scant coverage of many ancestry groups. It may be informative to consider lessons learned from the GWAS field, where—after modest success with smaller sample sizes—meaningful insights

and applications emerged after a major scale-up in power. Extending the sample size and diversity in eQTL studies has the potential to uncover trans-eQTLs with insights to causal network effects and gene function, improve fine-mapping of causal regulatory variants, and provide important insights into the effects of rare variants uncovered by WGS studies. We envision that complementing large-scale studies of human genomes and phenomes with molecular traits will add an important layer for interpretation of the genome and its links to human traits.

Conclusions

This is an exciting time in the field of human genetics and genomics. In recent years, we have witnessed historic advances in genome sequencing and analysis technologies, which are enabling the creation of ever larger and richer datasets, and the pursuit of ever more creative analyses. At present, the key challenges that need to be overcome for WGS studies to realize their full potential are comprehensive variant discovery and accurate prediction of functional effects of variants.

A solid framework and roadmap appears to be in place to meet the first challenge. We expect that further development of long-molecule sequencing technologies, high-quality haplotype catalogs, and pan-genome analysis methods will in combination enable reasonably comprehensive variant detection at the vast majority of functional genomic loci. However, making these approaches scalable and affordable enough for the massive sample sizes needed in human genetic studies will take considerable effort in years to come.

The challenges in the prediction of variant effects are more complex, and the roadmap and timeline are less clear. However, there is broad agreement that many different approaches will be required and that they need to be applied to diverse systems ranging from cellular, organoid, and animal models to human samples. Analysis of increasingly large and diverse cell types and human populations is essential. We envision that improvements in experimental methods, creation of large and comprehensive datasets, and algorithm development will go hand in hand to enable direct interrogation of variant effects and increasingly accurate computational prediction methods. These approaches will be complemented by information gained from large-scale WGS projects that provide not only catalogs of variants but also high-resolution maps of selective constraint in coding and non-coding regions and deeper genetic association data for rare and common disease.

In combination, knowledge of genome variation and its functional effects is an essential foundation for understanding human biology and improving human health. To ultimately achieve these goals, genomics will need to be integrated with population-scale phenotyping and clinical implementation.

ACKNOWLEDGMENTS

This work was supported by the NIH/NHGRI Center for Common Disease Genomics program (UM1HG008853 and UM1HG008901) and by a Mr. and Mrs. Spencer T. Olin Fellowship for Women in Graduate Study (to A.J.S.).

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

Abel, H.J., Larson, D.E., Chiang, C., Das, I., Kanchi, K., Layer, R.M., Neale, B.M., Salerno, W.J., Reeves, C., Buyske, S., et al. (2018). Mapping and characterization of structural variation in 17,795 deeply sequenced human genomes. bioRxiv. https://doi.org/10.1101/508515.

Alipanahi, B., Delong, A., Weirauch, M.T., and Frey, B.J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat. Biotechnol. *33*, 831–838.

Alkan, C., Coe, B.P., and Eichler, E.E. (2011). Genome structural variation discovery and genotyping. Nat. Rev. Genet. *12*, 363–376.

Audano, P.A., Sulovari, A., Graves-Lindsay, T.A., Cantsilieris, S., Sorensen, M., Welch, A.E., Dougherty, M.L., Nelson, B.J., Shah, A., Dutcher, S.K., et al. (2019). Characterizing the Major Structural Variant Alleles of the Human Genome. Cell *176*, 663–675.e19.

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., and Abecasis, G.R.; 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74.

Bakhtiari, M., Shleizer-Burko, S., Gymrek, M., Bansal, V., and Bafna, V. (2018). Targeted genotyping of variable number tandem repeats with adVNTR. Genome Res. *28*, 1709–1719.

Barash, Y., Calarco, J.A., Gao, W., Pan, Q., Wang, X., Shai, O., Blencowe, B.J., and Frey, B.J. (2010). Deciphering the splicing code. Nature *465*, 53–59.

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, and Biospecimen Collection Source Site—NDRI, et al.. (2017). Genetic effects on gene expression across human tissues. Nature *550*, 204–213.

