HHS Public Access

Author manuscript

Hum Hered. Author manuscript; available in PMC 2017 July 12.

Published in final edited form as:

Hum Hered. 2016 ; 81(2): 78–87. doi:10.1159/000447453.

Non-coding loss-of-function variation in human genomes

Zachary Zappala[1] and **Stephen B. Montgomery**[1,2,*]

[1]Department of Genetics, Stanford University, California, USA

[2]Department of Pathology, Stanford University, California, USA

Abstract

Whole genome and exome sequencing in human populations has revealed the tolerance of each gene for loss-of-function variation. By understanding this tolerance, it has become increasingly possible to identify genes that would make safe therapeutic targets and to identify rare genetic risk factors and phenotypes at the scale of individual genomes. To date, the vast majority of surveyed loss-of-function variants are in protein-coding regions of the genome mainly due to the focus on these regions by exome-based sequencing projects and their relative ease of interpretability. As whole genome sequencing becomes more prevalent, new strategies will be required to uncover impactful variation in non-coding regions of the genome where the architecture of genome function is more complex. In this review, we investigate recent studies of loss-of-function variation and emerging approaches for interpreting whole genome sequencing data to identify rare and impactful non-coding loss-of-function variants.

Keywords

non-coding variation; loss-of-function; rare variation; Mendelian disease genetics

Introduction

Humans have recently undergone explosive growth [1], leading to an excess of rare genetic variation across diverse global populations [2] with largely unknown impacts on human traits. Accurate characterization of these variants and their functional significance can reveal novel drivers of rare diseases [3–5], modifiers of common disease risk [6,7], and potentially protective variants and therapeutic targets [8,9]. Major efforts to identify impactful rare variants have focused on protein-coding regions of the genome due to the relative cost of exome sequencing compared to whole genome sequencing and the ease of identifying impactful variants in protein-coding regions from primary sequence data alone (Table 1). However, protein-coding variants comprise only a subset of potentially impactful variation in an individual's genome – to address this shortfall in genome interpretation, new approaches are needed that can identify impactful rare variation in the non-coding genome.

*Corresponding author: Stephen B. Montgomery, PhD, 300 Pasteur Drive, Stanford, CA, USA, 94305-5324, Telephone: 650-561-5628, smontgom@stanford.edu.

In this review, we summarize recent loss-of-function studies focused on coding variation – these studies have revealed both a gene's tolerance and intolerance for loss-of-function mutations, thereby aiding in identifying the genetic basis of multiple Mendelian disorders and guiding our understanding of the challenge in identifying and interpreting impactful variants in the non-coding genome [3–5]. We then address efforts to characterize rare variation in the non-coding genome by focusing on existing knowledge and emerging experimental and computational technologies that serve as a transition point to whole genome interpretation. We conclude by discussing how such improvements are deepening our understanding of the role of rare non-coding variation in disease risk and how more complete assessment of rare variation in the context of an individual's genomic background will be fundamental to personalized genomics and precision health.

Loss-of-function variation in human populations

Recent gene sequencing efforts in large population cohorts have identified an abundance of rare variants. Nelson *et al.* sequenced 202 genes in 14,002 individuals and found that more than 95% of the discovered variants had allele frequencies less than 0.5% [9]. Similarly, Tennessen *et al.* sequenced the exomes of 2,440 individuals and found that ~96% of predicted functionally important variants are rare [10]. With estimates of the *de novo* mutation rate between 1.1e-8 and 1.4e-8 and the current human population size of 7 billion people, it is possible to predict that every possible mutation is tested approximately 80 times a generation. Population-scale sequencing efforts such as the 1000 Genomes Project have shown that many protein-coding mutations, including nonsense, splice site-disrupting, or frameshift mutations, have a high probability of causing loss-of-function phenotypes [11]. Indeed, it was remarkable to find that there were often multiple protein-coding loss-of-function variants that would cause partial or full inactivation of their respective genes in every healthy individual sequenced.

Deciphering the prevalence of loss-of-function variants and their phenotypic consequences has required studies in well-phenotyped population cohorts with thousands to tens of thousands of individuals. In Finland, 83 loss-of-function variants enriched in Finnish individuals were identified through exome sequencing of 3,000 individuals and then tested for their association with 60 phenotypes in 36,262 Finns [12]. Five of these variants showed genome-wide significant associations, with the most notable being a splice variant in the LPA gene that conferred protection against cardiovascular disease. Two recent studies in Pakistani individuals have also highlighted tens of thousands of rare, protein-coding loss-of-function alleles. Narasimhan *et al.* exome sequenced 3,222 individuals and found 16,708 loss-of-function variants, 847 of which were homozygous in 781 genes [13]. On average, they observed 1.6 heterozygous recessive lethal loss-of-function variants per adult. Additionally, they identified 38 individuals who were homozygous for rare loss-of-function variants in known recessive Mendelian disease genes. However, they observed no relationship between health records for these individuals and the expected disease phenotypes for 32 of the 38 cases (84%) – they attributed this discrepancy to a variety of factors including incomplete penetrance, delayed onset of disease, and to the generally positive health status of the study participants. Saleheen *et al.* exome sequenced 7,078 individuals and found 36,850 loss-of-function variants that were homozygous in 961 genes

[14]. They found a total of 264 genes where two or more participants were homozygous for loss-of-function variants. When they tested these variants for association with 201 phenotypes, however, they only detected 14 significant associations after correcting for multiple testing. These studies demonstrate the difficulty in identifying the phenotypic consequences of genetic variation – despite these individuals carrying a large numbers of heterozygous and homozygous loss-of-function variants, particularly in well-known disease genes, only a handful of these variants can be associated with phenotypes and many do not confer any obvious clinical risks.

