

REVIEW

From variant to function in human disease genetics

Tuuli Lappalainen^{1,2*} and Daniel G. MacArthur^{3,4,5*}

Over the next decade, the primary challenge in human genetics will be to understand the biological mechanisms by which genetic variants influence phenotypes, including disease risk. Although the scale of this challenge is daunting, better methods for functional variant interpretation will have transformative consequences for disease diagnosis, risk prediction, and the development of new therapies. An array of new methods for characterizing variant impact at scale, using patient tissue samples as well as *in vitro* models, are already being applied to dissect variant mechanisms across a range of human cell types and environments. These approaches are also increasingly being deployed in clinical settings. We discuss the rationale, approaches, applications, and future outlook for characterizing the molecular and cellular effects of genetic variants.

Since the completion and publication of the human genome approximately two decades ago (1), the primary bottleneck in human genetics has shifted from the discovery of genetic variation to the large-scale characterization of the associations between genetic variation and phenotype, and more recently to the characterization of the causal mechanisms by which genetic variation influences human biology. Of the more than 200,000 genetic variants confidently linked with human complex traits through genome-wide association studies (GWAS), the majority remain mechanistically uncharacterized (2). Furthermore, because the majority of these variants fall in noncoding regions of the genome, there is often uncertainty about which gene is responsible for their biological effects (3). Similarly, of the nearly 1 million entries in the ClinVar database (4) of variants identified in patients with severe genetic disease, 47% are classified as having either uncertain or conflicting annotations, indicating a lack of clarity about variants' impact on molecular function and disease (5).

Resolving this pervasive uncertainty will require more accurate and scalable methods to unravel the molecular processes by which genetic variation influences phenotype. Such approaches will increase the accuracy of genetic diagnosis and prediction, especially for rare or unique disease-causing variants, by enhancing the direct inference of variants likely to disrupt the normal function of critical genes. These approaches will also accelerate therapeutic development, not only by highlighting the gene products directly involved in the causation of disease but also by revealing the direction of effect, the relevant cell types, and the overall biological pathways by which a gene influences disease risk.

¹Science for Life Laboratory, Department of Gene Technology, KTH Royal Institute of Technology, Stockholm, Sweden.

²New York Genome Center, New York, NY, USA. ³Centre for Population Genomics, Garvan Institute of Medical Research, and UNSW Sydney, Sydney, New South Wales, Australia.

⁴Centre for Population Genomics, Murdoch Children's Research Institute, Melbourne, Victoria, Australia. ⁵Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA.

*Corresponding author. Email: tuuli.lappalainen@scilifelab.se (T.L.); daniel.macarthur@populationgenomics.org.au (D.G.M.)

The genetic architecture of human disease

Over the past decade, it has become clear that the genetic architecture of human traits—the characteristics of the variants that causally influence those diseases—is highly variable. This architecture is the consequence of the interplay between the demographic and selective forces from our species' evolutionary history (6) and the differing distribution of the locations of variants within our genome. The action of natural selection means, in general, that variants with large effects on biological function and disease risk will tend to be removed from the population; any such variants will thus tend to be very rare, and if a variant is common, it is thus extremely unlikely to have large effects on disease (7). The structure of the genome means that variants with the largest effect will tend to be found in or very close to protein-coding regions, as changes in protein sequence and structure can affect a protein's normal function across all or most cell types where that gene is expressed.

In contrast, variants falling outside protein-coding regions (noncoding variants) are much more likely to be biologically neutral, and those that do affect biology likely do so through more subtle and potentially cell type-specific changes in gene regulation (8).

The interplay between evolutionary and biological forces explains the architecture of human traits that has emerged from large-scale genomic analysis. Causal variants across a wide range of disease fall along a spectrum of variant frequency and effect size (9) (Fig. 1A). One end of the spectrum represents very rare variants with large effects on human biology, virtually all of which are protein-coding or near-coding, including the majority of variants identified as causal for rare, severe genetic diseases. On the other end of the spectrum are variants that have small individual effects on disease risk, but which [because they experienced weaker selection (10)] have been able to reach higher frequencies in the population and generally have an impact on regulatory elements in noncoding sequences (11). Because of their higher frequency, these common noncoding variants contribute to a large fraction of the total explained heritability (genetic contribution to risk) of common diseases such as type 2 diabetes (12). These trends are not absolute, however: Rare variation, especially in protein-coding regions, nonetheless contributes disproportionately to the genetic architecture of common complex traits (12, 13), and common variants can also contribute to the risk and severity of rare diseases (14, 15).

