COLD SPRING HARBOR
Molecular Case Studies

Diagnosing rare diseases after the exome

Laure Frésard[1] and Stephen B. Montgomery[1,2]

[1]Department of Pathology, Stanford University, Stanford, California 94305, USA; [2]Department of Genetics, School of Medicine, Stanford, California 94305, USA

Abstract High-throughput sequencing has ushered in a diversity of approaches for identifying genetic variants and understanding genome structure and function. When applied to individuals with rare genetic diseases, these approaches have greatly accelerated gene discovery and patient diagnosis. Over the past decade, exome sequencing has emerged as a comprehensive and cost-effective approach to identify pathogenic variants in the protein-coding regions of the genome. However, for individuals in whom exome-sequencing fails to identify a pathogenic variant, we discuss recent advances that are helping to reduce the diagnostic gap.

High-throughput sequencing has helped to uncover the rates at which deleterious or damaging alleles accumulate in specific genes (Lek et al. 2016). For patients with shared rare diseases, it has aided in identifying novel causal genes by their excess of pathogenic variants (Ng et al. 2010; Muona et al. 2015; Lelieveld et al. 2016; Deciphering Developmental Disorders Study 2017). In clinical settings, depending on disease type and patient selection, exome sequencing has been estimated to lead to a diagnosis in 30%–50% of rare Mendelian diseases (McInerney-Leo et al. 2013; Veeramah et al. 2013; Clark et al. 2018). However, diagnoses of diseases involving novel mechanisms (Oláhová et al. 2018) can be more challenging to perform than well described—although rare—pathologies (Aartsma-Rus et al. 2016). For "exome-negative" cases, in which no such diagnosis is provided, there remain multiple approaches after the exome—but choosing that next step in an "experimental maze" still remains a major challenge.

Negative exome results can be explained in various ways (Fig. 1). To summarize, the causal variant can be missed in cases of somatic mosaicism (Priest et al. 2016): If it is in coding regions but not properly detected (Cornish and Guda 2015); if it is of unknown significance and was not selected (McCarthy et al. 2014; Hoffman-Andrews 2017); if multiple variants are responsible for the pathology, in two or more genes (Liu et al. 2007; Sambuughin et al. 2018); if it is a structural variant not properly caught through exome sequencing (Merker et al. 2018); and in cases in which it is not exonic (Short et al. 2018).

Without prior biological knowledge, exome sequencing a single case has a low probability of yielding a novel pathogenic variant or gene. The average individual carries approximately 20 rare, "loss-of-function" variants (MAF ≤ 0.1) (Dewey et al. 2016; Lek et al. 2016) among which four are splice disrupting. Subsequent filtering of these candidate variants to identify disease genes is generally easier for recessive inherited diseases than dominant inherited diseases in which one causal variant is expected instead of two. However, in both scenarios, identifying which variants are most likely to be pathogenic has required either large control data sets (Dewey et al. 2016; Lek et al. 2016) or family-based sequencing data such as trio sequencing, particularly in the context of de novo variant discovery (Zhu et al. 2015;

Corresponding authors:
smontgom@stanford.edu;
lfresard@stanford.edu

Published by Cold Spring Harbor Laboratory Press

doi: 10.1101/mcs.a003392

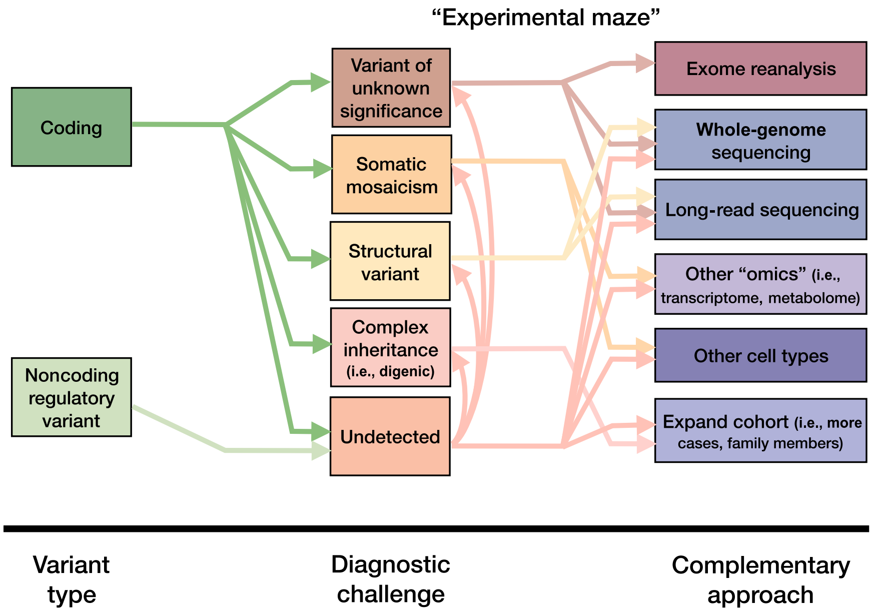

Figure 1. Challenges and approaches in "exome-negative" cases. Depending on the causal variant type, different options are available, influenced by both technical and biological diagnostic challenges. Without knowing the cause of the disease, it can be challenging to select a complementary approach in the postexome "experimental maze."