Understanding a gene's tolerance for loss-of-function mutations is likely a key to identifying mutations with strong phenotypic consequences. Larger and more systematic effort to characterize each gene's tolerance for loss-of-function mutations has recently been undertaken by the Exome Aggregation Consortium (ExAC) [15]. By amassing exomes from 60,706 individuals, the ExAC authors identify 3,230 genes that lack any known loss-of-function alleles – of note, 79% of these genes have no known disease phenotype. This list acts as a rich resource for identifying potential Mendelian disease genes when narrowing down possible targets for individuals with severe, hereditary rare diseases. Additionally, the presence of natural human knockouts provides a unique opportunity to understand the *in vivo* consequence of rare phenotypes, whether beneficial or detrimental, and further provides new opportunities to uncover potentially druggable genes that confer minimal individual risk.

However, a major concern with the detection of impactful loss-of-function mutations has been that such variants are enriched for false positives [16]. Dissecting loss-of-function variants from sequencing noise or from true mutations that have no functional consequence has involved defining strict thresholds for variant calling and leveraging high-quality gene annotations. Gene expression data from RNA-sequencing can aid in identifying the subset of loss-of-function mutations that induce functional consequences through nonsense-mediated decay (NMD) of the mRNA transcript [11]. For such analyses, an excess of allele-specific expression for loss-of-function alleles compared to their respective normal allele can inform degradation of the transcript by NMD. Studies of loss-of-function variation that have incorporated transcriptome data have demonstrated increased NMD surveillance for rare loss-of-function alleles compared to both common variants and predicted deleterious non-synonymous variants [17,18]. Furthermore, when studied across tissues, patterns of NMD for loss-of-function variants have exhibited only modest tissue heterogeneity with the vast majority of sites showing consistent patterns across tissues [17,19]. As will be further discussed for non-coding variants, these studies demonstrate how the application of both genome and transcriptome data provides increased utility to detect true loss-of-function alleles in personal genomes.

So far, the identification of loss-of-function variation has been largely restricted to protein-coding regions because exome sequencing is relatively cheap and annotating putative loss-of-function variants in protein-coding regions is relatively easy. These studies have guided our expectations of the frequency of loss-of-function alleles for particular genes and ongoing challenges with interpretation of their phenotypic consequences. To characterize the full scope of human loss-of-function variation will require interpreting the non-coding genome;

such variants have similar impacts on gene expression and protein abundance as coding loss-of-function variants but the identification of causal non-coding alleles from primary sequence data remains a major bottleneck.

Non-coding loss-of-function variation

Currently, most interpretable and actionable variants are discovered by sequencing efforts focused on the protein-coding regions of the genome – a small fraction comprising less than 2% of the entire genome. Rapid interpretation of protein-coding variants has had exceptional impact on identifying rare variants contributing to rare genetic diseases such as Miller syndrome [4], Kabuki syndrome [3], and Schinzel-Giedion syndrome [5]. Despite these successes, current clinical sequencing efforts only manage to identify a causal, pathogenic variant for about 25–50% of cases (Table 2), suggesting that non-coding loss-of-function variants may underlie the etiology of many rare diseases. This was well articulated in a genome-scale study of a family of four that identified a candidate locus for Miller syndrome and highlighted that an “unknown fraction of important phenotypes in humans are encoded by non-exonic variants identified only by means of whole-genome sequencing” [20].

So what evidence do we have for rare, pathogenic non-coding alleles? There are various mechanisms by which non-coding variants can elicit loss-of-function phenotypes, either by altering chromatin organization or disrupting proximal and distal regulatory elements (Table 1). Previous reviews have described a handful of examples of large-effect non-coding alleles underlying Mendelian disorders including beta-thalassemia and polydactyly [21,22]. A systematic analysis of disease variants curated within the Online Mendelian Inheritance in Man (OMIM) and ClinVar databases reported over 27,000 variants connected to Mendelian diseases [23]. While these variants are certain to be of varying quality [24], the authors reported that 29% were either upstream or downstream of their target gene, highlighting the fact that a significant fraction of variation associated to Mendelian disease resides in the non-coding genome. Indeed, specific examples of non-coding variants associated to Mendelian disorders continue to appear in the literature. Van Schil *et al.* identified four promoter mutations that alter gene expression of SAMD7 and may contribute to autosomal recessive retinitis pigmentosa [25]. Nellist *et al.* applied exome sequencing to mutation-negative individuals with tuberous sclerosis complex and identified compelling variants of unknown significance affecting promoter regions [26]. Likewise, Lin *et al.* applied exome sequencing to three mutation-negative familial adenomatous polyposis families to identify a promoter deletion in one family [27]. These authors highlight that many unresolved cases might be caused by yet-to-be discovered pathogenic non-coding variants and argue for wider genetic screening of these regions. Furthermore, additional pathogenic promoter and enhancer variants have been identified for autosomal dominant disorders such as congenital hereditary endothelial dystrophy 1 and adult-onset demyelinating leukodystrophy [28,29]. To date, however, there are no variants in non-coding regions that have a practice guideline curated by the ClinVar database or by the ACMG.

While there are several well-studied examples of large-effect non-coding variants involved in Mendelian disorders, large-effect non-coding variants could be the exception and not the rule. Whereas the genetic code provides an effective tool for identifying consequential

coding variants, there is no analogous resource for non-coding variants – that is, identifying non-coding mutations with strong priors on molecular consequences remains a significant challenge. While non-coding variants in promoters (like the examples given above) are often identified by exome sequencing, there is no specific method for identifying which of these mutations significantly impact expression from sequence data alone. Such problems are greatly exacerbated in more distal non-coding regions outside the core promoter where there is a generally weaker prior on variant function. Additionally, the functional significance of specific non-coding variants may only be exposed in specific cell- or tissue-types or under particular environmental conditions, making it exceedingly difficult to identify disease-causing non-coding variants from sequence data alone.

