



Published in final edited form as:

Nat Med. 2019 June ; 25(6): 911–919. doi:10.1038/s41591-019-0457-8.

Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts

Laure Frésard^{1,*}, Craig Smail², Nicole M. Ferraro², Nicole A. Teran³, Xin Li¹, Kevin S. Smith¹, Devon Bonner⁴, Kristin D. Kernohan⁵, Shruti Marwaha^{4,6}, Zachary Zappala³, Brunilda Balliu¹, Joe R. Davis³, Boxiang Liu⁷, Cameron J. Prybol³, Jennefer N. Kohler⁴, Diane B. Zastrow⁴, Chloe M. Reuter⁴, Dianna G. Fisk⁸, Megan E. Grove⁸, Jean M. Davidson⁴, Taila Hartley¹⁰, Ruchi Joshi⁸, Benjamin J. Strober⁹, Sowmithri Utiramerur⁸, Care4Rare Canada Consortium¹⁰, Undiagnosed Diseases Network¹¹, Lars Lind¹², Erik Ingelsson^{6,13}, Alexis Battle^{9,14}, Gill Bejerano^{15,16,17,18}, Jonathan A. Bernstein¹⁶, Euan A.

Users may view, print, copy, and download text and data-mine the content in such documents, for the purposes of academic research, subject always to the full Conditions of use:http://www.nature.com/authors/editorial_policies/license.html#terms

*Corresponding authors: Laure Frésard (lfrésard@stanford.edu) and Stephen B Montgomery (smontgom@stanford.edu).

Members of the Undiagnosed Diseases Network; David R. Adams; Aaron Aday; Mercedes E. Alejandro; Patrick Allard; Euan A. Ashley; Mahshid S. Azamian; Carlos A. Bacino; Eva Baker; Ashok Balasubramanyam; Hayk Barseghyan; Gabriel F. Batzli; Alan H. Beggs; Babak Behnam; Hugo J. Bellen; Jonathan A. Bernstein; Gerard T. Berry; Anna Bican; David P. Bick; Camille L. Birch; Devon Bonner; Braden E. Boone; Bret L. Bostwick; Lauren C. Briere; Elly Brokamp; Donna M. Brown; Matthew Brush; Elizabeth A. Burke; Lindsay C. Burrage; Manish J. Butte; Shan Chen; Gary D. Clark; Terra R. Coakley; Joy D. Cogan; Heather A. Colley; Cynthia M. Cooper; Heidi Cope; William J. Craigen; Precilla D'Souza; Mariska Davids; Jean M. Davidson; Jyoti G. Dayal; Esteban C. Dell'Angelica; Shweta U. Dhar; Katrina M. Dipple; Laurel A. Donnell-Fink; Naghmeh Dorrani; Daniel C. Dorset; Emilie D. Douine; David D. Draper; Annika M. Dries; Laura Duncan; David J. Eckstein; Lisa T. Emrick; Christine M. Eng; Gregory M. Enns; Ascia Eskin; Cecilia Esteves; Tyra Estwick; Liliana Fernandez; Carlos Ferreira; Elizabeth L. Fieg; Paul G. Fisher; Brent L. Fogel; Noah D. Friedman; William A. Gahl; Emily Glanton; Rena A. Godfrey; Alica M. Goldman; David B. Goldstein; Sarah E. Gould; Jean-Philippe F. Gourdine; Catherine A. Groden; Andrea L. Gropman; Melissa Haendel; Rizwan Hamid; Neil A. Hanchard; Frances High; Ingrid A. Holm; Jason Hom; Ellen M. Howerton; Yong Huang; Fariha Jamal; Yong-hui Jiang; Jean M. Johnston; Angela L. Jones; Lefkthea Karaviti; David M. Koeller; Isaac S. Kohane; Jennefer N. Kohler; Donna M. Krasnewich; Susan Korrick; Mary Kozlura; Joel B. Krier; Jennifer E. Kyle; Seema R. Lalani; C. Christopher Lau; Jozef Lazar; Kimberly LeBlanc; Brendan H. Lee; Hane Lee; Shawn E. Levy; Richard A. Lewis; Sharyn A. Lincoln; Sandra K. Loo; Joseph Loscalzo; Richard L. Maas; Ellen F. Macnamara; Calum A. MacRae; Valerie V. Maduro; Marta M. Majcherska; May Christine V. Malicdan; Laura A. Mamounas; Teri A. Manolio; Thomas C. Markello; Ronit Marom; Martin G. Martin; Julian A. Martínez-Agosto; Shruti Marwaha; Thomas May; Allyn McConkie-Rosell; Colleen E. McCormack; Alexa T. McCray; Jason D. Merker; Thomas O. Metz; Matthew Might; Paolo M. Moretti; Marie Morimoto; John J. Mulvihill; David R. Murdock; Jennifer L. Murphy; Donna M. Muzny; Michele E. Nehrebecky; Stan F. Nelson; J. Scott Newberry; John H. Newman; Sarah K. Nicholas; Donna Novacic; Jordan S. Orange; James P. Orengo; J. Carl Pallais; Christina GS. Palmer; Jeanette C. Papp; Neil H. Parker; Loren DM. Pena; John A. Phillips III; Jennifer E. Posey; John H. Postlethwait; Lorraine Potocki; Barbara N. Pusey; Genecee Renteria; Chloe M. Reuter; Lynette Rives; Amy K. Robertson; Lance H. Rodan; Jill A. Rosenfeld; Jacinda B. Sampson; Susan L. Samson; Kelly Schoch; Daryl A. Scott; Lisa Shakachite; Prashant Sharma; Vandana Shashi; Rebecca Signer; Edwin K. Silverman; Janet S. Sinshheimer; Kevin S. Smith; Rebecca C. Spillmann; Joan M. Stoler; Nicholas Stong; Jennifer A. Sullivan; David A. Sweetser; Queenie K.-G. Tan; Cynthia J. T₁; Camilo Toro; Alyssa A. Tran; Tiina K. Urv; Eric Vilain; Tiphonie P. Vogel; Daryl M. Waggott; Colleen E. Wahl; Nicole M. Walley; Chris A. Walsh; Melissa Walker; Jijun Wan; Michael F. Wangler; Patricia A. Ward; Katrina M. Waters; Bobbie-Jo M. Webb-Robertson; Monte Westerfield; Matthew T. Wheeler; Anastasia L. Wise; Lynne A. Wolfe; Elizabeth A. Worthey; Shinya Yamamoto; John Yang; Yaping Yang; Amanda J. Yoon; Guoyun Yu; Diane B. Zastrow; Chunli Zhao; Allison Zheng

Members of the Care4Rare Consortium: Steering Committee: Kym Boycott (lead; University of Ottawa), Alex MacKenzie (co-lead; University of Ottawa), Jacek Majewski (McGill University), Michael Brudno (University of Toronto), Dennis Bulman (University of Ottawa), David Dymnt (University of Ottawa).