This genetic architecture has consequences for the optimal approaches for discovering causal variants (Fig. 1A). Rare coding variants

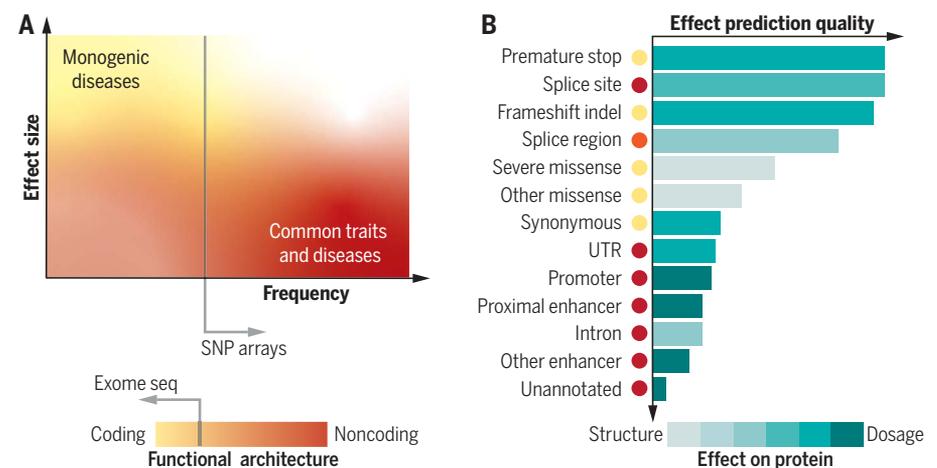


Fig. 1. Genetic variant effects on disease and gene function. (A) Functional genetic architecture of human disease ranging from a spectrum of monogenic diseases driven by rare predominantly coding variants with a strong effect on disease to common diseases and traits whose heritability is dominated by noncoding common variants with small effect sizes. However, these classifications are not exclusive: Noncoding causal variants in Mendelian disease exist, as do rare variants affecting common traits and diseases. **(B)** Quality of variant effect predictions by functional annotation class. The color of the bars indicates whether the variant classes tend to affect protein structure or dosage or whether a given class of variants can affect both—for example, via nonsense-mediated decay and splicing. The color of the labels indicates coding and noncoding annotation.

of large effect are currently efficiently identified with exome sequencing studies (an approach that cost-effectively sequences the protein-coding segments of the genome), either in families (for variants with extremely large effect sizes, such as those underlying most “monogenic” disorders) or in large case/control studies (16, 17). In contrast, common variants with smaller effect sizes are cost-effectively discovered with large-scale case/control studies with genotyping arrays or low-coverage whole-genome sequencing. Note that although the differential architecture of common and rare disorders is clear, its degree may have been exaggerated by the different methods predominantly used in each field that have hindered the discovery of noncoding variants in rare disease and rare variants in complex disease (12). Over the next decade, deep whole-genome sequencing will be applied to much larger cohorts of patients and population samples, providing an unbiased view of genetic architecture and improving the discovery of its shared features across different types of diseases (13, 18).

Genetic architecture also alters the optimal approaches for using functional data to understand variants’ effects on phenotype. In common disease, associated GWAS loci typically contain a large number of variants in linkage disequilibrium—meaning that they are close enough to each other that they are typically inherited together—of which only a small proportion are causal and affect the regulatory function of the genome (19). For such cases, a functional dissection typically requires assaying noncoding regulatory elements, often in a highly cell type-specific or environment-specific fashion. For rare disease variants, and for the small proportion of common disease variants found in coding regions, assays for functional impact generally focus on changes in the structure or dosage of mRNA transcripts and their encoded proteins. Below, we discuss high-throughput approaches spanning the complete range of potential biological impacts that will be needed for a comprehensive understanding of human variation.