To identify functionally significant non-coding regions and characterize the distribution of their effect sizes, massively parallel reporter gene assays (MPRAs) have recently been used to simultaneously test the impact of thousands of mutations on gene expression. By applying MPRAs and focusing on two enhancer regions, Melnikov *et al.* observed mostly subtle changes in enhancer activity with the most extreme effects being centered on transcription factor binding sites and inducing 2 to 4-fold expression changes [30]. Similarly, Patwardhan *et al.* focused on three enhancers and reported that most expression changes were subtle, with only 3% inducing a fold change greater than 2-fold [31]. In contrast, by focusing on the *Rhodopsin* promoter, Kwasnieski *et al.* reported that over 86% of tested substitutions had significant effects on expression, some as great as 30-fold [32]. These distinctions may reflect technological differences or differences in the distribution of effects for targeted regulatory elements. However, each assay highlighted a small fraction of mutations that could induce at least a 2-fold change in expression.

So how do these results translate to genome-wide analyses or to known human variants? So far these studies have not been performed genome-wide and have focused only on a small portion of the non-coding genome. By specifically coupling human genetic variation data to MPRAs, Vockley *et al.* were able to more systematically test the effects of both common and rare non-coding variants in a region associated to adiposity at birth. From 283 testable variants, they identified 36 variants that had significant effects on expression – of these, four common variants had a greater than 2-fold effect on gene expression [33]. Ignoring the fact that multiple alleles are co-inherited, the consequence of an individual being homozygous for one of these variants would reduce their expression by half and may be functionally equivalent to carrying a single loss-of-function protein-coding allele. MPRAs demonstrate that such large-effect non-coding changes are indeed possible in every regulatory sequence surveyed.

Because multiple alleles can be co-inherited, it is likely that deciphering combinations of allelic effects will be increasingly important when interpreting loss-of-function mutations. It is widely accepted that the genomic context of any mutation is important for understanding both its penetrance and phenotypic expressivity. This was well shown by a recent study of predicted deleterious protein-coding alleles that demonstrated how genes can harbor compensatory mutations, negating their detrimental effects [34]. Likewise, deciphering the interactions between regulatory variants and protein-coding loss-of-function mutation is important for interpreting the impact of any loss-of-function mutation. Two recent studies of

Mendelian disorders described compound heterozygosity of promoter and protein-coding loss-of-function mutations [35,36]. In both diseases, affected individuals are haploinsufficient; they carry a protein-coding loss-of-function mutation on one allele coupled with a mutation that reduces expression on the other allele. Furthermore, genome-wide analyses of deleterious alleles have shown that potentially harmful protein-coding alleles are more often lowly expressed and such patterns are likely driven by differences in purifying selection influenced by the level of expression [17,37,38]. These observations highlight how interpretation of protein-coding loss-of-function variants will require interpreting their genomic context with respect to both protein-coding and non-coding alleles. This is even more important when interpreting variation in the non-coding genome – both expression quantitative trait loci (eQTL) studies and MPRAs have demonstrated how frequently multiple regulatory alleles can influence the expression of their co-inherited gene. As any additional regulatory allele can nudge a gene outside its desired range of activity, interpretation of non-coding loss-of-function variants will undoubtedly include subtle effects that need to be contextualized on the activity of all other regulatory alleles carried by an individual.

From exome to genome interpretation

The catalogue of potentially functional regulatory variants has rapidly expanded in recent years. Due to falling sequencing costs, it has become increasingly common for large genomics projects to sequence entire genomes rather than exomes. The final phase of the 1000 Genomes Project saw the release of genome sequencing data for 2,504 individuals sampled from global populations [2]. At the same time, large-scale genome sequencing efforts have been underway in specific populations across the globe, for example: 1,070 Japanese individuals [39]; 2,120 Sardinian individuals [40]; 2,636 Icelandic individuals [41]; and 3,621 British individuals [42]. All in all, almost 12,000 people have been sequenced as part of a large, published research project. In order to better understand the contribution of rare variants to human traits, even more ambitious sequencing efforts are currently underway. The Saudi Human Genome Project is planning to sequence 20,000 individuals, Genomics England has launched the 100,000 Genomes Project, and the United States has announced the Trans-Omics for Precision Medicine Initiative (TOPMed) and the One Million Genomes Project. Researchers in China are planning to sequence a million human genomes as part of the 3-Million Genome Project, and the Korean Genome Project has a roadmap to sequence 1,000,000 genomes by 2020 and increasingly larger numbers in the subsequent years. These large collections of whole genomes will enable researchers to better understand the allele frequency distribution of both protein-coding and non-coding alleles, the distribution of loss-of-function effects, and the role of rare and extremely rare variants in complex and Mendelian disease.

With so much sequencing underway, functional interpretation of novel variants is rapidly becoming the bottleneck in research. Computational methods for predicting the functional impact of variation can be broken up into two major classes. The first class of algorithms, such as SIFT [43] and PolyPhen-2 [44], aimed at identifying the impact of mutations in protein-coding regions of the genome. These methods take advantage of gene models, the genetic code, and conservation identified by multiple species alignment to identify putative

loss-of-function mutations that introduce premature stop codons (stop-gains), abrogate existing stop codons (stop-loss), alter splice donors/acceptors, or dramatically alter the biochemical properties of individual residues. These methods play critical roles in exome interpretation pipelines [45–47], but are, of course, limited in their application to the non-coding genome. One complexity with variant interpretation in the non-coding genome is that there are various mechanisms by which non-coding variation can function – while many common regulatory elements act in *cis* [48], *trans* and long-range interactions dramatically increase the number of potential regulatory variants for any given gene and these interactions can have a dramatic impact on disease risk [49–51]. Additionally, these regulatory effects can be specific to certain cell/tissue types, certain developmental stages, or environmental conditions [52–54], making it difficult to predict phenotypic variation (e.g. gene expression) as a function of sequence-level variation without additional data.