Wang et al. 2016; Jin et al. 2017; Wright et al. 2018a). Hence, one of the most active areas of development has been the ongoing expansion and development of large exome resources. Databases such as ExAC or gnomAD (Lek et al. 2016) and DiscoverEHR (Dewey et al. 2016) currently provide data on more than 100,000 exomes from diverse populations combined. Using the ExAC database, Lek et al. developed the pLI score that reflects each gene's intolerance to loss-of-function variation. This score provides priors on the likelihood of seeing impactful variants in specific genes within a healthy population sample and has been particularly informative for discovery of genes contributing to dominant Mendelian disorders (Eilbeck et al. 2017; Rao and Nelson 2018). Recently, Coban-Akdemir et al. (2018) developed an NMD escape intolerance score to identify genes in which nonsense alleles may contribute to gain of function. As these exome resources grow, new and refined scoring metrics that account for different forms of inheritance and variant impacts are expected to provide new opportunities for variant interpretation.

Beyond enhancing sample sizes of case and controls, additional biological knowledge that aids in identifying candidate causal genes takes advantage of ongoing annotation of gene and phenotype relationships. Extensive and accurate phenotyping is essential to establish a proper link between potential candidate genes and disease characteristics. Tools such as DECIPHER (Firth et al. 2009), the Matchmaker Exchange (Philippakis et al. 2015), GeneMatcher (Sobreira et al. 2015), or PhenomeCentral (Buske et al. 2015) are allowing to match cases with similar phenotypes and/or genotypic profiles to help the diagnosis of rare diseases. Databases such as Human Phenotype Ontology (HPO; Köhler et al. 2017), OMIM (Hamosh et al. 2005), and Orphanet (INSERM 1997) provide useful information to narrow down candidate genes. HPO links phenotypes to genes and diseases. OMIM (https://www.ncbi.nlm.nih.gov/omim/) and Orphanet (https://www.orpha.net) catalog known rare disease genes. Complementing these resources, the AMELIE tool has been developed to curate knowledge of association between genes and phenotypes through extensive literature curation (Birgmeier et al. 2017). However, it is important to underline that these databases can rely on different semantics, and the variability of clinical terms when describing similar clinical features can lead to less accurate analyses. For genetic variants, several tools that incorporate prior biological knowledge are used to assess potential variant impacts including SIFT (Ng and Henikoff 2003), PolyPhen (Adzhubei et al. 2013), CADD (Kircher et al. 2014), LoFTEE (MacArthur et al. 2012), and M-CAP (Jagadeesh et al. 2016). As these tools and databases improve and expand, subsequent reanalysis of exomes continues to show improvement of diagnostic rate over time (Ewans et al. 2018; Nambot et al. 2018; Al-Nabhani et al. 2018; Wright et al. 2018b).

An inescapable limitation to exome sequencing is that it only characterizes the subset of the genome that encodes protein-coding genes. With rapid reductions in sequencing costs and a relatively unbiased survey of an individual's genetic variation, whole-genome sequencing (WGS) is a promising approach after the exome. For protein-coding regions, WGS provides the opportunity to find variants that are poorly captured using exome sequencing alone. In particular, WGS allows detection of structural variants more reliably (Meienberg et al. 2016; Stavropoulos et al. 2016). In a study aiming at comparing the diagnostic yield between conventional genetic tests and WGS on 103 patients with suspected genetic disorders, it was demonstrated that 18 diagnoses would not have been possible with exome sequencing alone as both structural and nonexonic variants were identified in disease associated genes (Lionel et al. 2018). It is expected that as costs decrease, WGS follow-up to exome-negative cases or as a first-line approach will become a new standard for clinical genomics (Stavropoulos et al. 2016; Lionel et al. 2018).

Whole-genome sequencing further provides access to a wealth of noncoding variants. Each individual carries approximately 30,000 rare variants across their genome (Li et al. 2017). A subset of these is expected to have important impacts on gene expression or alternative splicing. Recent estimates of the contribution of de novo mutation in fetal enhancer regions to neurodevelopmental disorders has identified that up to 3% of "exome-negative" cases could be explained by de novo mutations in regulatory regions (Short et al. 2018). This study was further notable in that it estimated that tens to hundreds of thousands of whole genomes would be required to comprehensively analyze the contributions of variants in noncoding elements to neurodevelopmental disorders. Exacerbating this challenge, there remains a lack of approaches to interpret increasingly complex scenarios in which coding and noncoding alleles act together to cause or modify a rare disease. The use of WGS as a routine will also require the development of robust methods to identify structural variations across control populations.