Observational approaches for characterizing functional variation

The foundation of genome-wide variant interpretation is the functional annotation of the genome, which has been established by projects including ENCODE (20) GenCode (21), and Roadmap Epigenomics (22), which have measured the biochemical activity of the genome in multiple cell types. The resulting annotation of genes, transcripts, and regulatory elements has enabled high-quality predictions of the most severe classes of likely gene-disrupting variants (Fig. 1B), which are particularly important for rare disease interpretation and therapeutic target discovery (23).

However, for most variant classes, accurate predictions of variant effects cannot be de-

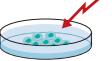
Approach	Biological systems	Genetic variation	Phenotyping	Interpretation
Genetic analysis of human populations 	 Primary cells in native physiological environments	Exact representation of natural variation	Access to physiological traits	Direct relevance to human physiology and health
	 Complexity of tissues, limited access to dynamic cellular states	Limited to study of existing genetic variation in the population	Confounders of molecular and cellular phenotypes limit interpretability	Interindividual variation and confounders complicate interpretation of causality
Cellular models in vitro 	 Access to and control of cell differentiation and state	Perturbation of any loci with different variants of diverse effect sizes	Cleaner data from molecular and cellular phenotyping	Cleaner data of causal molecular and cellular mechanisms
	 Limited set of cell lines or other models in cell culture media	SNP editing difficult; assays may lack genomic context	Physiological phenotypes not available	Potentially limited generalizability to physiological systems

Fig. 2. Functional characterization approaches relying on analysis of genetic variation in human populations versus experimental perturbations of the genome or its function.

rived solely from genome annotations. Even missense variants, where the amino acid changes are easy to identify and annotate, typically lack high-quality predictions of their effects on protein structure and function (24). For noncoding variants dominating complex disease heritability, predictions of regulatory effects currently have limited practical applicability and performance. This limitation is due to our generally poor knowledge of the complex regulatory code of the genome and the fact that such analyses are dependent on the specific cellular context. Variants affecting transcript splicing—an important class of post-transcriptional effects in both rare and common disease (25, 26)—fall between these extremes of prediction performance. Specifically, variants in canonical splice sites are easy to annotate, and predictive models have reasonable performance (27). However, changes in splicing (e.g., due to synonymous variants or variants deep in introns) remain difficult to predict (27).

Common genetic effects on gene regulation can be inferred by mapping quantitative trait loci (QTLs) for proximal molecular features. The most common type of molecular QTLs are cis-eQTLs (cis-associated QTLs associated to gene expression levels), but QTLs can also be mapped for splicing, chromatin accessibility, and protein levels, for example (28). Human eQTL studies have been pursued for more than 10 years, including by consortia such as GTEx (25) and eQTLGen (29), and have generated comprehensive catalogs of cis-eQTLs and splicing QTLs (sQTLs) (30) in diverse human tissues. Together with chromatin accessibility QTL (caQTL) and protein QTL (pQTL) maps, these data have been valuable in pinpointing potential causal genes and molecular mechanisms in individual loci.

However, the QTL approach has limitations. It can only capture the effects of existing common variation where linkage disequilibrium obscures the identity of the actual causal

variants (31), and this makes it suboptimal for scanning the functional effects of individual nucleotides. Furthermore, when the same variant or haplotype affects the expression of multiple genes or molecular traits, it is difficult to assess which one(s) have downstream effects that causally contribute to the disease association. The success of large studies in identifying eQTLs or sQTLs for nearly all human genes has made this problem of low specificity painfully clear (32). It is further exacerbated by the poor representation of non-European ancestries in QTL datasets (30), which also limits the utility of current QTL catalogs in the interpretation of GWAS data from diverse ancestries. In cases of rare variant analysis, the association-based QTL approach is not applicable, but analogous characterization of rare variants with likely molecular effects can identify situations where a rare variant coincides with an individual’s status as a population outlier for gene expression, splicing, or allelic expression (33).