Castel, S.E., Cervera, A., Mohammadi, P., Aguet, F., Reverter, F., Wolman, A., Guigo, R., Iossifov, I., Vasileva, A., and Lappalainen, T. (2018). Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk. Nat. Genet. *50*, 1327–1334.

Chaisson, M.J.P., Sanders, A.D., Zhao, X., Malhotra, A., Consortium, H.G.S.V., Mills, R.E., Marschall, T., Korbel, J.O., Eichler, E.E., and Lee, C. (2018). Multi-platform discovery of haplotype-resolved structural variation in human genomes. bioRxiv. https://doi.org/10.1101/193144.

Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., Cox, A.J., Kruglyak, S., and Saunders, C.T. (2016). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics *32*, 1220–1222.

Chiang, C., Scott, A.J., Davis, J.R., Tsang, E.K., Li, X., Kim, Y., Hadzic, T., Damani, F.N., Ganel, L., Montgomery, S.B., et al.; GTEx Consortium (2017). The impact of structural variation on human gene expression. Nat. Genet. *49*, 692–699.

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.

Church, D.M., Schneider, V.A., Steinberg, K.M., Schatz, M.C., Quinlan, A.R., Chin, C.S., Kitts, P.A., Aken, B., Marth, G.T., Hoffman, M.M., et al. (2015). Extending reference assembly models. Genome Biol. *16*, 13.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Cummings, B.B., Marshall, J.L., Tukiainen, T., Lek, M., Donkervoort, S., Foley, A.R., Bolduc, V., Waddell, L.B., Sandaradura, S.A., O'Grady, G.L., et al.; Genotype-Tissue Expression Consortium (2017). Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. Sci. Transl. Med. *9*, eaal5209.

Dashnow, H., Lek, M., Phipson, B., Halman, A., Davis, M., Lamont, P., Laing, N., MacArthur, D., and Oshlack, A. (2018). STRetch: detecting and discovering

pathogenic short tandem repeats expansions. bioRxiv. https://doi.org/10.1101/159228.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. *43*, 491–498.

Dilthey, A., Cox, C., Iqbal, Z., Nelson, M.R., and McVean, G. (2015). Improved genome inference in the MHC using a population reference graph. Nat. Genet. *47*, 682–688.

Dolzhenko, E., van Vugt, J.J.F.A., Shaw, R.J., Bekritsky, M.A., van Blitterswijk, M., Narzisi, G., Ajay, S.S., Rajan, V., Lajoie, B.R., Johnson, N.H., et al.; US–Venezuela Collaborative Research Group (2017). Detection of long repeat expansions from PCR-free whole-genome sequence data. Genome Res. *27*, 1895–1903.

Fairfax, B.P., Humburg, P., Makino, S., Naranbhai, V., Wong, D., Lau, E., Jostins, L., Plant, K., Andrews, R., McGee, C., and Knight, J.C. (2014). Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. Science *343*, 1246949.

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.

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.; ReproGen Consortium; Schizophrenia Working Group of the Psychiatric Genomics Consortium; RACI Consortium (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet. *47*, 1228–1235.

Fresard, L., Smail, C., Smith, K.S., Ferraro, N.M., Teran, N.A., Kernohan, K.D., Bonner, D., Li, X., Marwaha, S., Zappala, Z., et al. (2018). Identification of rare-disease genes in diverse undiagnosed cases using whole blood transcriptome sequencing and large control cohorts. boirXiv. https://doi.org/10.1101/408492.

Freund, M.K., Burch, K.S., Shi, H., Mancuso, N., Kichaev, G., Garske, K.M., Pan, D.Z., Miao, Z., Mohlke, K.L., Laakso, M., et al. (2018). Phenotype-Specific Enrichment of Mendelian Disorder Genes near GWAS Regions across 62 Complex Traits. Am. J. Hum. Genet. *103*, 535–552.

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

Gamazon, E.R., Segrè, A.V., van de Bunt, M., Wen, X., Xi, H.S., Hormozdiari, F., Ongen, H., Konkashbaev, A., Derks, E.M., Aguet, F., et al.; GTEx Consortium (2018). Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. Nat. Genet. *50*, 956–967.