Regardless of DNA sequencing approach, it is estimated that the diagnostic rate using exome or genome sequencing is plateauing between 35% and 50% (Taylor et al. 2015; Wortmann et al. 2015; Clark et al. 2018; Powis et al. 2018). There is a need for other approaches to highlight the molecular signature of the disease and therefore target more efficiently the causal gene(s). A potential route is to focus on the consequences of pathogenic variants on different cellular products, among which are gene transcripts, proteins, or metabolites. By adding those layers of information, it is possible to identify aberrant products or activities that may further narrow the list of candidate genes and variants.

RNA-sequencing (RNA-seq), for example, has now become the gold standard to measure RNA levels and quantify transcript diversity. RNA-seq shows considerable diagnostic promise for rare diseases, as it provides a measurement of the consequences of both coding and noncoding variants on gene expression levels and alternative splicing (Byron et al. 2016). Under the assumption that only one or very few genes are impacted, RNA-seq makes it possible to detect and narrow investigation to the subset of genes with aberrant expression or splicing in affected individuals when compared to unaffected controls. There is further growing evidence that rare, genetic variants influence these aberrant events, providing further data to localize causal variants (Zhao et al. 2016; Li et al. 2017; Pala et al. 2017). By focusing in on a small set of outlier genes and their regulatory elements, there can be significant power advantages to detecting causal noncoding variants compared to studies that use only DNA-based sequencing data alone. Using the DDD model, we have estimated that pinpointing recurrent outlier activity for a specific regulatory element can substantially reduce the number of genomes required to associate de novo mutations with disease (Short et al. 2018; Fig. 2A). In the context of rare disease, recent applications of RNA-seq have aided in the diagnosis of multiple Mendelian diseases (Estivill 2015; UK10K Consortium et al. 2015; Xiong et al. 2015; Cummings et al. 2017; Kernohan et al. 2017; Kremer et al. 2017). Notably, it has helped to detect pathogenic splice variants in neuromuscular and mitochondrial diseases (Cummings et al. 2017; Kremer et al. 2017) some of which were previously undetected by exome sequencing alone (Kernohan et al. 2017). However, RNA-seq for "exome-negative" cases has only been used in very specific tissues and subsets of diseases so far and there remains a need to understand the potential applications of RNA-seq in the diagnosis of Mendelian diseases in a broader context.

A potential limitation of RNA-seq (and other functional genomics assays) is that gene expression is dependent on environment, cell type, and state. This has direct consequences on the tissue to use for RNA-seq as a diagnostic tool for Mendelian diseases. Although any single tissue can be used for WGS, genes involved in a specific disease might not be expressed in a ubiquitous manner. Fortunately, the accumulation of RNA-seq studies in multiple tissues and databases provides an extensive resource to evaluate the expression status of a particular set of genes across a range of normal biological contexts (Su et al. 2004; Liu et al. 2008;

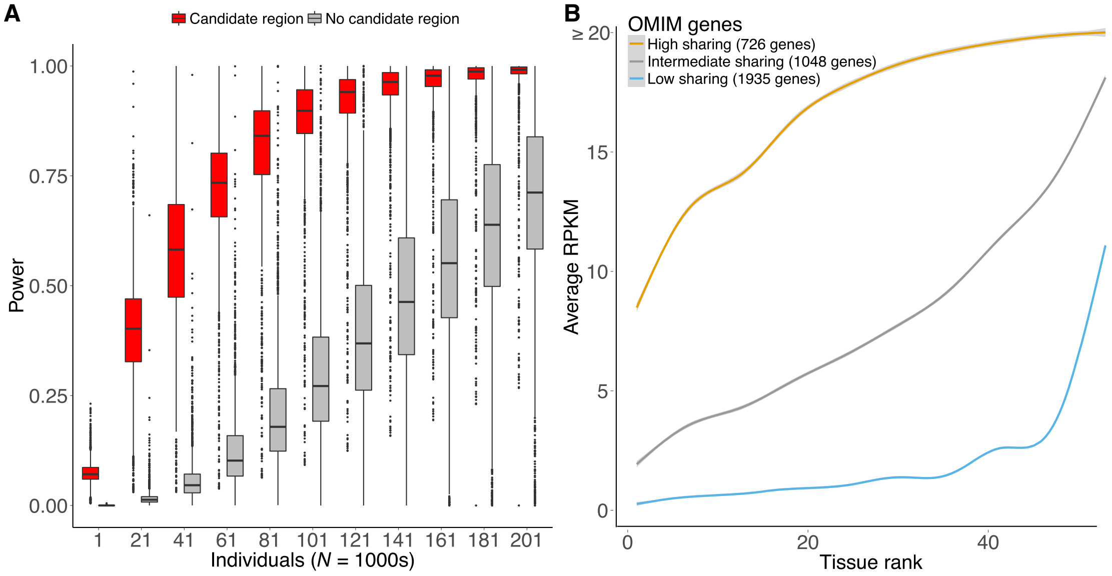

Figure 2. Using functional genomics to interpret rare diseases. (*A*) Power estimates for detecting a candidate regulatory region associated with a disease by the presence of recurrent de novo mutations. We compare power differences when there is a single candidate region (red) versus no candidate region (gray) using the model of Short et al. (2018). This demonstrates that additional biological knowledge of regions that are dysregulated in disease may significantly reduce the number of genomes required for their detection. (*B*) Ranked expression of OMIM genes across 53 tissues (GTEx v7). We used k-means to summarize expression data across all genes. Using three clusters, we show that we can separate our data as follow: high sharing, with high expression across all tissues (20% of OMIM genes); intermediate sharing, in which genes are expressed in a reasonable amount of tissues (28%); and low sharing, in which genes tend to be specifically expressed in a few tissues (52%). For 48% of OMIM genes, several tissues can be selected for RNA-seq analysis.