Currently, QTL datasets are limited by their derivation primarily from a limited number of easily accessible tissues (such as blood) or post mortem tissue samples (30). These biospecimens are a mixture of multiple, often unknown, steady-state cell types. This can lead to poor resolution and detection power for genetic regulatory effects that are active in specific cell types or dynamic cell states that may be particularly important in disease risk (34). Thus, expanding QTL studies to computationally resolved cell types (35) and single-cell datasets from tissue samples is a major ongoing focus of the field (36), especially because many cell types are not represented by available cell lines. A complementary approach to analyzing developmental lineages and environmental responses is to extract primary cells or cell lines from donors, induce pluripotency, and grow or differentiate the cells in vitro to study molecular phenotypes by bulk or single-cell sequencing. This has allowed access to key cell types that

are next to impossible to sample at scale, and also provides a renewable and scalable source for molecular assays (28, 37–39). However, not all primary cell types can be easily extracted or differentiated, and thus the map of regulatory associations of standing human variation across all cell types is still very incomplete.

Genome perturbation approaches

In addition to analysis of naturally existing genetic variation in human populations, we now can perform scalable experimental introduction of genetic variants in a cellular model system, followed by molecular phenotyping of their effects, often at a single-cell level (40). The assays that pair potential regulatory sequences with sequencing barcodes in artificial DNA sequences, so that the ability of the different alleles to drive gene expression can be compared (41, 42), are the best suited to test thousands of variants. However, this approach only detects the cis-regulatory potential of short sequences. Pooled CRISPR-based perturbation of the genome or its regulatory activity, followed by single-cell RNA sequencing and cellular phenotyping (43–46), has enabled the characterization of target genes in cis and downstream cellular effects of regulatory or coding variants.

These approaches are rapidly expanding in scale and scope, including deeper phenotyping of diverse cellular functions by high-throughput imaging, multi-omics, as well as application not only to cell lines but also to organoid and animal models. However, there are substantial limitations to the scale and cost of precision genome perturbation, cellular phenotyping, and the availability of good biological model systems. The latter represents a major challenge: Easy-to-use cancer cell lines represent a limited range of human cell types and differ from primary cells in many ways. The extent to which different model systems can recapitulate variant effects is largely unknown, poorly defined, and likely to vary not only between the cells used but also between different types of variants (25). For example, a variant introducing a premature stop in a constitutively expressed exon of a ubiquitously expressed gene is likely to have much lower context specificity than a variant in a cell state-specific enhancer. Thus, integration of experimental approaches with human genetic studies remains valuable as a way to leverage the benefits of both approaches (Fig. 2).

The clinical impact of variant-to-function approaches

The molecular diagnosis of patients affected by severe genetic diseases has accelerated rapidly over the past decade as a result of the increased availability of exome and genome sequencing, improved methods for variant interpretation, and global data sharing. Current standards for

clinical variant annotation (47) require the integration of multiple classes of evidence to assess the probability that a variant is pathogenic (disease-causing). These fall broadly into three categories: (i) evidence that the variant is rare in the general population, (ii) evidence that the variant has previously been seen in other patients with similar clinical appearance, and (iii) evidence (either direct or indirect) that the variant has a functional impact on a gene previously implicated in the same class of disorders. Assessment of the first two categories has been empowered by the availability of large databases of population variation (48) and of variants previously seen and interpreted in disease patients (4, 49). However, assessing the probability that a variant is biologically damaging is much more challenging and less standardized.

Because most variants found in patients have not yet been subjected to well-calibrated functional assays, most clinical interpretation relies on indirect measures of functional impact, such as evolutionary conservation (47). Historically, when functional information was incorporated, it was generated with custom assays for specific variants, often were only weakly quantitative, and lacking information regarding the probability of obtaining the same result from randomly selected variants in the same gene (50). However, large-scale genomic approaches to inferring variant function are now feasible, and are beginning to be applied in clinical settings.