Ganel, L., Abel, H.J., and Hall, I.M.; FinMetSeq Consortium (2017). SVScore: an impact prediction tool for structural variation. Bioinformatics *33*, 1083–1085.

Garrison, E., and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. arXiv, arXiv:1203.3907v2, https://arxiv.org/abs/12073907.

Garrison, E., Sirén, J., Novak, A.M., Hickey, G., Eizenga, J.M., Dawson, E.T., Jones, W., Garg, S., Markello, C., Lin, M.F., et al. (2018). Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nat. Biotechnol. *36*, 875–879.

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.

Glusman, G., Rose, P.W., Prlić, A., Dougherty, J., Duarte, J.M., Hoffman, A.S., Barton, G.J., Bendixen, E., Bergquist, T., Bock, C., et al. (2017). Mapping genetic variations to three-dimensional protein structures to enhance variant interpretation: a proposed framework. Genome Med. *9*, 113.

Gulko, B., Hubisz, M.J., Gronau, I., and Siepel, A. (2015). A method for calculating probabilities of fitness consequences for point mutations across the human genome. Nat. Genet. *47*, 276–283.

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

Gymrek, M., Golan, D., Rosset, S., and Erlich, Y. (2012). lobSTR: A short tandem repeat profiler for personal genomes. Genome Res. *22*, 1154–1162.

Gymrek, M., Willems, T., Guilmatre, A., Zeng, H., Markus, B., Georgiev, S., Daly, M.J., Price, A.L., Pritchard, J.K., Sharp, A.J., and Erlich, Y. (2016). Abundant contribution of short tandem repeats to gene expression variation in humans. Nat. Genet. *48*, 22–29.

Handsaker, R.E., Korn, J.M., Nemesh, J., and McCarroll, S.A. (2011). Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. Nat. Genet. *43*, 269–276.

Handsaker, R.E., Van Doren, V., Berman, J.R., Genovese, G., Kashin, S., Boettger, L.M., and McCarroll, S.A. (2015). Large multiallelic copy number variations in humans. Nat. Genet. *47*, 296–303.

Hormozdiari, F., van de Bunt, M., Segrè, A.V., Li, X., Joo, J.W.J., Bilow, M., Sul, J.H., Sankararaman, S., Pasaniuc, B., and Eskin, E. (2016). Colocalization of GWAS and eQTL Signals Detects Target Genes. Am. J. Hum. Genet. *99*, 1245–1260.

Hormozdiari, F., Gazal, S., van de Geijn, B., Finucane, H.K., Ju, C.J., 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.

Hsi-Yang Fritz, M., Leinonen, R., Cochrane, G., and Birney, E. (2011). Efficient storage of high throughput DNA sequencing data using reference-based compression. Genome Res. *21*, 734–740.

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.

Jain, M., Koren, S., Miga, K.H., Quick, J., Rand, A.C., Sasani, T.A., Tyson, J.R., Beggs, A.D., Dilthey, A.T., Fiddes, I.T., et al. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat. Biotechnol. *36*, 338–345.

Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2019). Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. bioRxiv. https://doi.org/10.1101/531210.

Kazazian, H.H., Jr., and Moran, J.V. (2017). Mobile DNA in Health and Disease. N. Engl. J. Med. *377*, 361–370.

Kidd, J.M., Cooper, G.M., Donahue, W.F., Hayden, H.S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F., et al. (2008). Mapping and sequencing of structural variation from eight human genomes. Nature *453*, 56–64.

Kirby, A., Gnirke, A., Jaffe, D.B., Barešová, V., Pochet, N., Blumenstiel, B., Ye, C., Aird, D., Stevens, C., Robinson, J.T., et al. (2013). Mutations causing medullary cystic kidney disease type 1 lie in a large VNTR in MUC1 missed by massively parallel sequencing. Nat. Genet. *45*, 299–303.

Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. *46*, 310–315.