Despite these challenges, a significant amount of work has been done in order to develop methods that can more readily identify functional genetic variation, regardless of whether said variation is coding or non-coding. The first major research in this area focused on identifying functionally significant genomic regions using phylogenetics – that is, by comparing orthologous sequences across several species, methods such as PhyloP [55], phastCons [56], and GERP [57] can estimate the evolutionary constraint at individual base pairs. While such conservation metrics have been very successful at identifying regions of the genome under strong purifying selection (e.g. coding sequences), they fail to effectively identify functional regulatory variants that are generally under weaker purifying selection than coding variants [58].

To overcome this limitation, newer methods have sought to integrate conservation data with other relevant genomic annotations that have been produced by large-scale functional genomics projects like the Encyclopedia of DNA Elements (ENCODE) project [59] and the Roadmap Epigenomics project [60]. These two projects have produced genome-wide maps of transcription factor binding sites, chromatin accessibility, and various histone modifications across a wide range of human cell types. Together, these data can be used to demarcate regions of functional significance in the genome. The public release of these datasets has led to several novel computational methods that attempt to identify functionally significant variants. For example, the combined annotation-depletion dependent (CADD) software [61] and genome-wide annotation of variants (GWAVA) software [62] integrate a variety of functional data regarding chromatin state & accessibility, conservation, and transcription factor (TF) binding in order to build discriminative classifiers to identify functionally significant variants. Another method, fitCons [63], estimates the extent of selective pressure on similarly annotated regions of the genome as a proxy for the functional relevance of variants in these regions. Eigen [57] uses an unsupervised method to calculate a functional score for both coding and non-coding variants based on relevant annotations in each region, based on the assumption that variants are either functional or non-functional. Finally, the gkm-SVM [64] algorithm can identify functional regulatory variants in specific cell types after being trained on appropriate regulatory sequences (e.g. open chromatin regions for the cell-type of interest). It is important to note that many of these supervised learning programs (GWAVA, CADD, and gkm-SVM) rely on training data sets, which can be of varying quality and can limit technical performance in non-coding regions of the

genome. Altogether, this rapidly expanding set of tools is enabling the functional interpretation of both rare and common non-coding variants. For convenience, we provide a table summarizing the reference datasets and tools we have discussed and how to access them (Table 3).

Integrating functional genomics data to interpret rare alleles

Unfortunately, even with these new, integrative approaches it is still difficult to identify the regulatory impact of specific variants in the absence of additional functional data from carriers. In order to characterize variants that define the regulatory architecture of the human genome, there has been a significant push to identify eQTL in increasingly large and diverse cohorts, across different environmental contexts, and across human tissues [65–70]. However, these studies have largely focused on the impact of common variants on gene expression due to statistical limitations in how eQTLs are conventionally identified.

While methods for identifying rare regulatory variants are still evolving, the typical approach is to identify individuals that have extreme gene expression phenotypes – for example, an individual might significantly over- or under-express a gene or have a unique signature of allele-specific expression. Early studies of Europeans and other populations from the 1000 Genomes project found an excess of rare variants near genes with extreme expression phenotypes [37,65]. Since the frequency of rare variants is increased in related individuals, analyses of rare expression phenotypes in families can be particularly effective at identifying the impact of rare regulatory variants. That being said, however, it can still be difficult to identify the causal variant if there is linkage between rare variants segregating in a family. A study of a large European family demonstrated that family members who shared extreme regulatory phenotypes were more likely to share rare variants near affected genes and that genes affected by rare variants were more likely to be essential or associated with disease [71]. Recently, methods based on the identification of expression outliers have emerged [72] and have been used to identify autism genes where rare variants are associated with extreme variability in gene expression [73]. Similar analyses have been used to reveal the impact of rare regulatory variants and rare copy number variants in schizophrenia [74]. Furthermore, novel rare variant burden tests have demonstrated the dramatic enrichment of rare variants in the promoters of genes with outlier gene expression [75]. By using these outlier methods to identify rare regulatory variants and the growing number of reference transcriptomes being made available, it is increasingly possible to identify rare non-coding variants that cause rare Mendelian diseases [76,77].

Given the growing interest in using personal genomes and transcriptomes to diagnose rare disease, there is an increasing demand for methods that can more effectively identify rare regulatory variants. For example, current methods do not effectively integrate variability in total expression, allele-specific expression, or the effects of other common and regulatory alleles in order to identify rare regulatory variants. Additionally, improved efficiency of CRISPR/Cas9-mediated genome editing has made it increasingly feasible to validate putative rare regulatory variants [75]. While these experimental advances will surely drive the identification and validation of causal variants, increased data sharing amongst large genomics institutes through efforts like the Global Alliance for Genomics and Health

(GA4GH) is expected to enable interpretation of rare variants on a global scale. For instance, the GA4GH Beacon project allows researchers anywhere to identify if a rare variant they are interested in has been previously observed. Similar technology for securely sharing clinical measurements, transcriptomes, and other functional genomics data will create an invaluable resource for researchers trying to characterize the function and pathogenicity of rare variants.