Such approaches fall into two broad classes. First, there are direct assays run on patient tissue samples or cell lines, which provide a direct functional readout of variants discovered in that patient. The most widely used functional genomic assay for the diagnosis of rare genetic diseases is bulk transcriptome sequencing. This approach can be used to assess the impact of variants on gene expression and transcript splicing; its use improves diagnosis rates over those derived from DNA sequencing alone, although these improvements vary by disease and tissue type [$>30\%$ when applied to muscle biopsies from severe muscle disease patients (51); 16% in a study of fibroblast samples from patients with mitochondrial disease (52)]. Proteomic analysis has also proven valuable both for the assessment of individual candidate genes (53) and in combination with transcriptomics for diagnosis (54). Other genome-scale technologies including metabolomics and epigenomics have also been deployed to improve the identification of causal genes (55).

These approaches are powerful, but access to disease-relevant tissues and cell types in a clinical setting remains challenging; post mortem tissue collection is of limited use, and differentiation of induced pluripotent stem cells from patient tissue comes with a considerable cost. Other hurdles to widespread clinical adoption include the heterogeneity of clinical sam-

ples, the lack of standardization of assays and analysis pipelines, and the need for experienced data interpretation.

A second broad category of approaches consists of *in vitro* assays in which an artificially created model of the variant is tested for functional impact in a well-established cell line, as outlined above. These approaches can be rapidly scaled into multiplexed assays of variant effect (MAVEs), allowing for the rapid testing of candidate variants, and potentially of all possible genetic variants for a given gene (40). Such assays have already proved effective in inferring the pathogenicity of variants for a subset of well-established disease genes, in some cases eclipsing standard approaches to variant interpretation (56). For MAVEs to accurately infer pathogenicity, the precise functional assay used must be both highly scalable and tightly correlated with the disease-relevant function(s) of the target gene and the mechanisms of action of pathogenic variants, and thus will be extremely challenging or impossible for a subset of disease-causing variation. Although MAVE data do not yet exist for the vast majority of clinically relevant genes, current approaches include assays exquisitely customized to individual genes, as well as generalizable assays of protein stability (40, 57). Efforts are now under way to systematically develop well-calibrated assays for clinically relevant genes, and to harmonize and release the resulting data (58).

The future of variant function

The human population, through explosive growth, has performed a comprehensive saturation mutagenesis experiment on itself. It is now the case that any single base substitution that is compatible with life is expected to be present somewhere among the nearly 8 billion living humans (59). Humanity has thus, in effect, done many of the natural experiments required to understand our own genotype-phenotype map; this leaves geneticists to catalog the outcomes of those experiments, and to leverage both observational and experimental approaches to understand the mechanisms by which variants alter biology. Over the next decade, unless hampered by major obstacles to data sharing, increasingly massive cohorts of disease patients and deeply phenotyped population samples should produce the requisite catalog, and sophisticated and scalable tools will be applied to characterize the underlying functional mechanisms in both research and clinical settings (Table 1). These approaches should generate large, standardized datasets that can be combined with machine learning tools to provide predictions of variant effects on biology and disease.

We believe that the consequences of this process will be profound. In the clinical setting, the assessment of the potential effects of

Table 1. Variant-to-function challenges in common and rare disease, and approaches to address them.

	● Common disease	● Rare disease
	Challenge	Approaches
Causal variant	Linkage disequilibrium	Fine-mapping, improvements from diverse populations and functional priors; scalable experimental assays
	Few observations of any given variant ($N \sim 1$)	Gene and variant prioritization with larger datasets and functional priors; scalable experimental assays
Immediate functional effect	“Regulatory code” of variant effects on regulatory elements mostly unknown	Regulatory region annotations, chromatin/splicing QTL mapping, scalable experimental assays, predictive models
	Difficulty of predicting effects on protein 3D structure and function	Predictive models, scalable experimental assays
Causal gene and effect direction	Difficulty of enhancer target prediction	eQTL mapping, predictive models, scalable experimental assays
	Small number of patients with genetic disruption of a given gene	Data aggregation; better priors for gene and variant function
Downstream cellular effects	Small effect sizes of individual loci; relevant cell type/state often unknown	Cellular phenotype characterization in (selected) population samples or model systems
	Few observations of any given variant ($N \sim 1$)	

a sequence variant on disease risk will become increasingly quantitative, driven by access to three critical strands of information: the frequency of that variant across hundreds of millions of humans; the phenotype observed in any other human observed to carry it; and the results of well-calibrated experiments assessing the precise impact of that variant on gene function. In effect, variant interpreters should have access to a quantitative functional look-up table for any possible variant in a wide variety of clinically relevant genes. This will empower the diagnosis of genetic disease while also improving the prediction of risk for clinical phenotypes that have not yet manifested (e.g., in the assessment of genomes sequenced before or at birth, or those with environmental responses).