Author Contributions

S.B.M., M.T.W., J.D.M., E.A.A. and K.M.B. conceived and planned the experiments. K.S.M., D.B., J.N.K., D.B.Z., D.G.F., M.E.G., C.M.R., J.M.D., R.J. contributed to sample preparation and cases review. L.L. and E.I. provided phenotypic data together with blood RNA-seq of PIVUS control samples. S.M., X.L., K.K., R.J., S.U. helped processing the variant data. L.F., C.S., N.M.F., N.A.T., Z.Z., X.L., B.B., J.R.D., B.L. carried out the analyses. K.D.K., B.J.S., A.B., G.B. and J.A.B. contributed to the interpretation of the results. K.D.K., T.H., C.J.P., D.B., J.N.K., D.Z., D.G.F., M.E.G. performed the validation of results. L.F. and S.B.M. wrote the manuscript with support from C.S., N.M.F. and N.A.T. All authors provided critical feedback and helped shape the research, analysis and manuscript.

Competing Interests Statement

J.D.M. is on Genoox Scientific Advisory Board and Rainbow Genomics Clinical Advisory Board and consults for Illumina. E.A.A. is Co-Founder to Personalis, DeepCell and Advisor to Genome Medical and Sequence Bio. E.I. is a scientific advisor for Precision Wellness for work unrelated to the present project. S.B.M. is on the Scientific Advisory Board for Prime Genomics.

Ashley^{3,4,13}, Kym M. Boycott¹⁰, Jason D. Merker^{1,8,19}, Matthew T. Wheeler^{4,6}, and Stephen B. Montgomery^{1,3,*}

¹Department of Pathology, Stanford University, Stanford, California, USA ²Biomedical Informatics Program, Stanford, California, USA ³Department of Genetics, School of Medicine, Stanford, California, USA ⁴Stanford Center for Undiagnosed Diseases, Stanford, California, USA ⁵Newborn Screening Ontario (NSO), Children's Hospital of Eastern Ontario, Ottawa, Ontario, Canada ⁶Stanford Cardiovascular Institute, Stanford University, Stanford, California, USA ⁷Department of Biology, School of Humanities and Sciences, Stanford, California, USA ⁸Stanford Medicine Clinical Genomics Program, Stanford, California, USA ⁹Department of Biomedical Engineering, Johns Hopkins University, Baltimore, Maryland, USA ¹⁰Children's Hospital of Eastern Ontario Research Institute, University of Ottawa, Ottawa, Ontario, Canada ¹¹NIH Undiagnosed Diseases Network, National Institutes of Health, Bethesda, Maryland, USA ¹²Department of Medical Sciences, Cardiovascular epidemiology, Uppsala University, Sweden ¹³Department of Medicine, Division of Cardiovascular Medicine, School of Medicine, Stanford, California, USA ¹⁴Department of Computer Science, Johns Hopkins University, Baltimore, Maryland, USA ¹⁵Department of Computer Science, Stanford, California, USA ¹⁶Department of Pediatrics, Stanford School of Medicine, Stanford, California, USA ¹⁷Department of Developmental Biology, Stanford School of Medicine, Stanford, California, USA ¹⁸Department of Biomedical Data Science, Stanford School of Medicine, Stanford, California, USA ¹⁹Present address: Departments of Pathology and Laboratory Medicine & Genetics, Lineberger Comprehensive Cancer Center, University of North Carolina School of Medicine, Chapel Hill, North Carolina, USA

NOTE

Introductory Paragraph

It is estimated that 350 million individuals worldwide suffer from rare diseases, which are predominantly caused by mutation in a single gene¹. The current molecular diagnostic rate is estimated at 50% with whole exome sequencing (WES) among the most successful approaches²⁻⁵. For patients in whom WES is uninformative, RNA-seq has shown diagnostic utility in specific tissues and diseases⁶⁻⁸. This includes muscle biopsies from patients with undiagnosed rare muscle disorders^{6,9}, and cultured fibroblasts from patients with mitochondrial disorders⁷. However, in many cases, biopsies are not performed for clinical care, and tissues are difficult to access. We sought to assess the utility of RNA-seq from blood as a diagnostic tool for rare diseases of different pathophysiologies. We generated whole blood RNA-seq from 94 cases with undiagnosed rare diseases spanning 16 diverse disease categories. We developed a robust approach to compare cases to large sets of RNA-seq controls (N=1,594 external and N=49 family-based controls) and demonstrated the impacts of expression, splicing, gene and variant filtering strategies on disease gene identification. Across our cohort, we observed that RNA-seq yields a 7.5% diagnostic rate and an extra 16.7% with improved candidate gene resolution.

We obtained RNA sequencing data from samples from 143 individuals, 94 affected by rare diseases and 49 unaffected family members (Supplementary Table 1) and whole exome or

genome sequencing for 112 of them. In total, whole exome sequencing did not identify the causal variant in 88.8% of patients. Patients represented 80 different diseases and were broadly classified into 16 distinct disease categories, with neurology, musculoskeletal, hematology and ophthalmology as the most frequent (Fig. 1a, Supplementary Table 2). We integrated these data with RNA-seq data from healthy individuals from the Depression Genes and Network (DGN) cohort (N=909)¹⁰, the Prospective Investigation of the Vasculature in Uppsala Seniors (PIVUS) project (N=65)¹¹ and the Genotype-Tissue Expression consortium (GTEx version 7) (N=620) cohorts¹²(Supplementary Table 3). By comparison to large healthy cohorts, we demonstrate how extreme gene expression and splicing events can aid in identifying candidate genes and variants.

We first evaluated the extent that whole blood RNA-seq captured gene expression of known rare disease genes in each major disease category (Fig. 1b, Supplementary Table 4). When broadly considering disease genes from the Online Mendelian Inheritance in Man (OMIM) database¹³, we observed 70.6% were expressed in blood and 50% of corresponding gene splicing junctions were covered with at least 5 reads in 20% of samples (Extended Data Fig. 1A–B). Notably, for a panel of genes known to be involved in neurological disorders (N=284), we observed that 76% were expressed. Using scores from ExAC¹⁴, we further observed that genes expressed across multiple tissues were more depleted in missense or loss-of-function (LoF) mutations¹⁴(two-sided Wilcoxon test, p-value 2×10^{-16} Extended Data Fig. 1C, Supplementary Table 3). This suggests that mutations that have more severe consequences occur more often in genes for which expression is not restricted to one tissue. Indeed, we observed that 66% of LoF-intolerant genes (probability of being intolerant to LoF mutations (pLI = 0.9) are expressed in blood samples (average transcripts per million, TPM = 1)(Extended Data Fig. 1D).