Future perspective

With the prospect of millions of genomes on the horizon, integration of rare, non-coding variants in the interpretation of personal genomes will require new paradigms in genomic analysis. Key to successful interpretation will be the development of *in silico* whole genome interpretation approaches that leverage diverse and continually expanding genome and epigenome annotation data to propose candidate impactful alleles and their tissues, developmental-stages, and environments of context. Integrating the impact of these alleles into personal genetic risk profiles will require multi-omics assays that can indicate if the non-coding allele significantly affects gene expression. Furthermore, revisiting family studies will improve our power to identify the distribution of large, segregating effects on gene expression and decipher their causal genetic architecture. These approaches will leverage expanding genome and cellular engineering techniques to measure effects in specific genome and cellular contexts. For instance, one can easily foresee leveraging the expanding epigenomics compendium to identify a potentially impactful non-coding variant, determining its context of activity, and then engineer cells from the same individual to measure genomic activity in relevant stages or environments.

Interpreting non-coding loss-of-function variants will require moving away from the narrow view of specific sites to considering the genomic context of multiple variants. An individual who is homozygous for a common variant that lowers expression may be more impacted by a rare variant that further reduces activity than an individual who naturally carries the higher expressing alleles. Understanding these epistatic effects will require understanding the thresholds at which gene expression transitions outside of the normal range of activity. Furthermore, it will require moving away from single genes to gene networks and pathways in order to identify how various genes can compensate each other or cumulatively add to an individual's risk of disease. The study of non-coding variants will transition the study of single variants to systems of variants and genes undoubtedly providing higher resolution and understanding of the impact of our genome on our health.

Acknowledgments

Z.Z. is supported by the National Science Foundation (NSF) GRFP (DGE-114747). Z.Z. also acknowledges support from the National Institute of Health (NIH) (T32HG000044). S.B.M. is supported by the National Institutes of Health through R01HG008150, R01MH101814 and U01HG007436.

References

1. Keinan A, Clark AG. Recent explosive human population growth has resulted in an excess of rare genetic variants. Science. 2012 May.336:740–743. [PubMed: 22582263]

2. 1000 Genomes Project Consortium. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015 Oct 1.526:68–74. [PubMed: 26432245]
3. Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, Gildersleeve HI, et al. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nat Genet. 2010 Sep. 42:790–793. [PubMed: 20711175]
4. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, et al. Exome sequencing identifies the cause of a mendelian disorder. Nat Genet. 2010 Jan.42:30–35. [PubMed: 19915526]
5. Hoischen A, van Bon BWM, Gilissen C, Arts P, van Lier B, Steehouwer M, et al. De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. Nat Genet. 2010 Jun.42:483–485. [PubMed: 20436468]
6. Singh T, Kurki MI, Curtis D, Purcell SM, Crooks L, McRae J, et al. Rare loss-of-function variants in SETD1A are associated with schizophrenia and developmental disorders. Nat Neurosci. 2016 Mar 14.19:571–577. [PubMed: 26974950]
7. Rivas MA, Beaudoin M, Gardet A, Stevens C, Sharma Y, Zhang CK, et al. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat Genet. 2011 Nov 1.43:1066–1073. [PubMed: 21983784]
8. Chen R, Shi L, Hakenberg J, Naughton B, Sklar P, Zhang J, et al. Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases. Nat Biotechnol. 2016 Apr 11.34:531–538. [PubMed: 27065010]
9. Nelson MR, Wegmann D, Ehm MG, Kessner D, St Jean P, Verzilli C, et al. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science. 2012 Jul 6.337:100–104. [PubMed: 22604722]
10. Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science. 2012 Jul 6.337:64–69. [PubMed: 22604720]
11. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, et al. A systematic survey of loss-of-function variants in human protein-coding genes. Science. 2012 Feb 17.335:823–828. [PubMed: 22344438]
12. Lim ET, Würtz P, Havulinna AS, Palta P, Tukiainen T, Rehnström K, et al. Distribution and medical impact of loss-of-function variants in the Finnish founder population. PLoS Genet. 2014 Jul.10:e1004494. [PubMed: 25078778]
13. Narasimhan VM, Hunt KA, Mason D, Baker CL, Karczewski KJ, Barnes MR, et al. Health and population effects of rare gene knockouts in adult humans with related parents. Science. 2016 Mar 3.:aac8624.
14. Saleheen D, Natarajan P, Zhao W, Rasheed A, Khetarpal S, Won HH, et al. Human knockouts in a cohort with a high rate of consanguinity. 2015; doi: 10.1101/031518
15. Lek M, Karczewski K, Minikel E, Samocha K, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. bioRxiv. 2015 Oct 30.doi: 10.1101/030338
16. MacArthur DG, Tyler-Smith C. Loss-of-function variants in the genomes of healthy humans. Hum Mol Genet. 2010 Oct 15.19:R125–30. [PubMed: 20805107]
17. Kukurba KR, Zhang R, Li X, Smith KS, Knowles DA, How Tan M, et al. Allelic Expression of Deleterious Protein-Coding Variants across Human Tissues. PLoS Genet. 2014 May.10:e1004304. [PubMed: 24786518]
18. Lappalainen T, Sammeth M, Friedländer MR, 't Hoen PAC, Monlong J, Rivas MA, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013 Sep 26.501:506–511. [PubMed: 24037378]
19. Rivas MA, Pirinen M, Conrad DF, Lek M, Tsang EK, Karczewski KJ, et al. Human genomics. Effect of predicted protein-truncating genetic variants on the human transcriptome. Science. 2015 May 8.348:666–669. [PubMed: 25954003]
20. Roach JC, Glusman G, Smit AFA, Huff CD, Hubley R, Shannon PT, et al. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science. 2010 Apr 30.328:636–639. [PubMed: 20220176]