For complex traits, rapid *in silico* assessment of likely functional effects should speed the identification of causal variant(s) responsible for multiple uncharacterized risk loci and their likely target genes in *cis*, and will increase the proportion of accurate predictions of the affected pathways and relevant cell types. It will also enhance individual risk prediction, as genome-wide functional annotation can already improve the accuracy and portability of polygenic scores for complex disease (60, 61), and may allow the partitioning of individual risk into components driven by different physiological mechanisms and potentially amenable to different therapeutic approaches (62).

The functional characterization of human variants is likely to continue to increasingly influence the development of new therapeutics or repurposing of already established drugs. Drug targets with genetic evidence have high-

er success rates (63, 64). For a trait-associated gene affected by multiple functional variants, robust functional data of their effects can provide allelic series, allowing for functional dosages for each gene to be linked to phenotypic outcomes (65) and directly support specific therapeutic hypotheses. The convergence of genome-wide disease signals into genetically implicated pathways (66) should provide targets beyond individual GWAS genes. Characterization of the cell types where genetic risk effects manifest—which are often different from the organs affected by the disease—is also important for being able to target causal pathways rather than treating consequences. Understanding the molecular mechanisms underlying pathology will also pinpoint pathways that mediate the contribution of environmental and developmental risk factors to disease, and thus advance the integration of genetic and nongenetic factors in predicting and treating disease.

Finally, a shift from pure genetic approaches toward high-throughput functional data as evidence for variant interpretation will also contribute to greater equity in the utility of genomic medicine across communities. Allele frequencies and linkage disequilibrium patterns reflect population history and are thus highly population-specific; hence, they can contribute to biases due to the current underrepresentation of non-European samples (67). In contrast, although individual variants affecting disease risk often vary by ancestry, their molecular and cellular mechanisms are more likely to be shared among all humans. This hypothesis should be systematically addressed by performing functional experiments in mod-

el systems that represent different ancestries, sexes, ages, and environmental exposures.

Despite these potentialities, the challenges ahead for high-throughput biologists are daunting. Although human genetic datasets are increasingly rich, the power to assess trait associations for all genetic variants—let alone their combinations—has its limits. The complexity of highly correlated and easily confounded human phenotype measurements greatly complicates this task. The paths by which genetic variants lead to phenotypic outcomes are complex and dynamic: They take place across multiple cell types, in response to many environmental conditions and developmental cues. They act through a wide range of molecular and cellular processes, and potentially interact with other genetic variants present in the same individual, including somatic variants that have arisen during cell division over that person’s lifetime. As such, developing an understanding of the physiological mechanisms of disease will require deep molecular characterization across interacting cell types and dynamic cell states, as well as access to informative biospecimens and/or biological model systems. Across these, we must harness the power of a diverse and complementary toolkit, ranging from interrogation of natural genetic variation to genome perturbations, followed by multimodal cellular assays. Although we are optimists about the transformative power of these approaches, it is always important to be humble in the face of biology’s complexity, and we acknowledge that interactions and higher-order effects can complicate conclusions drawn from simple experimental models.

Different approaches, such as those outlined above, have unique advantages and limitations. Embracing a diversity of scalable approaches, including highly generalizable assays (of gene expression, protein structure, and protein stability), well-calibrated assays of gene-specific functions (such as enzyme activity), and analytical methods for integrating complex information across molecular classes and cell types, will be necessary to deeply characterize the mechanisms by which genome sequence shapes human biology.