Outlier (or aberrant) expression of a gene in a sample when comparing to all tested samples has previously been shown to help identify large-effect rare variants and rare disease genes in blood^{15–17}. We assessed the differences between outlier genes in cases versus controls after correcting the data for batch effects (see Methods, Extended Data Fig. 2, Extended Data Fig. 3). We observed an enrichment of case under-expression outliers in genes more sensitive to LoF mutations (Fig. 2a, red). This observation corroborates previous evidence that new LoF mutations are more likely to lower expression level through nonsense mediated decay (NMD)^{18–20}. As we increased the number of controls, the enrichment became stronger, demonstrating the impact of large control datasets (Extended Data Fig. 4).

We observed an average of 343 outliers per sample ($|Z\text{-score}| \geq 2$, Fig. 2b). We tested different variant and gene-level filters that could aid in further narrowing down the lists of candidate genes (Fig. 2c). We filtered for genes that were LoF intolerant (Filter 1; pLI = 0.9), likely to have a regulatory variant impacting gene expression (Filter 2: RIVER score 0.85^{21}); showed allele-specific expression (ASE) (Filter 3); linked to the phenotype (Filter 4: Human Phenotype Ontology²² (HPO) match), with a rare variant with MAF = 0.01% within 10kb upstream of the gene (Filter 5); and with a rare variant that was likely deleterious (Filter 7; CADD score ≥ 10). Other filters were combinations of these sets. We observed that when restricting to under-expression outlier genes with HPO matches and a deleterious rare variant nearby, we were able to reduce the candidate genes list to less than 1% of the initial

set of outliers with 80% of cases having at least one candidate gene (Filter 11; Fig. 2c; Extended Data Fig. 5A). Overall, we are able to reduce the number of expression outliers to less than 10 on average for all cases (Fig. 2d).

Outlier splicing is also an important contributor to Mendelian disease^{6,7,23–26}. To evaluate splicing events across rare disease samples, we corrected junction data for batch effects (Extended Data Fig. 6) and obtained Z-scores in all samples (Fig. 3a, see Methods). On average, we detected 540 splicing outlier genes for each sample at $|Z\text{-score}| \geq 2$ (Fig. 3b). We observed that the number of splicing outliers was influenced by the number of junction in each gene, was higher in cases and, like expression outliers, was enriched in genes sensitive to LoF mutation (data not shown). From both exome and genome data alone, we observed that the number of rare variants impacting splicing was large but could be significantly reduced when combined with outlier splicing information from RNA-seq (Fig. 3c). From our pool of candidate genes with splicing outliers, we looked at the proportion remaining after different filters (Fig. 3d). We observed that limiting to genes relevant to the phenotype (Filter 2) and with a deleterious rare variant within 20 bp of the splicing junction (Filter 5), we were able to narrow down to only 0.14% of potential candidate genes (Filter 7). Overall, 32% of cases had at least one gene matching these criteria (Extended Data Fig. 5B). Furthermore, genes selected after filtering carried more deleterious rare variants than unfiltered outliers suggesting an enrichment of disease-genes with compound heterozygous mutations (Fig. 3e). Overall, the filtering steps are reducing candidates to less than 10 per case on average (Fig. 3f).

RNA-seq provides the ability to measure allele-specific expression. ASE can inform the presence of a large-effect heterozygous regulatory, splicing or nonsense variant, or epimutation aiding the identification of candidate rare disease genes and variants^{7,27–29}. Out of all possible heterozygous sites ($\sim 10^5$ to 10^6 per sample for exome and genome, respectively), 10^4 variants had sufficient coverage for analysis (Extended Data Fig. 7A). Independent of sequencing technology, we observed 10^3 sites displaying allelic imbalance with an allelic ratio < 0.35 or > 0.65 . To highlight ASE events that might be disease-related, we focused on the subset of genes outlier ASE sites within case individuals when compared to all other rare disease individuals and GTEx samples (Extended Data Fig. 7A). We found an average of 94 ASE outliers per individual. We observed that the top 20 ASE outlier genes are enriched for overlap with HPO-associated genes per case, regardless of the filters applied to the extreme ASE genes and background genes (ie pLI > 0.9 , Rare variant nearby, Rare variant with CADD score > 10 nearby; (Extended Data Fig. 7B). We also tested whether ASE would allow us to identify deleterious variants that were over-represented as this may be a marker for compound events or haploinsufficiency. Here, we focused on rare deleterious variants where the alternative allele is more abundant than the reference allele (Extended Data Fig. 7C). In total, 111 rare variants show allelic imbalance biased toward the deleterious alternative allele (96 splice and 15 stop-gain). Among those, one variant is in *EFHD2*, a gene coding for Ca^{2+} adapter protein involved in B-cell apoptosis, NF-kB mediated in inflammatory response, and immune cell activation and motility^{30–33}. The carrier of this event was diagnosed with idiopathic cardiomyopathy, where accompanying symptoms (elevated inflammatory markers, Raynaud's disease, and alopecia) are indicative of auto-immune issues.

By integrating expression, splicing and ASE signals, we were able to identify and validate the causal gene in 6/80 independent cases (7.5%, 4 expression outliers, 2 splicing outliers), identify candidate genes potentially linked to the disease phenotype (gene matching HPO terms for the symptoms of the proband) in 5/30 cases with candidate gene information (16.7%) (Extended Data Fig. 8A, Supplementary Table 1). We did not find highly relevant candidate genes for 69 cases (86%). Notably, candidates were identified for 5 neurological cases where blood is not assumed to be a representative tissue. Furthermore, we observed that for cases where a candidate gene set was provided based on previous literature, we had a higher percentage of overlaps with an RNA-based filtered gene set than a DNA-based filtered gene set (Extended Data Fig. 8B).

Three cases exemplify the use of RNA-seq in causal gene identification. In a first case, two brothers ages 4 and 5 years, each presented at 6 months with delayed motor milestones and hypotonia, which evolved to include spasticity, an ataxic gait, and progressive loss of motor skills. Genome sequencing identified biallelic heterozygous pathogenic variants in the *MECR* gene present in both siblings: c.830+2dupT and c.-39G>C. Pathogenic variants in *MECR* are associated with mitochondrial enoyl CoA reductase protein-associated neurodegeneration (MEPAN), a rare disorder characterized by childhood-onset movement disorder, signal hyperintensity in the basal ganglia, optic atrophy, and relatively preserved cognition. To date only seven individuals with MEPAN have been reported in the literature³⁴. While the c.830+2dupT variant has been described previously, the c.-39G>C variant is novel. After applying our pipeline on expression outliers, we found *MECR* as a candidate in both siblings in a list of 11 and 15 genes respectively (with respectively 1 and 0 candidates left after the splicing pipeline, True Positive Rate (TPR) between 6.7 and 8.3%) (Extended Data Fig. 9A–B). Without expression information, there was 245 and 302 genes (including *MECR*) and (with 111 and 161 additional after the splicing based pipeline) linked to the phenotype with a rare deleterious variant within 10kb (TPR between 0.28 and 0.21%).