Krupp et al. 2012; GTEx Consortium 2013; Kim et al. 2014, 2018; Uhlén et al. 2015; Papatheodorou et al. 2018). Use of these data increases the number of normal samples and can aid in differentiating between rare events that are random versus biological. The recent work of Cummings and collaborators (Cummings et al. 2017) took advantage of such an approach by comparing 180 control skeletal muscle samples from the GTEx Project to 50 cases with muscle disorders. They noted that many of their discoveries could not have been found in blood gene expression alone. With resources that describe cell type expression and gene/phenotype relationships, we expect that future computational approaches will further aid in diagnosis by predicting the relevant cell types to study. Indeed, analysis of gene expression shows 20% of OMIM genes are expressed across multiple tissues (Fig. 2B).

Induced pluripotent stem cells (iPSCs) constitute another promising avenue to overcome tissue specificity (Sterneckert et al. 2014; Zhang et al. 2015; Yamasaki et al. 2017). iPSCs provide access to expression of genes of interest not expressed in the patient's most accessible tissues. They have been shown to be extremely useful in the context of heart disease (Bellin and Mummery 2016) when derived to cardiomyocytes. However, iPSC lines are known to be heterogeneous (Lund et al. 2012; Germain and Testa 2017); thus, in the context of rare diseases, it is essential to separate expression variability due to the cell line from that due to the disease. There is also a trade-off to be found between the time and cost to develop iPSC lines for a patient and the likelihood that genes of interest are expressed and involved in the mechanisms of disease. The growing availability of iPSC transcriptome data from multiple individuals constitutes a new promising source to help mitigate these challenges and guide the decision of generating iPSCs for specific cases (Kilpinen et al. 2017; Panopoulos et al. 2017).

Regardless of the follow-up approach, a major challenge resides in the rarity of a patient's sample. In contrast to association studies for complex traits in which we can evaluate a trait across data from multiple individuals, patients with rare disease are often isolated cases. For gene expression data, when comparing data from one case to multiple controls, one needs to disentangle what is caused by inherent noise in the sample itself from the actual variation due to the disease. Analyses are further impacted by the sample imbalance of case versus controls. Indeed, sample imbalance has been shown to impact differential expression results (Yang et al. 2006). Robust methods have been developed for genome-wide association analyses from single-case exomes (Wilfert et al. 2016). There is a need in adapting those methods for other data modalities. In practice, increasing the number of controls from different studies not only helps to identify aberrant signals in an $N = 1$ sample but can run the risk of increasing batch effects by adding variability from individual studies. A crucial step when combining data sets is therefore to perform an adequate correction for those hidden batch effects, without compromising the relevant information in each sample (Stegle et al. 2012; Risso et al. 2014; Brechtmann et al. 2018).

The challenge of what to do next for "exome-negative" cases resides mainly in how effectively we can detect signal in either noisy or heterogeneous data. This is where adding layers of information and combining clues from different data modalities and statistical approaches can help narrow down from multiple candidate genes to a handful of meaningful ones. Gene expression is a promising next-step approach but is only one potential window to look for the effects of variants in exome-negative cases. Some variants, like missense, will not necessarily affect gene expression. Other scales of analysis, like proteome (Costanzo et al. 2017), metabolome (Gülbakan et al. 2016), or epigenome (Bjornsson 2015), have the potential to give access to consequences of those other types of variants. Adding up layers of information will drastically increase the probability of finding the causal gene. However, there will be a trade-off between large-scale functional genomics or more targeted approaches to confirm a suspected mechanism of action. The balance between processing time and cost can be poorly defined with respect to the value of added biological knowledge.

In the era of personalized medicine and high-throughput technologies, opportunities to diagnose the causes of rare diseases are abundant. Future efforts will take advantage of a growing ability to identify and classify different categories of variants. Intriguingly, current comparisons of genome and exome sequencing have shown similar diagnostic rates (Clark et al. 2018). However, much of the focus remains on protein-coding "loss-of-function" alleles, and new approaches are still required to identify different classes and categories of pathogenic variants, notably in conserved noncoding regions. Indeed, we have observed scenarios in which longer-read technologies are providing access to structural variants missed through exome sequencing alone (Merker et al. 2018) or family-based data help to identify inherited expression outliers (Pala et al. 2017). Furthermore, growing functional genomics resources such as TOPMed and GTEx provide data from thousands of healthy individuals and multiple biological contexts. Increasing the scale of these resources in natural and ex vivo contexts like iPSCs will eventually enable characterization of functional events that are present at frequencies of 1 in 10,000 or less. As such data types improve diagnostic yield in monogenic disorders, we expect that they will also provide new tools to identify the molecular causes of variable penetrance and oligogenic disorders. For "exome-negative" cases, with ongoing access to diverse data types complemented by data from healthy controls, our knowledge and ability to diagnose rare diseases only has room to grow.