REFERENCES AND NOTES

1. International Human Genome Sequencing Consortium, *Nature* **409**, 860–921 (2001).
2. M. Claussnitzer *et al.*, *Nature* **577**, 179–189 (2020).
3. M. T. Maurano *et al.*, *Science* **337**, 1190–1195 (2012).
4. M. J. Landrum *et al.*, *Nucleic Acids Res.* **46**, D1062–D1067 (2018).
5. ClinVar; www.ncbi.nlm.nih.gov/clinvar.
6. G. Sella, N. H. Barton, *Annu. Rev. Genomics Hum. Genet.* **20**, 461–493 (2019).
7. J. Zeng *et al.*, *Nat. Commun.* **12**, 1164 (2021).
8. E. A. Boyle, Y. I. Li, J. K. Pritchard, *Cell* **169**, 1177–1186 (2017).
9. M. I. McCarthy *et al.*, *Nat. Rev. Genet.* **9**, 356–369 (2008).
10. E. C. Glassberg, Z. Gao, A. Harpak, X. Lan, J. K. Pritchard, *Genetics* **211**, 757–772 (2019).
11. H. K. Finucane *et al.*, *Nat. Genet.* **47**, 1228–1235 (2015).
12. S. Gazal *et al.*, *Nat. Genet.* **50**, 1600–1607 (2018).
13. P. Wainschein *et al.*, bioRxiv 588020 [preprint]. 11 June 2021.

14. A. R. Harper *et al.*, *Nat. Genet.* **53**, 135–142 (2021).
15. S. S. Shringarpure *et al.*, medRxiv 21258643 [preprint]. 16 June 2021.
16. O. Zuk *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **111**, E455–E464 (2014).
17. J. E. Posey *et al.*, *Genet. Med.* **21**, 798–812 (2019).
18. E. Turro *et al.*, *Nature* **583**, 96–102 (2020).
19. O. Weissbrod *et al.*, *Nat. Genet.* **52**, 1355–1363 (2020).
20. ENCODE Project Consortium, *Nature* **489**, 57–74 (2012).
21. A. Frankish *et al.*, *Nucleic Acids Res.* **47**, D766–D773 (2019).
22. Roadmap Epigenomics Consortium, *Nature* **518**, 317–330 (2015).
23. E. V. Minikel *et al.*, *Nature* **581**, 459–464 (2020).
24. B. J. Livesey, J. A. Marsh, *Mol. Syst. Biol.* **16**, e9380 (2020).
25. GTEx Consortium, *Science* **369**, 1318–1330 (2020).
26. R. Soemedi *et al.*, *Nat. Genet.* **49**, 848–855 (2017).
27. K. Jaganathan *et al.*, *Cell* **176**, 535–548.e24 (2019).
28. Y. I. Li *et al.*, *Science* **352**, 600–604 (2016).
29. U. Vösa *et al.*, bioRxiv 447367 [preprint]. 19 October 2018.
30. N. Kerimov *et al.*, bioRxiv 924266 [preprint]. 9 January 2021.
31. Q. S. Wang *et al.*, *Nat. Commun.* **12**, 3394 (2021).
32. A. N. Barbeira *et al.*, *Genome Biol.* **22**, 49 (2021).
33. X. Li *et al.*, *Nature* **550**, 239–243 (2017).
34. B. D. Umans, A. Battle, Y. Gilad, *Trends Genet.* **37**, 109–124 (2021).
35. S. Kim-Hellmuth *et al.*, *Science* **369**, eaaz8528 (2020).
36. M. van der Wijst *et al.*, *eLife* **9**, e52155 (2020).
37. D. Liang *et al.*, *Nat. Neurosci.* **24**, 941–953 (2021).
38. M. C. Ward, N. E. Banovich, A. Sarkar, M. Stephens, Y. Gilad, *eLife* **10**, e57345 (2021).
39. B. P. Fairfax *et al.*, *Science* **343**, 1246949 (2014).
40. G. M. Findlay, *Hum. Mol. Genet.* ddab219 (2021).
41. R. Tewhey *et al.*, *Cell* **165**, 1519–1529 (2016).
42. M. Kircher *et al.*, *Nat. Commun.* **10**, 3583 (2019).
43. A. Dixit *et al.*, *Cell* **167**, 1853–1866.e17 (2016).
44. J. A. Morris *et al.*, bioRxiv 438882 [preprint]. 8 April 2021.
45. M. C. Canver *et al.*, *Nature* **527**, 192–197 (2015).
46. B. Adamson *et al.*, *Cell* **167**, 1867–1882.e21 (2016).
47. ACMG Laboratory Quality Assurance Committee, *Genet. Med.* **17**, 405–424 (2015).
48. K. J. Karczewski *et al.*, *Nature* **581**, 434–443 (2020).
49. H. L. Rehm *et al.*, *N. Engl. J. Med.* **372**, 2235–2242 (2015).
50. D. G. MacArthur *et al.*, *Nature* **508**, 469–476 (2014).
51. B. B. Cummings *et al.*, *Sci. Transl. Med.* **9**, eaal5209 (2017).
52. V. A. Yépez *et al.*, medRxiv 21254633 [preprint]. 5 April 2021.
53. N. J. Lake *et al.*, *Am. J. Hum. Genet.* **101**, 239–254 (2017).
54. R. Kopajtic *et al.*, medRxiv 21253187 [preprint]. 12 March 2021.
55. K. Kerr *et al.*, *Orphanet J. Rare Dis.* **15**, 107 (2020).
56. G. M. Findlay *et al.*, *Nature* **562**, 217–222 (2018).
57. K. A. Matreyek *et al.*, *Nat. Genet.* **50**, 874–882 (2018).
58. The Atlas of Variant Effects (AVE) Alliance: Understanding Genetic Variation at Nucleotide Resolution (2021); <https://zenodo.org/record/4989960>.
59. B. H. Shirts, C. C. Pritchard, T. Walsh, *Trends Mol. Med.* **22**, 925–934 (2016).
60. T. Amariuta *et al.*, *Nat. Genet.* **52**, 1346–1354 (2020).
61. P. M. Visscher, L. Yengo, N. J. Cox, N. R. Wray, *Science* **373**, 1468–1473 (2021).
62. M. S. Udler, M. I. McCarthy, J. C. Florez, A. Mahajan, *Endocr. Rev.* **40**, 1500–1520 (2019).
63. M. R. Nelson *et al.*, *Nat. Genet.* **47**, 856–860 (2015).
64. E. A. King, J. W. Davis, J. F. Degner, *PLOS Genet.* **15**, e1008489 (2019).
65. R. M. Plenge, E. M. Scolnick, D. Altshuler, *Nat. Rev. Drug Discov.* **12**, 581–594 (2013).
66. N. Sinnott-Armstrong, S. Naqvi, M. Rivas, J. K. Pritchard, *eLife* **10**, e58615 (2021).
67. A. Buniello *et al.*, *Nucleic Acids Res.* **47**, D1005–D1012 (2019).