In a second case, a 12-year-old Hispanic female presented with developmental regression after typical development until age 18 months, manifesting with loss of milestones including head control, and speech. Tremors developed at 21 months; and seizures at 22 months. She also suffered from occasional myoclonus. She has a 5-year-old brother with onset at 13 months of ataxia, autism, developmental delay, recurrent febrile seizures, and absent speech. Without expression data, we were able to filter the number of candidate genes from 1,034 genes to 96 genes (with an additional 105 candidate genes from the expression based side of the pipeline, TPR 0.49%), when looking only at genes associated with the phenotypes (from HPO terms [24]) and containing rare variants within 20bp of annotated junctions with a CADD score $> 10^{35}$. The causal gene is missing from the most stringent filter because there is no rare deleterious variants within 20bp of known junctions. Adding splicing outlier information from RNA-seq data left us with one gene (in addition to 7 filtered expression outliers, TPR 12.5%); *KCTD7*, containing a non-annotated junction in the affected sample (Fig. 4a, left panel). A synonymous mutation was found responsible for the creation of a new splicing junction in this gene (p.V152V, Fig. 4b). RT-PCR from RNA extracted from fibroblasts from exons 2–4 regions of the gene confirmed a difference in fragment size in the probands (Fig. 4c, Source Data). In addition, this variant exhibited monoallelic expression towards the reference allele as a consequence of the premature splicing event.

In a third case, we reprocessed a solved case for which we had found an exon-skipping event in a previous study⁸. In this case, the patient presented with a sporadic form of spinal muscular atrophy. After filtering for splicing outliers ($|Z\text{-score}| \geq 2$) and selecting only genes relevant to the symptoms (HPO), only one gene was left (Extended Data Fig. 10A), *ASAH1*, for which we subsequently identified with Sanger sequencing a splice-loss mutation leading to the creation of a new transcript, skipping exon 6 (Extended Data Fig. 10B). While many of our cases had genetic data, this case demonstrated that use of RNA-seq alone can aid in disease gene identification.

In summary, the use of whole blood RNA-seq in combination with variants and phenotype relevant gene filters was able to identify the causal gene and variant(s) in 7.5% of cases or to further highlight candidate genes linked to phenotype in 16.7% of cases (see methods). We recommend using our most stringent set of filters from splicing and expression outliers. As with exome sequencing, we expect this to be a baseline rate that will grow through ongoing case reanalysis^{36,37}. Similar to the utility of large databases of control exomes for Mendelian disease diagnoses^{14,38–40}, we demonstrated the utility of large control RNA-seq data to identify aberrant expression, splicing and ASE events in candidate rare disease genes. Furthermore, this work demonstrates the utility of performing RNA-Seq on peripheral blood, which is a readily available specimen type in clinical practice. Throughout our study, a trade-off needed to be found between strictly filtering the data and losing candidates of interest. It is worth noting that this combination of information is not expected to lead to the causal gene successfully in the following situations: first, if the causal gene is not expressed in the analyzed tissue; second, if the effects of the causal variant do not affect the expression of the gene; and third, if the filters are too strict. Therefore, expert evaluation remains required when prioritizing candidate genes using RNA-seq data. We can expect that combining information from multiple “omics” sources will only further improve diagnosis of unsolved rare disease cases in the future.

Material and Methods

We sequenced 143 whole blood samples, 94 extracted from affected individuals and 49 unaffected family members. The 94 cases represent a total of 80 independent diseases. Samples were collected from 3 different institutions, the Children’s Hospital of Eastern Ontario (CHEO), the Stanford Clinical Genomics Program (CGP) and the Undiagnosed Disease Network (UDN). Ethical and research approval was obtained by CHEO Research Ethics Board (REB Protocol Number 11/04E), NHGRI IRB (Protocol 15-HG-0130) and Stanford University IRB (Protocols 23066, 32641 and 38046).

Whole blood samples were collected and shipped in Paxgene RNA tubes or as isolated RNA for processing. Paxgene RNA tubes were processed manually per manufacturer’s protocol and 1.0 μg RNA was used for further processing. Isolated total RNA was analyzed on an Agilent Bioanalyzer 2100 by pico RNA chip for RIN quality. Globin mRNA was removed using GLOBINclear prior to cDNA library construction. cDNA libraries were constructed following the Illumina TrueSeq Stranded mRNA Sample Prep Kit protocol and dual indexed. The average size and quality of each cDNA library was determined by Bioanalyzer and concentrations were determined by Qubit for proper dilutions and balancing across

samples. On average, twenty pooled samples were run simultaneously on an Illumina NextSeq 500 (high output cartridge). Pooled samples were run in 9 distinct sequencing runs: two runs generated 75bp paired end reads and seven runs generated 150 bp paired end reads. Output bcl files were converted to fastq files and demultiplexed using bcl2fastq version 2.15.0.4 from Illumina. Overall, we obtained around 50M reads per sample (median 52M +/- 20M).

Reads were trimmed and adapters were removed using cutadapt v1.11 (<https://github.com/marcelm/cutadapt>). Reads were then aligned to the reference human genome (hg19) with STAR v2.4.0j (https://github.com/alexdobin/STAR/releases/tag/STAR_2.4.0j). We used gencode v19 for reference annotation (<https://www.gencodegenes.org/releases/19.html>). We removed reads with a mapping quality under 30 and filtered duplicate reads with Picard Tools MarkDuplicates v1.131 (<http://broadinstitute.github.io/picard/>). Gene-level and transcript-level quantifications were generated with RSEM v1.2.21⁴¹ (<https://github.com/deweylab/RSEM/releases/tag/v1.2.21>). Junctions files generated by STAR were filtered: to consider a junction, a minimum of 10 reads uniquely spanning was required. For faster processing of samples, we used GNU parallel⁴².

Independent control cohorts for expression, splicing and ASE analyses

We used whole blood transcriptome data of 909 samples from the DGN cohort¹⁰ as well as 65 samples (age 70) from the PIVUS cohort¹¹ to serve as independent healthy controls for expression analysis and splicing respectively. DGN samples are single-end 50bp reads and PIVUS samples are 75bp paired-end reads. Sequences were aligned, quantified and filtered following the same protocol used for rare disease cases and controls. We determined outlier ASE events at the gene level per individual by comparing our data to 620 individuals in GTEx v7¹² across 48 tissues. Allele-specific expression in GTEx was processed as in²¹, and only sites with a minimum of 20 reads overlapping and not entirely monoallelically expressed were analyzed.