Kremer, L.S., Bader, D.M., Mertes, C., Kopajtich, R., Pichler, G., Iuso, A., Haack, T.B., Graf, E., Schwarzmayr, T., Terrile, C., et al. (2017). Genetic diagnosis of Mendelian disorders via RNA sequencing. Nat. Commun. *8*, 15824.

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J., et al.; Roadmap Epigenomics Consortium (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.

Layer, R.M., Chiang, C., Quinlan, A.R., and Hall, I.M. (2014). LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. *15*, R84.

Lee, D., Gorkin, D.U., Baker, M., Strober, B.J., Asoni, A.L., McCallion, A.S., and Beer, M.A. (2015). A method to predict the impact of regulatory variants from DNA sequence. Nat. Genet. *47*, 955–961.

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al.; Exome Aggregation Consortium (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv, arXiv: 1303.3997v2, https://arxiv.org/abs/1303.3997.

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.

Li, X., Kim, Y., Tsang, E.K., Davis, J.R., Damani, F.N., Chiang, C., Hess, G.T., Zappala, Z., Strober, B.J., Scott, A.J., et al.; 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; Biospecimen Collection Source Site—RPCI; Biospecimen Core Resource—VARI; Brain Bank Repository—University of Miami Brain Endowment Bank; Leidos Biomedical—Project Management; ELSI Study; Genome Browser Data Integration &Visualization—EBI; Genome Browser Data Integration &Visualization—UCSC Genomics Institute, University of California Santa Cruz (2017). The impact of rare variation on gene expression across tissues. Nature *550*, 239–243.

Marks, P., Garcia, S., Martinez Barrio, A., Belhocine, K., Bernate, J., Bharadwaj, R., Bjornson, K., Catalanotti, C., Delaney, J., Fehr, A., et al. (2018). Resolving the Full Spectrum of Human Genome Variation using Linked-Reads. bioRxiv. https://doi.org/10.1101/230946.

McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biol. *17*, 122.

Mills, R.E., Walter, K., Stewart, C., Handsaker, R.E., Chen, K., Alkan, C., Abyzov, A., Yoon, S.C., Ye, K., Cheetham, R.K., et al.; 1000 Genomes Project (2011). Mapping copy number variation by population-scale genome sequencing. Nature *470*, 59–65.

Mirkin, S.M. (2007). Expandable DNA repeats and human disease. Nature *447*, 932–940.

Mousavi, N., Shleizer-Burko, S., and Gymrek, M. (2018). Profiling the genome-wide landscape of tandem repeat expansions. bioRxiv. https://doi.org/10.1101/361162.

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.

Niemi, M.E.K., Martin, H.C., Rice, D.L., Gallone, G., Gordon, S., Kelemen, M., McAloney, K., McRae, J., Radford, E.J., Yu, S., et al. (2018). Common genetic variants contribute to risk of rare severe neurodevelopmental disorders. Nature *562*, 268–271.

Numanagić, I., Malikić, S., Pratt, V.M., Skaar, T.C., Flockhart, D.A., and Sahinalp, S.C. (2015). Cypiripi: exact genotyping of CYP2D6 using high-throughput sequencing data. Bioinformatics *31*, i27–i34.

Ongen, H., Brown, A.A., Delaneau, O., Panousis, N.I., Nica, A.C., and Dermitzakis, E.T.; GTEx Consortium (2017). Estimating the causal tissues for complex traits and diseases. Nat. Genet. *49*, 1676–1683.

Pasaniuc, B., Rohland, N., McLaren, P.J., Garimella, K., Zaitlen, N., Li, H., Gupta, N., Neale, B.M., Daly, M.J., Sklar, P., et al. (2012). Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. Nat. Genet. *44*, 631–635.

Paten, B., Novak, A.M., Eizenga, J.M., and Garrison, E. (2017). Genome graphs and the evolution of genome inference. Genome Res. *27*, 665–676.

Petrovski, S., Gussow, A.B., Wang, Q., Halvorsen, M., Han, Y., Weir, W.H., Allen, A.S., and Goldstein, D.B. (2015). The Intolerance of Regulatory Sequence to Genetic Variation Predicts Gene Dosage Sensitivity. PLoS Genet. *11*, e1005492.