21. Li X, Montgomery SB. Detection and impact of rare regulatory variants in human disease. Front Genet. 2013; 4:67. [PubMed: 23755067]
22. Makrythanasis P, Antonarakis SE. Pathogenic variants in non-protein-coding sequences. Clin Genet. 2013 Nov.84:422–428. [PubMed: 24007299]
23. Ma M, Ru Y, Chuang L-S, Hsu N-Y, Shi L-S, Hakenberg J, et al. Disease-associated variants in different categories of disease located in distinct regulatory elements. BMC Genomics. 2015; 16(Suppl 8):S3.
24. Daneshjou R, Zappala Z, Kukurba K, Boyle SM, Ormond KE, Klein TE, et al. Path-scan: a reporting tool for identifying clinically actionable variants. Pac Symp Biocomput. 2014:229–240. [PubMed: 24297550]
25. Van Schil K, Karlstetter M, Aslanidis A, Dannhausen K, Azam M, Qamar R, et al. Autosomal recessive retinitis pigmentosa with homozygous rhodopsin mutation E150K and non-coding cis-regulatory variants in CRX-binding regions of SAMD7. Sci Rep. 2016; 6:21307. [PubMed: 26887858]
26. Nellist M, Brouwer RWW, Kockx CEM, van Veghel-Plandsoen M, Withagen-Hermans C, Prins-Bakker L, et al. Targeted Next Generation Sequencing reveals previously unidentified TSC1 and TSC2 mutations. BMC Med Genet. 2015; 16:10. [PubMed: 25927202]
27. Lin Y, Lin S, Baxter MD, Lin L, Kennedy SM, Zhang Z, et al. Novel APC promoter and exon 1B deletion and allelic silencing in three mutation-negative classic familial adenomatous polyposis families. Genome Medicine. 2015; 7:42. [PubMed: 25941542]
28. Davidson AE, Liskova P, Evans CJ, Dudakova L, Nosková L, Pontikos N, et al. Autosomal-Dominant Corneal Endothelial Dystrophies CHED1 and PPCD1 Are Allelic Disorders Caused by Non-coding Mutations in the Promoter of OVOL2. – PubMed – NCBI. The American Journal of Human Genetics. 2016 Jan.98:75–89. [PubMed: 26749309]
29. Giorgio E, Robyr D, Spielmann M, Ferrero E, Di Gregorio E, Imperiale D, et al. A large genomic deletion leads to enhancer adoption by the lamin B1 gene: a second path to autosomal dominant adult-onset demyelinating leukodystro. – PubMed – NCBI. Hum Mol Genet. 2015 May 8.24:3143–3154. [PubMed: 25701871]
30. Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, et al. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nat Biotechnol. 2012 Mar.30:271–277. [PubMed: 22371084]
31. Patwardhan RP, Lee C, Litvin O, Young DL, Pe'er D, Shendure J. High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. Nat Biotechnol. 2009 Dec. 27:1173–1175. [PubMed: 19915551]
32. Kwasnieski JC, Mogno I, Myers CA, Corbo JC, Cohen BA. Complex effects of nucleotide variants in a mammalian cis-regulatory element. Proc Natl Acad Sci USA. 2012 Nov 20.109:19498–19503. [PubMed: 23129659]
33. Vockley CM, Guo C, Majoros WH, Nodzenski M, Scholtens DM, Hayes MG, et al. Massively parallel quantification of the regulatory effects of noncoding genetic variation in a human cohort. Genome Res. 2015 Aug.25:1206–1214. [PubMed: 26084464]
34. Jordan DM, Frangakis SG, Golzio C, Cassa CA, Kurtzberg J, Task Force for Neonatal Genomics. et al. Identification of cis-suppression of human disease mutations by comparative genomics. Nature. 2015 Aug 13.524:225–229. [PubMed: 26123021]
35. Albers CA, Paul DS, Schulze H, Freson K, Stephens JC, Smethurst PA, et al. Compound inheritance of a low-frequency regulatory SNP and a rare null mutation in exon-junction complex subunit RBM8A causes TAR syndrome. Nat Genet. 2012 Apr.44:435–9. S1–2. [PubMed: 22366785]
36. Wieczorek D, Newman WG, Wieland T, Berulava T, Kaffe M, Falkenstein D, et al. Compound heterozygosity of low-frequency promoter deletions and rare loss-of-function mutations in TXNL4A causes Burn-McKeown syndrome. Am J Hum Genet. 2014 Dec 4.95:698–707. [PubMed: 25434003]
37. Montgomery SB, Lappalainen T, Gutierrez-Arcelus M, Dermitzakis ET. Rare and common regulatory variation in population-scale sequenced human genomes. PLoS Genet. 2011 Jul. 7:e1002144. [PubMed: 21811411]