We tested the tolerance to different types of mutations (from ExAC) in function of the expression status in a single versus multiple tissues using a two-sided Wilcoxon rank sum test on 620 individuals from GTEx v7 across 22 tissues.

Disease gene lists

Disease gene lists for neurology (n=284 genes), ophthalmology (n=380 genes), hematology (n=50 genes) and musculoskeletal and orthopedics (n=395 genes) and disease categories were obtained from curators for genes of interest in regards of the disease (Table S4). We obtained OMIM genes list (n=3,766 genes) from <https://omim.org/downloads/>. Gene expression of disease genes in our samples was restricted to protein coding genes.

Genetic data

Variant data was produced according to recommended protocols for exome or genome data. VCFs obtained from UDN were generated through Hudson Alpha and Baylor pipelines. In short, DNA reads alignment was performed using BWA-mem v0.7.12⁴³ and variant calling was made using GATK v3.3⁴⁴. For CGS samples, variant calling was performed using

GATK v3.4. We filtered variants according to the following criteria from previous studies^{14,45}:

- Filter field is PASS
- At least 20 reads covering the position (DP field)
- Genotype quality greater than 20 (GQ field)
- Normalized Phred-scaled likelihoods of the predicted genotypes lower than 20 (PL field)\
- $\frac{Allele\ depth}{Total\ depth} > 0.8$ for homozygous calls and > 0.2 for each allele for heterozygous calls.
- Exclude variants with Hardy-Weinberg Equilibrium p-value $< 1 \times 10^{-6}$
- Exclude variants with call rate < 0.80 (missing $> 20\%$)

We obtained genetic information for 112 samples (out of 143, 54 from whole genome sequencing, 58 from whole exome sequencing) (Supplementary Table 1). The number of LoF rare variants is variable across samples, and institutions. We merged all VCF files from those different institutions and homogenized their format for further analysis. We used BEDtools (version 2.26.0–112-gd8c0fe4)⁴⁶ to filter for junction or gene with a rare variant within 20 bp of a tested splicing junction. We filtered for rare variants with minor allele frequency 0.1%. We kept the singletons in the analysis.

Genetic data annotation

We annotated genetic data with allele frequency from the Genome Aggregation Database (gnomAD)¹⁴ and Combined Annotation Dependent Depletion scores (CADD)³⁵ scores using Vcfanno (version 0.2.7)⁴⁷. We used CADD scores v1.3 and gnomAD genomes release 2.0.2.

Ancestry inference

VCF files were processed for ancestry inference using BCFtools v1.8 as following. They were normalized (fixing strand flips and left aligning indel records) and merged. We then subset this file to only variants in exonic regions, and filtered out variant with $> 25\%$ missingness. Missing variants were set to homozygous reference. A total of 2,666 variants remained after filtering. To perform ancestry inference, we used all individuals from 1000 Genomes phase 3 version 5 populations. For computational feasibility, we used genotypes from chromosomes 1,4,12,15,16, and 19. We used the prcomp function in R to extract principal components and plotted the first three principal components.

Expression levels normalization

We filtered out genes for which less than 50% of samples from each origin (i.e. rare disease individuals and unaffected family members sequenced in-house, external controls) had TPM > 0.5 and/or variance equal to zero. This resulted in 14,988 genes being retained in the dataset. We performed Surrogate Variable Analysis (SVA) using the “two-step” method on a centered and scaled matrix of log₁₀-transformed (log₁₀(TPM+1)) RNA-seq count data

output by RSEM⁴¹. We did not provide any known covariates to SVA. To control for nonlinearity in uncorrected gene expression data, we added regression splines for the top two Surrogate Variables (SVs) significantly associated with batch and institution (p-value $< 1e^{-30}$ from univariate linear regression of batch and institution against all significant SVs), removing the untransformed SV in each case. Linear regression splines had knots positioned at every 1.66% of samples, resulting in approximately 17 individuals per region - which is around the average number of individuals in each batch sequenced in-house (Extended Data Fig. 3). Significant surrogate variables (SVs) and regression splines were then used as covariates in a regression model. The residuals of this model were centered and scaled to generate Z-scores for use in all subsequent analyses using gene expression data.

We tested the impact of adding splines in the model using a per-gene likelihood ratio test comparing linear regression model fit with and without regression splines. We used 1052 samples and corrected p-values for multiple testing (Benjamini & Hochberg adjustment).

Global outliers

To control for potential residual technical artifacts impacting outlier expression, we removed samples for which 100 or more genes had normalized expression values of $|Z\text{-score}| > 4$ after SVA correction (54 samples). We tested the model described in Figure 2 for several global outlier thresholds and observed a similar enrichment profile.

Gene expression outlier enrichment analyses

We used the union of DGN samples and healthy family members that passed the global outliers criteria as the control set (n=899 and n=32 individuals, respectively). We assessed enrichment for case outliers at increasingly stringent percentiles of gene expression in genes intolerant to mutations using a logistic regression model. As features in this model we used ExAC gene constraint metrics for LoF, missense and synonymous mutations [16]. For each gene in the dataset that had ExAC gene constraint metrics (n genes = 10,605), we calculated a binomial outcome variable corresponding to the proportion of case expression outliers found in each gene: $Y_i \sim B(n_i; p_i)$, where n_i is the number of outlier samples in gene i at a given percentile tested (the number of 'trials'), and p_i is the proportion of case outlier samples (which can be thought of as the probability of 'success' (or all outliers being case outliers) for gene i). Then we modeled the relationship between the observed proportion of cases for each gene, and the corresponding gene constraint Z-score from ExAC. Specifically, we wanted to find $\Pr(Y_i = \text{AllCases} | X_i)$, where X_i is the gene constraint Z-score for gene i . We assessed the effect of X using logistic regression: $\text{logit}(p(X)) = \beta_0 + \beta_1 X$. A positive β_1 value indicates that a step change in constraint metric X (toward genes less tolerant to mutations) is associated with an increase in the log odds of $Y_i = 1$ (i.e. all outliers being case outliers). A separate model was fit for each mutation class. We reported results as the log odds (+/- 1.96*SE) associated with each feature for each percentile. P-values were calculated based on the z-statistic.