ADDITIONAL INFORMATION

Acknowledgments

This work was supported by National Institutes of Health (NIH) grants R01HG008150 (NoVa) and U01HG009080 (GSPAC). We thank Matthew T. Wheeler, Craig Smail, and anonymous reviewers for their critical review of the manuscript.

Competing Interest Statement

S.B.M. is on the SAB for Prime Genomics.

REFERENCES

Aartsma-Rus A, Ginjaar IB, Bushby K. 2016. The importance of genetic diagnosis for Duchenne muscular dystrophy. *J Med Genet* **53:** 145–151.

Adzhubei I, Jordan DM, Sunyaev SR. 2013. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* **Chapter 7:** Unit7.20.

Al-Nabhani M, Al-Rashdi S, Al-Murshedi F, Al-Kindi A, Al-Thihli K, Al-Saegh A, Al-Futaisi A, Al-Mamari W, Zadjali F, Al-Maawali A. 2018. Re-analysis of exome sequencing data of intellectual disability samples: yields and benefits. *Clin Genet* doi: 10.1111/cge.13438.

Bellin M, Mummery CL. 2016. Inherited heart disease—what can we expect from the second decade of human iPS cell research? *FEBS Lett* **590:** 2482–2493.

Birgmeier J, Haeussler M, Deisseroth CA, Jagadeesh KA, Ratner AJ, Guturu H, Wenger AM, Stenson PD, Cooper DN, Re C, et al. 2017. AMELIE accelerates Mendelian patient diagnosis directly from the primary literature. *bioRxiv* doi: 10.1101/171322.

Bjornsson HT. 2015. The Mendelian disorders of the epigenetic machinery. *Genome Res* **25:** 1473–1481.

Brechtmann F, Matuseviciute A, Mertes C, Yepez VA, Avsec Z, Herzog M, Bader DM, Prokisch H, Gagneur J. 2018. OUTRIDER: a statistical method for detecting aberrantly expressed genes in RNA sequencing data. *bioRxiv* doi: 10.1101/322149.

Buske OJ, Girdea M, Dumitriu S, Gallinger B, Hartley T, Trang H, Misyura A, Friedman T, Beaulieu C, Bone WP, et al. 2015. PhenomeCentral: a portal for phenotypic and genotypic matchmaking of patients with rare genetic diseases. *Hum Mutat* **36:** 931–940.

Byron SA, Van Keuren-Jensen KR, Engelthaler DM, Carpten JD, Craig DW. 2016. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nat Rev Genet* **17:** 257–271.

Clark MM, Stark Z, Farnaes L, Tan TY, White SM, Dimmock D, Kingsmore SF. 2018. Meta-analysis of the diagnostic and clinical utility of genome and exome sequencing and chromosomal microarray in children with suspected genetic diseases. *NPJ Genom Med* **3:** 16.

Coban-Akdemir Z, White JJ, Song X, Jhangiani SN, Fatih JM, Gambin T, Bayram Y, Chinn IK, Karaca E, Punetha J, et al. 2018. Identifying genes whose mutant transcripts cause dominant disease traits by potential gain-of-function alleles. *Am J Hum Genet* **103:** 171–187.

Cornish A, Guda C. 2015. A comparison of variant calling pipelines using genome in a bottle as a reference. *Biomed Res Int* **2015:** 456479.

Costanzo M, Zacchia M, Bruno G, Crisci D, Caterino M, Ruoppolo M. 2017. Integration of proteomics and metabolomics in exploring genetic and rare metabolic diseases. *Kidney Dis* **3:** 66–77.

Cummings BB, Marshall JL, Tukiainen T, Lek M, Donkervoort S, Foley AR, Bolduc V, Waddell LB, Sandaradura SA, O'Grady GL, et al. 2017. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Sci Transl Med* **9:** eaal5209.

Deciphering Developmental Disorders Study. 2017. Prevalence and architecture of de novo mutations in developmental disorders. *Nature* **542:** 433–438.

Dewey FE, Murray MF, Overton JD, Habegger L, Leader JB, Fetterolf SN, O'Dushlaine C, Van Hout CV, Staples J, Gonzaga-Jauregui C, et al. 2016. Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. *Science* **354:** aaf6814.

Eilbeck K, Quinlan A, Yandell M. 2017. Settling the score: variant prioritization and Mendelian disease. *Nat Rev Genet* **18:** 599–612.

Estivill X. 2015. Genetic variation and alternative splicing. *Nat Biotechnol* **33:** 357–359.

Ewans LJ, Schofield D, Shrestha R, Zhu Y, Gayevskiy V, Ying K, Walsh C, Lee E, Kirk EP, Colley A, et al. 2018. Whole-exome sequencing reanalysis at 12 months boosts diagnosis and is cost-effective when applied early in Mendelian disorders. *Genet Med* doi: 10.1038/gim.2018.39.