38. Lappalainen T, Montgomery SB, Nica AC, Dermitzakis ET. Epistatic selection between coding and regulatory variation in human evolution and disease. Am J Hum Genet. 2011 Sep 9.89:459–463. [PubMed: 21907014]
39. Nagasaki M, Yasuda J, Katsuoka F, Nariai N, Kojima K, Kawai Y, et al. Rare variant discovery by deep whole-genome sequencing of 1,070 Japanese individuals. Nat Commun. 2015; 6:8018. [PubMed: 26292667]
40. Sidore C, Busonero F, Maschio A, Porcu E, Naitza S, Zoledziewska M, et al. Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. Nat Genet. 2015 Nov.47:1272–1281. [PubMed: 26366554]
41. Gudbjartsson DF, Helgason H, Gudjonsson SA, Zink F, Oddson A, Gylfason A, et al. Large-scale whole-genome sequencing of the Icelandic population. Nat Genet. 2015 May.47:435–444. [PubMed: 25807286]
42. UK10K Consortium. Walter K, Min JL, Huang J, Crooks L, Memari Y, et al. The UK10K project identifies rare variants in health and disease. Nature. 2015 Oct 1.526:82–90. [PubMed: 26367797]
43. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucl Acids Res. 2003 Jul 1.31:3812–3814. [PubMed: 12824425]
44. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet. 2013 Jan. Jan;Chapter 7:Unit7.20–7.20.41.
45. Li M-X, Gui H-S, Kwan JSH, Bao S-Y, Sham PC. A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. Nucl Acids Res. 2012 Apr.40:e53–e53. [PubMed: 22241780]
46. Li M-X, Kwan JSH, Bao S-Y, Yang W, Ho S-L, Song Y-Q, et al. Predicting mendelian disease-causing non-synonymous single nucleotide variants in exome sequencing studies. PLoS Genet. 2013; 9:e1003143. [PubMed: 23341771]
47. Robinson PN, Köhler S, Oellrich A, Sanger Mouse Genetics Project. Wang K, Mungall CJ, et al. Improved exome prioritization of disease genes through cross-species phenotype comparison. Genome Res. 2014 Feb.24:340–348. [PubMed: 24162188]
48. Veyrieras J-B, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, Stephens M, et al. High-Resolution Mapping of Expression-QTLs Yields Insight into Human Gene Regulation. PLoS Genet. 2008 Oct 10.4:e1000214. [PubMed: 18846210]
49. Westra H-J, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat Genet. 2013 Oct.45:1238–1243. [PubMed: 24013639]
50. Claussnitzer M, Dankel SN, Kim K-H, Quon G, Meuleman W, Haugen C, et al. FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. N Engl J Med. 2015 Sep 3.373:895–907. [PubMed: 26287746]
51. Flavahan WA, Drier Y, Liau BB, Gillespie SM, Venteicher AS, Stemmer-Rachamimov AO, et al. Insulator dysfunction and oncogene activation in IDH mutant gliomas. Nature. 2016 Jan 7.529:110–114. [PubMed: 26700815]
52. Lee MN, Ye C, Villani AC, Raj T, Li W, Eisenhaure TM, et al. Common genetic variants modulate pathogen-sensing responses in human dendritic cells. Science. 2014 Mar.343:1246980. [PubMed: 24604203]
53. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015 May 8.348:648–660. [PubMed: 25954001]
54. Spies N, Smith CL, Rodriguez JM, Baker JC, Batzoglou S, Sidow A. Constraint and divergence of global gene expression in the mammalian embryo. Elife. 2015; 4:e05538. [PubMed: 25871848]
55. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 2010 Jan.20:110–121. [PubMed: 19858363]
56. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005 Aug. 15:1034–1050. [PubMed: 16024819]
57. Cooper GM, Stone EA, Asimenos G, NISC Comparative Sequencing Program. Green ED, Batzoglou S, et al. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. 2005 Jul.15:901–913. [PubMed: 15965027]

58. ENCODE Project Consortium. Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, Gingeras TR, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007 Jun 14.447:799–816. [PubMed: 17571346]
59. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012 Sep 6.489:57–74. [PubMed: 22955616]
60. Roadmap Epigenomics Consortium. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015 Feb 19.518:317–330. [PubMed: 25693563]
61. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014 Mar.46:310–315. [PubMed: 24487276]
62. Ritchie GRS, Dunham I, Zeggini E, Flicek P. Functional annotation of noncoding sequence variants. Nat Methods. 2014 Mar.11:294–296. [PubMed: 24487584]
63. Gulko B, Hubisz MJ, Gronau I, Siepel A. A method for calculating probabilities of fitness consequences for point mutations across the human genome. Nat Genet. 2015 Mar.47:276–283. [PubMed: 25599402]
64. Lee D, Gorkin DU, Baker M, Strober BJ, Asoni AL, McCallion AS, et al. A method to predict the impact of regulatory variants from DNA sequence. Nat Genet. 2015 Aug.47:955–961. [PubMed: 26075791]
65. Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, et al. Transcriptome genetics using second generation sequencing in a Caucasian population. Nature. 2010 Apr 1.464:773–777. [PubMed: 20220756]
66. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature. 2010 Apr 1.464:768–772. [PubMed: 20220758]
67. Stranger BE, Montgomery SB, Dimas AS, Parts L, Stegle O, Ingle CE, et al. Patterns of cis regulatory variation in diverse human populations. PLoS Genet. 2012; 8:e1002639. [PubMed: 22532805]
68. Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, McCormick C, et al. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Res. 2014 Jan.24:14–24. [PubMed: 24092820]
69. Lee MN, Ye C, Villani AC, Raj T, Li W, Eisenhaure TM, et al. Common genetic variants modulate pathogen-sensing responses in human dendritic cells. Science. 2014 Mar.343:1246980. [PubMed: 24604203]
70. Fairfax BP, Humburg P, Makino S, Naranbhai V, Wong D, Lau E, et al. Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. Science. 2014 Mar 7.343:1246949–1246949. [PubMed: 24604202]
71. Li X, Battle A, Karczewski KJ, Zappala Z, Knowles DA, Smith KS, et al. Transcriptome sequencing of a large human family identifies the impact of rare noncoding variants. Am J Hum Genet. 2014 Sep 4.95:245–256. [PubMed: 25192044]
72. Zeng Y, Wang G, Yang E, Ji G, Brinkmeyer-Langford CL, Cai JJ. Aberrant gene expression in humans. PLoS Genet. 2015 Jan.11:e1004942. [PubMed: 25617623]
73. Guan J, Yang E, Yang J, Wang G, Zeng Y, Ji G, et al. Aberrant gene expression in autism. 2015; doi: 10.1101/029488
74. Duan J, Sanders AR, Moy W, Drigalenko EI, Brown EC, Freda J, et al. Transcriptome outlier analysis implicates schizophrenia susceptibility genes and enriches putatively functional rare genetic variants. Hum Mol Genet. 2015 May.28:ddv199.
75. Zhao J, Akinsanmi I, Arafat D, Cradick TJ, Lee CM, Banskota S, et al. A Burden of Rare Variants Associated with Extremes of Gene Expression in Human Peripheral Blood. Am J Hum Genet. 2016 Feb 4.98:299–309. [PubMed: 26849112]
76. Gonorazky H, Liang M, Cummings B, Lek M, Micallef J, Hawkins C, et al. RNAseq analysis for the diagnosis of muscular dystrophy. Ann Clin Transl Neurol. 2016 Jan.3:55–60. [PubMed: 26783550]