RIVER

RIVER (RNA-informed variant effect on regulation) is a hierarchical Bayesian model to infer rare variants of their regulatory effects. Compared with other variant scoring methods,

RIVER takes the advantage of utilizing both genomic information and transcriptome information²¹. We used GTEx v7 whole genome sequencing and cross-tissue RNA-seq data as training data for the model. The trained model (with learned parameters) is subsequently applied on UDN data to predict effects of rare variants. The model uses rare variants and the genomic annotations at those variants as predictors, uses RNA status (for this case is outlier status based on total gene expression levels) as the target/response variable. Rare variants here are defined as those with a minor allele frequency (MAF) <0.01 in 1000 Genomes Project phase III all populations combined. For variants in GTEx, we additionally require MAF <0.01 within the GTEx cohort itself and for variants in the rare disease samples we additionally require MAF <0.02 within the rare disease samples themselves. We considered all rare variants 10kb near genes (10kb before transcription start site until 10kb after transcription end site). Overall, there were a median of 2 rare variants per gene for GTEx subjects and rare disease subjects. For this analysis, we considered protein-coding and lincRNA genes only. We used the following genomic annotations: Ensembl VEP⁴⁸, CADD³⁵, DANN⁴⁹, conservation score (Gerp⁵⁰, PhyloP⁵¹, PhastCons⁵²), CpG content, GC content, chromHMM⁵³ and Encode chromatin-openness track. We selected those features based on their prior evidence of association with regulatory effects²¹. Features were aggregated over each gene and individual pair, using either max(), min() for quantitative features, or any() for categorical features. Expression outliers (the response variable) were defined as those with $|Z\text{-score}|>2$. Z-scores were calculated based on total gene expression level RPKM from RNA-seq. In addition, for GTEx training data, gene expression levels were corrected by PEER⁵⁴ to remove technical artifacts and major common-variant eQTL effects are also removed. Z-scores for GTEx are median over all available tissues²¹.

Junctions coverage ratios

Reference junctions were derived from Gencode v19 annotation file on known protein coding genes (142,246 in 14,296 genes). For each junction donor (then acceptor), all possible acceptors (then donors) were screened in the samples junctions files. The distribution of reads spanning those junction sets was evaluated by calculating the set ratios (Fig. 3a). We restricted the analysis to junctions for which several acceptors (donors) were associated to one donor (acceptor). In total, 13,109 groups of junctions were generated. In total, 34,060 junctions in 6,261 protein coding genes across all samples fulfilled those criteria. We performed the analysis on all PIVUS samples (n=65) and rare disease samples (n=143).

Splicing data normalization and analysis

We used the union of PIVUS samples (n=65) and healthy family members (n=49) as a control set together with all the cases samples (n=94). To remove possible noise, and to allow missing values imputation, we removed junctions for which there was no more than 30 samples with data in the junction group. We analyzed coverage ratios for a total of 25,612 junctions.

Missing values in junction coverage ratios were imputed using missMDA R package. PCA analysis was then performed using prcomp R package. We regressed out principal components (PCs) accounting for 95% of the variation in our imputed dataset (176 PCs). We

then put back original missing values in the dataset and derived Z-scores used in the outlier analysis. We looked at the correlation pattern between the 10 first PCs and known covariates from our dataset. In brief, PC1 was mainly separating the source of the data (UDN, CGS, CHEO or PIVUS). PC2 was highlighting differences between the first batch and the other batches. Overall, we observed some level of correlation between all known covariates and the PCs that were regressed out from the data.

We tested the impact of our filters on the median number of rare variants with CADD score 10 with a two-sided Wilcoxon rank sum test on all samples with splicing outliers and genetic information (n=74).

Allele specific expression

ASEReadCounter⁵⁵ version 3.8-0-ge9d806836 from Genome Analysis Tool Kit (GATK)⁵⁶ was run on single nucleotide variants from VCFs provided by the UDN, CHEO and CGS and corresponding RNA-seq data, using all samples for which we had genetic information (n=112). Only sites with a minimum read depth of 10, mapping quality of 10 and base quality of 2 were integrated in the analysis. For a gene to be considered with ASE, we required that at least 5 samples had heterozygous sites in the gene, that the heterozygous site was covered by at least 20 reads for the individual and to be considered imbalanced for allelic ratios 0.65 or 0.35. We eliminated total mono-allelic expression from the analysis (ie allelic ratio = 0 or 1).

To detect ASE outliers we restricted our analysis to sites and genes common to our samples and GTEx dataset including 11,224 genes and 87,739 sites, subject to the same site filters above. After this step, 108 individuals were left. We scaled the reference ratios for all sites within a gene across samples to obtain Z-scores per site. To summarize GTEx data per individual, we considered the maximum ratio ($|0.5 - \text{reference ratio}|$) across all tissues for which the individual had data at that site. Then, for each case, we selected the top N genes by $|Z\text{-score}|$ as ASE outliers. We assessed the overlap of this set of genes with the genes associated with that individual case's listed HPO terms, as well as the parent and child terms. To determine whether the overlap of ASE outliers with HPO-associated genes is significant, we selected 20 genes at random for each individual and assessed the overlap with the same HPO-associated genes. This was repeated 100 times. We then layered in additional filters, and took the top N most extreme ASE genes with a pLI > 0.9 or with a nearby rare variant with that individual, and finally, a nearby deleterious (CADD 10) rare variant. In each instance, we matched the background for that filter, thereby comparing the overlap for extreme ASE + pLI > 0.9 with HPO-associated genes to the overlap seen in a random background of genes, also with pLI > 0.9, and the same with the rare variant criteria.

Phenotypic data

For each case we have RNA-seq data (n=94), we also obtained HPO terms corresponding to the symptoms of the affected individual. We extended this list of HPO terms to terms that were hierarchically one level lower (child terms), one level higher (parent terms) or alternative terms for the same phenotype. To do so, we used the Human Phenotypic Ontology (HPO, downloaded 10-23-18)²² (<http://human-phenotype-ontology.github.io/>

[downloads.html](#)). To link HPO terms to genes, we used the genes to phenotype and phenotype annotation files provided by the Human Phenotypic Ontology.

Diagnostic rate

We labeled “solved” cases for which we found candidates from RNA-seq data for which the causal mutation was found and validated. In order to evaluate the number of cases for which we have strong candidates, we took a subset of 30 cases from the same institutions for which we obtained a list of candidate genes from curators. If any of those candidates were in the final set of filtered genes, they were labelled as “strong candidate”. Cases for which no strong candidate genes were found after analyzing RNA-seq data are labeled “no candidate”.

You can find detailed information on protocol in the Life Science Reporting Summary.