Pugliese, A., Zeller, M., Fernandez, A., Jr., Zalcberg, L.J., Bartlett, R.J., Ricordi, C., Pietropaolo, M., Eisenbarth, G.S., Bennett, S.T., and Patel, D.D. (1997). The insulin gene is transcribed in the human thymus and transcription levels correlated with allelic variation at the INS VNTR-IDDM2 susceptibility locus for type 1 diabetes. Nat. Genet. *15*, 293–297.

Quinlan, A.R., and Hall, I.M. (2012). Characterizing complex structural variation in germline and somatic genomes. Trends Genet. *28*, 43–53.

Raraigh, K.S., Han, S.T., Davis, E., Evans, T.A., Pellicore, M.J., McCague, A.F., Joynt, A.T., Lu, Z., Atalar, M., Sharma, N., et al. (2018). Functional Assays Are Essential for Interpretation of Missense Variants Associated with Variable Expressivity. Am. J. Hum. Genet. *102*, 1062–1077.

Rausch, T., Zichner, T., Schlattl, A., Stütz, A.M., Benes, V., and Korbel, J.O. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics *28*, i333–i339.

Regier, A.A., Farjoun, Y., Larson, D.E., Krasheninina, O., Kang, H.M., Howrigan, D.P., Chen, B.J., Kher, M., Banks, E., Ames, D.C., et al. (2018). Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. Nat. Commun. *9*, 4038.

Rivas, M.A., Pirinen, M., Conrad, D.F., Lek, M., Tsang, E.K., Karczewski, K.J., Maller, J.B., Kukurba, K.R., DeLuca, D.S., Fromer, M., et al.; GTEx Consortium; Geuvadis Consortium (2015). Human genomics. Effect of predicted protein-truncating genetic variants on the human transcriptome. Science *348*, 666–669.

Samocha, K.E., Robinson, E.B., Sanders, S.J., Stevens, C., Sabo, A., McGrath, L.M., Kosmicki, J.A., Rehnström, K., Mallick, S., Kirby, A., et al. (2014). A framework for the interpretation of de novo mutation in human disease. Nat. Genet. *46*, 944–950.

Savisaar, R., and Hurst, L.D. (2018). Exonic splice regulation imposes strong selection at synonymous sites. Genome Res. *28*, 1442–1454.

Sedlazeck, F.J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler, A., and Schatz, M.C. (2018). Accurate detection of complex structural variations using single-molecule sequencing. Nat. Methods *15*, 461–468.

Sherman, R.M., Forman, J., Antonescu, V., Puiu, D., Daya, M., Rafaels, N., Boorgula, M.P., Chavan, S., Vergara, C., Ortega, V.E., et al. (2019). Assembly of a pan-genome from deep sequencing of 910 humans of African descent. Nat. Genet. *51*, 30–35.

Soemedi, R., Cygan, K.J., Rhine, C.L., Wang, J., Bulacan, C., Yang, J., Bayrak-Toydemir, P., McDonald, J., and Fairbrother, W.G. (2017). Pathogenic variants that alter protein code often disrupt splicing. Nat. Genet. *49*, 848–855.

Spies, N., Weng, Z., Bishara, A., McDaniel, J., Catoe, D., Zook, J.M., Salit, M., West, R.B., Batzoglou, S., and Sidow, A. (2017). Genome-wide reconstruction of complex structural variants using read clouds. Nat. Methods *14*, 915–920.

Stefansson, H., Helgason, A., Thorleifsson, G., Steinthorsdottir, V., Masson, G., Barnard, J., Baker, A., Jonasdottir, A., Ingason, A., Gudnadottir, V.G., et al. (2005). A common inversion under selection in Europeans. Nat. Genet. *37*, 129–137.

Sudmant, P.H., Mallick, S., Nelson, B.J., Hormozdiari, F., Krumm, N., Huddleston, J., Coe, B.P., Baker, C., Nordenfelt, S., Bamshad, M., et al. (2015a). Global diversity, population stratification, and selection of human copy-number variation. Science *349*, aab3761.

Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M.H., et al.; 1000 Genomes Project Consortium (2015b). An integrated map of structural variation in 2,504 human genomes. Nature *526*, 75–81.

Sun, B.B., Maranville, J.C., Peters, J.E., Stacey, D., Staley, J.R., Blackshaw, J., Burgess, S., Jiang, T., Paige, E., Surendran, P., et al. (2018). Genomic atlas of the human plasma proteome. Nature *558*, 73–79.

Tewhey, R., Kotliar, D., Park, D.S., Liu, B., Winnicki, S., Reilly, S.K., Andersen, K.G., Mikkelsen, T.S., Lander, E.S., Schaffner, S.F., and Sabeti, P.C. (2016). Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. Cell *165*, 1519–1529.

Trynka, G., Sandor, C., Han, B., Xu, H., Stranger, B.E., Liu, X.S., and Raychaudhuri, S. (2013). Chromatin marks identify critical cell types for fine mapping complex trait variants. Nat. Genet. *45*, 124–130.

van Arensbergen, J., Pagie, L., FitzPatrick, V., de Haas, M., Baltissen, M., Comoglio, F., van der Weide, R., Teunissen, H., Võsa, U., Franke, L., et al. (2018). Systematic identification of human SNPs affecting regulatory element activity. boiRxiv. https://doi.org/10.1101/460402.

Vaz-Drago, R., Custódio, N., and Carmo-Fonseca, M. (2017). Deep intronic mutations and human disease. Hum. Genet. *136*, 1093–1111.

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.

Võsa, U., Claringbould, A., Westra, H., Jan Bonder, M.J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Kasela, S., et al. (2018). Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis. bioRxiv. https://doi.org/10.1101/447367.

Waszak, S.M., Delaneau, O., Gschwind, A.R., Kilpinen, H., Raghav, S.K., Witwicki, R.M., Orioli, A., Wiederkehr, M., Panousis, N.I., Yurovsky, A., et al. (2015). Population Variation and Genetic Control of Modular Chromatin Architecture in Humans. Cell *162*, 1039–1050.

Weiner, D.J., Wigdor, E.M., Ripke, S., Walters, R.K., Kosmicki, J.A., Grove, J., Samocha, K.E., Goldstein, J.I., Okbay, A., Bybjerg-Grauholm, J., et al.; iPSYCH-Broad Autism Group; Psychiatric Genomics Consortium Autism Group (2017). Polygenic transmission disequilibrium confirms that common and rare variation act additively to create risk for autism spectrum disorders. Nat. Genet. *49*, 978–985.

Weisenfeld, N.I., Kumar, V., Shah, P., Church, D.M., and Jaffe, D.B. (2017). Direct determination of diploid genome sequences. Genome Res. *27*, 757–767.

Wen, X., Pique-Regi, R., and Luca, F. (2017). Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. PLoS Genet. *13*, e1006646.

Willems, T., Gymrek, M., Highnam, G., Mittelman, D., and Erlich, Y.; 1000 Genomes Project Consortium (2014). The landscape of human STR variation. Genome Res. *24*, 1894–1904.

Xiong, H.Y., Alipanahi, B., Lee, L.J., Bretschneider, H., Merico, D., Yuen, R.K., Hua, Y., Gueroussov, S., Najafabadi, H.S., Hughes, T.R., et al. (2015). RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. Science *347*, 1254806.

Yeo, G., and Burge, C.B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J. Comput. Biol. *11*, 377–394.

Zhernakova, D.V., Deelen, P., Vermaat, M., van Iterson, M., van Galen, M., Arindrarto, W., van 't Hof, P., Mei, H., van Dijk, F., Westra, H.J., et al. (2017). Identification of context-dependent expression quantitative trait loci in whole blood. Nat. Genet. *49*, 139–145.

Zhou, J., Theesfeld, C.L., Yao, K., Chen, K.M., Wong, A.K., and Troyanskaya, O.G. (2018). Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. Nat. Genet. *50*, 1171–1179.

Zook, J.M., Chapman, B., Wang, J., Mittelman, D., Hofmann, O., Hide, W., and Salit, M. (2014). Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nat. Biotechnol. *32*, 246–251.