Firth HV, Richards SM, Bevan AP, Clayton S, Corpas M, Rajan D, Van Vooren S, Moreau Y, Pettett RM, Carter NP. 2009. DECIPHER: database of chromosomal imbalance and phenotype in humans using ensembl resources. *Am J Hum Genet* **84:** 524–533.

Germain PL, Testa G. 2017. Taming human genetic variability: transcriptomic meta-analysis guides the experimental design and interpretation of IPSC-based disease modeling. *Stem Cell Reports* **8:** 1784–1796.

GTEx Consortium. 2013. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45:** 580–585.

Gülbakan B, Özgül RK, Yüzbaşıoğlu A, Kohl M, Deigner H-P, Özgüç M. 2016. Discovery of biomarkers in rare diseases: innovative approaches by predictive and personalized medicine. *EPMA J* **7:** 24.

Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. 2005. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* **33:** D514–D517.

Hoffman-Andrews L. 2017. The known unknown: the challenges of genetic variants of uncertain significance in clinical practice. *J Law Biosci* **4:** 648–657.

INSERM. 1997. *Orphanet: an online database of rare diseases and orphan drugs.* http://www.orpha.net.

Jagadeesh KA, Wenger AM, Berger MJ, Guturu H, Stenson PD, Cooper DN, Bernstein JA, Bejerano G. 2016. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat Genet* **48:** 1581–1586.

Jin ZB, Wu J, Huang XF, Feng CY, Cai XB, Mao JY, Xiang L, Wu KC, Xiao X, Kloss BA, et al. 2017. Trio-based exome sequencing arrests de novo mutations in early-onset high myopia. *Proc Natl Acad Sci* **114:** 4219–4224.

Kernohan KD, Frésard L, Zappala Z, Hartley T, Smith KS, Wagner J, Xu H, McBride A, Bourque PR, Consortium CRC, et al. 2017. Whole-transcriptome sequencing in blood provides a diagnosis of spinal muscular atrophy with progressive myoclonic epilepsy. *Hum Mutat* **38:** 611–614.

Kilpinen H, Goncalves A, Leha A, Afzal V, Alasoo K, Ashford S, Bala S, Bensaddek D, Casale FP, Culley OJ, et al. 2017. Common genetic variation drives molecular heterogeneity in human IPSCs. *Nature* **546:** 370–375.

Kim M-S, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, Madugundu AK, Kelkar DS, Isserlin R, Jain S, et al. 2014. A draft map of the human proteome. *Nature* **509:** 575–581.

Kim P, Park A, Han G, Sun H, Jia P, Zhao Z. 2018. TissGDB: tissue-specific gene database in cancer. *Nucleic Acids Res* **46:** D1031–D1038.

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46:** 310–315.

Köhler S, Vasilevsky NA, Engelstad M, Foster E, McMurry J, Aymé S, Baynam G, Bello SM, Boerkoel CF, Boycott KM, et al. 2017. The human phenotype ontology in 2017. *Nucleic Acids Res* **45:** D865–D876.

Kremer LS, Bader DM, Mertes C, Kopajtich R, Pichler G, Iuso A, Haack TB, Graf E, Schwarzmayr T, Terrile C, et al. 2017. Genetic diagnosis of Mendelian disorders via RNA sequencing. *Nature Commun* **8:** 15824.

Krupp M, Marquardt JU, Sahin U, Galle PR, Castle J, Teufel A. 2012. RNA-Seq Atlas—a reference database for gene expression profiling in normal tissue by next-generation sequencing. *Bioinformatics* **28:** 1184–1185.

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, et al. 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536:** 285–291.

Lelieveld SH, Reijnders MR, Pfundt R, Yntema HG, Kamsteeg E-J, de Vries P, de Vries BB, Willemsen MH, Kleefstra T, Löhner K, et al. 2016. Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability. *Nat Neurosci* **19:** 1194–1196.

Li X, Kim Y, Tsang EK, Davis JR, Damani FN, Chiang C, Hess GT, Zappala Z, Strober BJ, Scott AJ, et al. 2017. The impact of rare variation on gene expression across tissues. *Nature* **550:** 239–243.

Lionel AC, Costain G, Monfared N, Walker S, Reuter MS, Hosseini SM, Thiruvahindrapuram B, Merico D, Jobling R, Nalpathamkalam T, et al. 2018. Improved diagnostic yield compared with targeted gene sequencing panels suggests a role for whole-genome sequencing as a first-tier genetic test. *Genet Med* **20:** 435–443.

Liu Y, Zaghloul NA, Katsanis N. 2007. Bardet-Biedl syndrome, an oligogenic disease. In *Encyclopedia of life sciences* (ed. John Wiley & Sons, Ltd). Wiley, Chichester.

Liu X, Yu X, Zack DJ, Zhu H, Qian J. 2008. TiGER: a database for tissue-specific gene expression and regulation. *BMC Bioinformatics* **9:** 271.