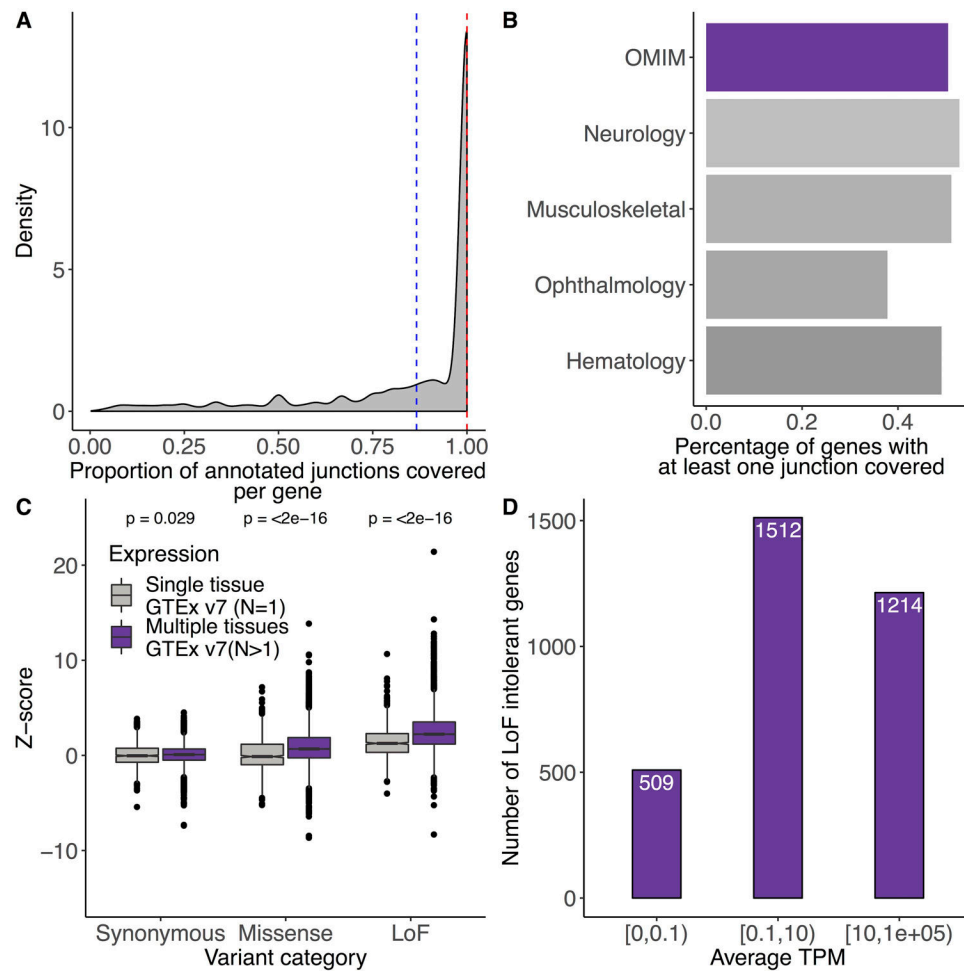
Code availability

Code for running analysis and producing the figures throughout the manuscript is available at https://github.com/lfresard/blood_rnaseq_rare_disease_paper. Our pipeline to highlight candidate variants is available at https://github.com/lfresard/blood_rnaseq_rare_disease_paper/blob/master/pipeline.md

Data availability

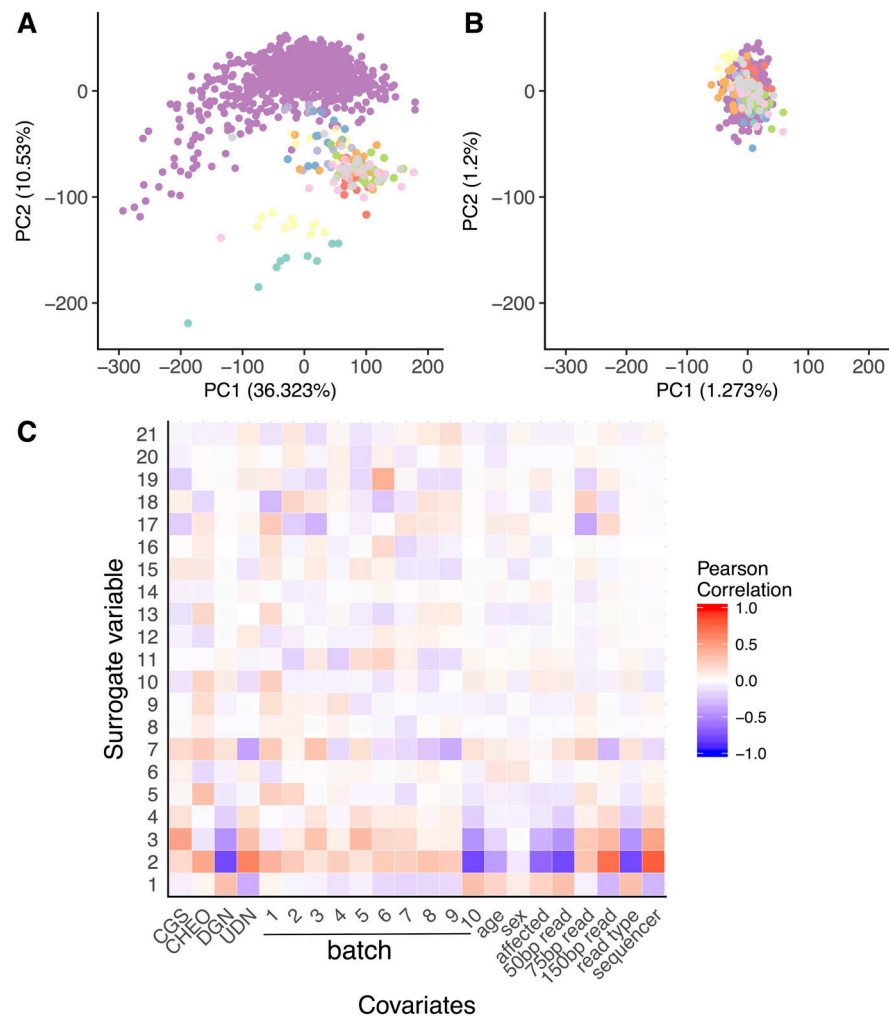
UDN data is accessible through the UDN Gateway and through dbGaP entry at phs001232.v1.p1. DGN RNA-seq data is available by application through the NIMH Center for Collaborative Genomic Studies on Mental Disorders. Instructions for requesting access to data can be found at https://www.nimhgenetics.org/access_data_biomaterial.php, and inquiries should reference the “Depression Genes and Networks study (D. Levinson, PI).” The GTEx Analysis V7 release allele-specific expression data is available from dbGaP (dbGaP Accession phs000424.v7.p2). PIVUS RNA-seq data is accessible on the European Genome-Phenome Archive (EGAS00001003583). The Care4Rare data is available through Genomics4RD.