ACKNOWLEDGMENTS

We thank A. Pinho for her contributions to the preparation of this manuscript. **Funding:** Supported by the KTH Royal Institute of Technology, New York Genome Center, and NIH grants R01GM122924, R01HL142028, R01AG057422, and R01MH106842 (T.L.) and the Garvan Institute of Medical Research, the Murdoch Children's Research Institute, and NIH grants U24HG011450 and U01HG011755 (D.G.M.). **Author contributions:** T.L. and D.G.M. contributed equally to the writing of this manuscript; author order was determined by a coin toss. **Competing interests:** T.L. is an advisor for Variant Bio, Goldfinch Bio, and GSK and has stock in Variant Bio. D.G.M. is a founder with equity in Goldfinch Bio and is an advisor to Insitro, Variant Bio, GSK, and Foresite Labs.

10.1126/science.abi8207

From variant to function in human disease genetics

Tuuli Lappalainen Daniel G. MacArthur

Science, 373 (6562), • DOI: 10.1126/science.abi8207

View the article online

<https://www.science.org/doi/10.1126/science.abi8207>

Permissions

<https://www.science.org/help/reprints-and-permissions>

Use of think article is subject to the [Terms of service](#)

Science (ISSN) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science* is a registered trademark of AAAS.

Copyright © 2021 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works