Lund RJ, Närvä E, Lahesmaa R. 2012. Genetic and epigenetic stability of human pluripotent stem cells. *Nat Rev Genet* **13:** 732–744.

MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, Jostins L, Habegger L, Pickrell JK, Montgomery SB, et al. 2012. A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335:** 823–828.

McCarthy DJ, Humburg P, Kanapin A, Rivas MA, Gaulton K, Cazier JB, Donnelly P. 2014. Choice of transcripts and software has a large effect on variant annotation. *Genome Med* **6:** 26.

McInerney-Leo AM, Marshall MS, Gardiner B, Coucke PJ, van Laer L, Loeys BL, Summers KM, Symoens S, West JA, West MJ, et al. 2013. Whole exome sequencing is an efficient, sensitive and specific method of mutation detection in osteogenesis imperfecta and Marfan syndrome. *Bonekey Rep* **2:** 456.

Meienberg J, Bruggmann R, Oexle K, Matyas G. 2016. Clinical sequencing: is WGS the better WES? *Hum Genet* **135:** 359–362.

Merker JD, Wenger AM, Sneddon T, Grove M, Zappala Z, Fresard L, Waggott D, Utiramerur S, Hou Y, Smith KS, et al. 2018. Long-read genome sequencing identifies causal structural variation in a Mendelian disease. *Genet Med* **20:** 159–163.

Muona M, Berkovic SF, Dibbens LM, Oliver KL, Maljevic S, Bayly MA, Joensuu T, Canafoglia L, Franceschetti S, Michelucci R, et al. 2015. A recurrent de novo mutation in *KCNC1* causes progressive myoclonus epilepsy. *Nat Genet* **47:** 39–46.

Nambot S, Thevenon J, Kuentz P, Duffourd Y, Tisserant E, Bruel A-L, Mosca-Boidron A-L, Masurel-Paulet A, Lehalle D, Jean-Marçais N, et al. 2018. Clinical whole-exome sequencing for the diagnosis of rare disorders with congenital anomalies and/or intellectual disability: substantial interest of prospective annual reanalysis. *Genet Med* **20:** 645–654.

Ng PC, Henikoff S. 2003. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31:** 3812–3814.

Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, et al. 2010. Exome sequencing identifies the cause of a Mendelian disorder. *Nat Genet* **42:** 30–35.

Oláhová M, Yoon WH, Thompson K, Jangam S, Fernandez L, Davidson JM, Kyle JE, Grove ME, Fisk DG, Kohler JN, et al. 2018. Biallelic mutations in *ATP5F1D*, which encodes a subunit of ATP synthase, cause a metabolic disorder. *Am J Hum Genet* **102:** 494–504.

Pala M, Zappala Z, Marongiu M, Li X, Davis JR, Cusano R, Crobu F, Kukurba KR, Gloudemans MJ, Reinier F, et al. 2017. Population- and individual-specific regulatory variation in Sardinia. *Nat Genet* **49:** 700–707.

Panopoulos AD, D'Antonio M, Benaglio P, Williams R, Hashem SI, Schuldt BM, DeBoever C, Arias AD, Garcia M, Nelson BC, et al. 2017. iPSCORE: a resource of 222 iPSC lines enabling functional characterization of genetic variation across a variety of cell types. *Stem Cell Reports* **8:** 1086–1100.

Papatheodorou I, Fonseca NA, Keays M, Tang YA, Barrera E, Bazant W, Burke M, Füllgrabe A, Fuentes AM, George N, et al. 2018. Expression Atlas: gene and protein expression across multiple studies and organisms. *Nucleic Acids Res* **46:** D246–D251.

Philippakis AA, Azzariti DR, Beltran S, Brookes AJ, Brownstein CA, Brudno M, Brunner HG, Buske O, Carey K, Doll C, et al. 2015. The Matchmaker Exchange: a platform for rare disease gene discovery. *Hum Mutat* **36:** 915–921.

Powis Z, Farwell Hagman KD, Speare V, Cain T, Blanco K, Mowlavi LS, Mayerhofer EM, Tilstra D, Vedder T, Hunter JM, et al. 2018. Exome sequencing in neonates: diagnostic rates, characteristics, and time to diagnosis. *Genet Med* doi: 10.1038/gim.2018.11.

Priest JR, Gawad C, Kahlig KM, Yu JK, O'Hara T, Boyle PM, Rajamani S, Clark MJ, Garcia ST, Ceresnak S, et al. 2016. Early somatic mosaicism is a rare cause of long-QT syndrome. *Proc Natl Acad Sci* **113:** 11555–11560.

Rao AR, Nelson SF. 2018. Calculating the statistical significance of rare variants causal for Mendelian and complex disorders. *BMC Med Genomics* **11:** 53.

Risso D, Ngai J, Speed TP, Dudoit S. 2014. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat Biotechnol* **32:** 896–902.