77. Zhang H, Xue C, Shah R, Bermingham K, Hinkle CC, Li W, et al. Functional analysis and transcriptomic profiling of iPSC-derived macrophages and their application in modeling Mendelian disease. Circ Res. 2015 Jun 19.117:17–28. [PubMed: 25904599]

Table 1

Common types of genetic variation and their ability to be discovered by exome and genome sequencing.

	Coding/Non-coding	Disruptive (* canonical loss-of-function)	Found by exome sequencing	Found by genome sequencing
5′/3′ UTR variation	Non-coding	No	Yes	Yes
Synonymous variation	Coding	No	Yes	Yes
Non-synonymous variation	Coding	Yes	Yes	Yes
Out-of-frame insertion/deletions	Coding	Yes*	Yes	Yes
Stop loss/gains	Coding	Yes*	Yes	Yes
In-frame insertion/deletions	Coding	Yes	Yes	Yes
Splice donor/acceptors	Coding	Yes*	Yes	Yes
Intronic variation	Non-coding	No	Sometimes	Yes
Upstream promoter variation	Non-coding	No	Sometimes	Yes
Proximal downstream variation	Non-coding	No	Sometimes	Yes
Distal enhancer/repressor variation	Non-coding	No	No	Yes
microRNA variation	Non-coding	No	No	Yes
cis-Regulatory element (CRE) variation	Non-coding	No	No	Yes
Large structural variation	Both	Yes	Sometimes	Yes

Table 2

Diagnostic success rates for clinical exome sequencing of patients with rare, undiagnosed disorders.

Diagnostic Yield	Year	DOI	Journal	Title
50% (inferred)	2012	10.1136/jmedgenet-2012-100819	Journal of Medical Genetics	Clinical application of exome sequencing in undiagnosed genetic conditions
16%	2012	10.1056/NEJMoa1206524	New England Journal of Medicine	Diagnostic Exome Sequencing in Persons with Severe Intellectual Disability
25%	2013	10.1056/NEJMoa1306555	New England Journal of Medicine	Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders
25.2%	2014	10.1001/jama.2014.14601	Journal of the American Medical Association	Molecular Findings Among Patients Referred for Clinical Whole-Exome Sequencing
26%	2014	10.1001/jama.2014.14604	Journal of the American Medical Association	Clinical Exome Sequencing for Genetic Identification of Rare Mendelian Disorders
41%	2014	10.1002/ana.24251	Annals of Neurology	Clinical whole exome sequencing in child neurology practice
24%	2015	10.1038/gim.2014.191	Genetics in Medicine	Whole-exome sequencing in undiagnosed genetic diseases: interpreting 119 trios
60%	2015	10.1007/s00439-015-1575-0	Human Genetics	High diagnostic yield of clinical exome sequencing in Middle Eastern patients with Mendelian disorders
25.8%	2016	10.1111/cge.12569	Clinical Genetics	Practical considerations in the clinical application of whole-exome sequencing
29%	2016	10.1016/j.mayocp.2015.12.018	Mayo Clinic Proceedings	Outcome of Whole Exome Sequencing for Diagnostic Odyssey Cases of an Individualized Medicine Clinic: The Mayo Clinic Experience
25–40%	2016	10.1111/cge.12654	Clinical Genetics	Utility of whole-exome sequencing for those near the end of the diagnostic odyssey: time to address gaps in care

Table 3

Referenced datasets and software.

Project	Data	URL
1000 Genomes	Genetic variation in human populations	http://www.1000genomes.org/
UK10K	Genetic variation in the UK with health records; data available upon request	http://www.uk10k.org/
ExAC	Collection of variants discovered by exome sequencing	http://exac.broadinstitute.org/
SIFT	Predicts impact of amino acid substitutions	http://sift.jcvi.org/
PolyPhen-2	Predicts impact of amino acid substitutions	http://genetics.bwh.harvard.edu/pph2/
phastCons	Measure of evolutionary conservation	Software: http://compgen.cshl.edu/phast/(br/)Data: http://hgdownload.cse.ucsc.edu/goldenpath/hg19/phastCons100way/
PhyloP	Measure of evolutionary conservation	Software: http://compgen.cshl.edu/phast/(br/)Data: http://hgdownload.cse.ucsc.edu/goldenpath/hg19/phyloP100way/
GERP	Measure of evolutionary conservation	http://mendel.stanford.edu/SidowLab/downloads/gerp/index.html
ENCODE	ChIP-seq data for various cell types	https://genome.ucsc.edu/ENCODE/
Roadmap Epigenomics	Epigenome data for various cell types (DNA methylation, histone modifications, etc.)	http://www.roadmapepigenomics.org/
CADD	Supervised method for predicting functional significance	http://cadd.gs.washington.edu/
Eigen	Unsupervised method for predicting functional variants	http://www.columbia.edu/~ii2135/eigen.html
fitCons	Measure of selective pressure	http://compgen.cshl.edu/fitCons/

Project	Data	URL
gkm-SVM	Supervised method for predicting functional significance	http://www.beerlab.org/gkmsvm/
GWAVA	Supervised method for predicting functional significance	https://www.sanger.ac.uk/sanger/StatGen_Gwava