Sambuughin N, Mungunsukh O, Ren M, Capacchione JF, Horkayne-Szakaly I, Chuang K, Muldoon SM, Smith JK, O'Connor FG, Deuster PA. 2018. Pathogenic and rare deleterious variants in multiple genes suggest oligogenic inheritance in recurrent exertional rhabdomyolysis. *Mol Genet Metab Rep* **16:** 76–81.

Short PJ, McRae JF, Gallone G, Sifrim A, Won H, Geschwind DH, Wright CF, Firth HV, FitzPatrick DR, Barrett JC, et al. 2018. De novo mutations in regulatory elements in neurodevelopmental disorders. *Nature* **555:** 611–616.

Sobreira N, Schiettecatte F, Valle D, Hamosh A. 2015. GeneMatcher: a matching tool for connecting investigators with an interest in the same gene. *Human Mutat* **36:** 928–930.

Stavropoulos DJ, Merico D, Jobling R, Bowdin S, Monfared N, Thiruvahindrapuram B, Nalpathamkalam T, Pellecchia G, Yuen RKC, Szego MJ, et al. 2016. Whole genome sequencing expands diagnostic utility and improves clinical management in pediatric medicine. *NPJ Genom Med* **1:** 15012.

Stegle O, Parts L, Piipari M, Winn J, Durbin R. 2012. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat Protoc* **7:** 500–507.

Sterneckert JL, Reinhardt P, Schöler HR. 2014. Investigating human disease using stem cell models. *Nat Rev Genet* **15:** 625–639.

Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, et al. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci* **101:** 6062–6067.

Taylor JC, Martin HC, Lise S, Broxholme J, Cazier JB, Rimmer A, Kanapin A, Lunter G, Fiddy S, Allan C, et al. 2015. Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. *Nat Genet* **47:** 717–726.

Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson Å, Kampf C, Sjöstedt E, Asplund A, et al. 2015. Proteomics. Tissue-based map of the human proteome. *Science* **347:** 1260419.

UK10K Consortium, Walter K, Min JL, Huang J, Crooks L, Memari Y, McCarthy S, Perry JR, Xu C, Futema M, et al. 2015. The UK10K project identifies rare variants in health and disease. *Nature* **526:** 82–90.

Veeramah KR, Johnstone L, Karafet TM, Wolf D, Sprissler R, Salogiannis J, Barth-Maron A, Greenberg ME, Stuhlmann T, Weinert S, et al. 2013. Exome sequencing reveals new causal mutations in children with epileptic encephalopathies. *Epilepsia* **54:** 1270–1281.

Wang C, Sun L, Zheng H, Hu YQ. 2016. Detecting multiple variants associated with disease based on sequencing data of case-parent trios. *J Hum Genet* **61:** 851–860.

Wilfert AB, Chao KR, Kaushal M, Jain S, Zöllner S, Adams DR, Conrad DF. 2016. Genome-wide significance testing of variation from single case exomes. *Nat Genet* **48:** 1455–1461.

Wortmann SB, Koolen DA, Smeitink JA, van den Heuvel L, Rodenburg RJ. 2015. Whole exome sequencing of suspected mitochondrial patients in clinical practice. *J Inherit Metab Dis* **38:** 437–443.

Wright CF, FitzPatrick DR, Firth HV. 2018a. Paediatric genomics: diagnosing rare disease in children. *Nat Rev Genet* **19:** 253–268.

Wright CF, McRae JF, Clayton S, Gallone G, Aitken S, FitzGerald TW, Jones P, Prigmore E, Rajan D, Lord J, et al. 2018b. Making new genetic diagnoses with old data: iterative reanalysis and reporting from genome-wide data in 1,133 families with developmental disorders. *Genet Med* doi: 10.1038/gim.2017.246.

Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RK, Hua Y, Gueroussov S, Najafabadi HS, Hughes TR, et al. 2015. RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. *Science* **347:** 1254806.

Yamasaki AE, Panopoulos AD, Belmonte JCI. 2017. Understanding the genetics behind complex human disease with large-scale IPSC collections. *Genome Biol* **18:** 135.

Yang K, Li J, Gao H. 2006. The impact of sample imbalance on identifying differentially expressed genes. *BMC Bioinformatics* **7:** S8.

Zhang H, Xue C, Shah R, Bermingham K, Hinkle CC, Li W, Rodrigues A, Tabita-Martinez J, Millar JS, Cuchel M, et al. 2015. Functional analysis and transcriptomic profiling of IPSC-derived macrophages and their application in modeling Mendelian disease. *Circulation Res* **117:** 17–28.

Zhao J, Akinsanmi I, Arafat D, Cradick TJ, Lee CM, Banskota S, Marigorta UM, Bao G, Gibson G. 2016. A burden of rare variants associated with extremes of gene expression in human peripheral blood. *Am J Hum Genet* **98:** 299–309.

Zhu X, Petrovski S, Xie P, Ruzzo EK, Lu YF, McSweeney KM, Ben-Zeev B, Nissenkorn A, Anikster Y, Oz-Levi D, et al. 2015. Whole-exome sequencing in undiagnosed genetic diseases: interpreting 119 trios. *Genet Med* **17:** 774–781.