Extended Data



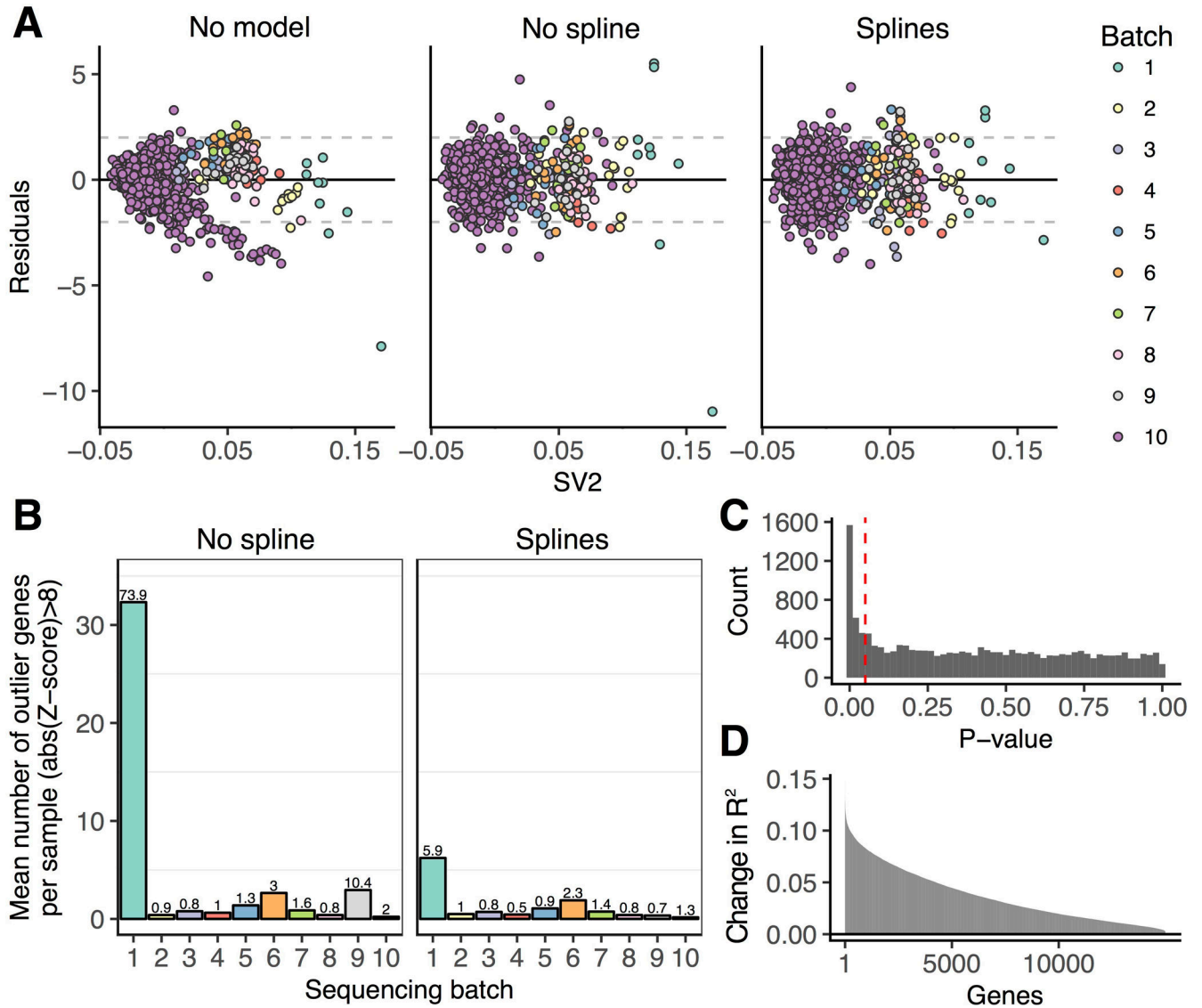
Extended Data Fig. 1. Gene expression patterns across whole blood samples.

We used a total of 1,061 whole blood samples from our controls cohorts and rare disease samples. **a**, Density plot representing the proportion of annotated junctions covered per gene. Those are a subset of genes for which at least one junction is covered with at least five uniquely mapped reads across at least 20% of the samples. On average (blue dashed line) 86%, (median of 100%—red dashed line) of junctions fulfil those criteria. **b**, Percentage of genes from disease genes panels in which at least one junction is covered with at least five uniquely mapped reads in at least 20% of samples. We observe that about 50% of genes from OMIM, Neurology, Musculoskeletal, Ophthalmology or Hematology panels are fulfilling this criteria. **c**, Tolerance to different types of mutations (from ExAC) in function of the expression status in a single versus multiple tissues (two-sided Wilcoxon test, P value 2×10^{-16}). Analysis performed on 620 individuals from GTEx v.7 across 22 tissues. Boxplots represent median value, with lower and upper hinges corresponding to the 25th and 75th percentiles, and lower and upper whiskers extend from the hinge to the smallest and largest value at most $1.5 \times$ interquartile range of the hinge, respectively. Genes that are expressed in multiple tissues tend to be more sensitive to missense and LoF mutations. **d**, Number of LoF intolerant genes stratified by expression level in blood. We considered genes with pLI score ≥ 0.9 as LoF intolerant.



Extended Data Fig. 2. Correction for batch effects: Expression data.

Analyses performed on $n = 909$ DGN samples and 143 rare diseases (cases and family controls). **a**, Plot of first two principal components run on uncorrected gene expression data. Samples are coloured by batch. Largest cluster (green dots) are DGN control samples ($n = 909$). **b**, Plot of first two principal components run on gene expression data after regressing out significant surrogate variables found by SVA. **c**, Correlation between known covariates and all significant surrogate variables (SVs). We observed that SV2 is highly correlated with the read type, and the sequencing technology corresponding to differences between DGN and the other samples.



Extended Data Fig. 3. Use of regression splines in expression data normalization.

a, Normalized gene expression residuals from 1,052 samples in an example gene without correction (left panel), after regressing out significant surrogate variables (SVs) (middle panel) and significant SVs plus regression splines on top SVs significantly associated with batch and study (right panel). Residuals were plotted against SV2 for illustration purposes (SV2 is significantly associated with batch (P value $< 1 \times 10^{-30}$, two-sided t -test from linear regression, no adjustment for multiple correction). **b**, Mean number of outlier genes per sample ($n = 990$) in each batch (absolute Z -score > 8) after correction with SVs (left panel) and SVs with regression splines (right panel). Standard deviation is displayed above each bar. Regression splines resulted in a more consistent number of outlier genes across samples in all batches. **c**, Benjamini & Hochberg adjusted P values resulting from a per-gene likelihood ratio test comparing linear regression model fit both with and without regression splines. Regression splines improve the model fit for 2,644 genes (P value 0.05, 17.6% of all genes in dataset). Red dashed line indicates P value = 0.05 cutoff. **d**, Change in R^2 , in

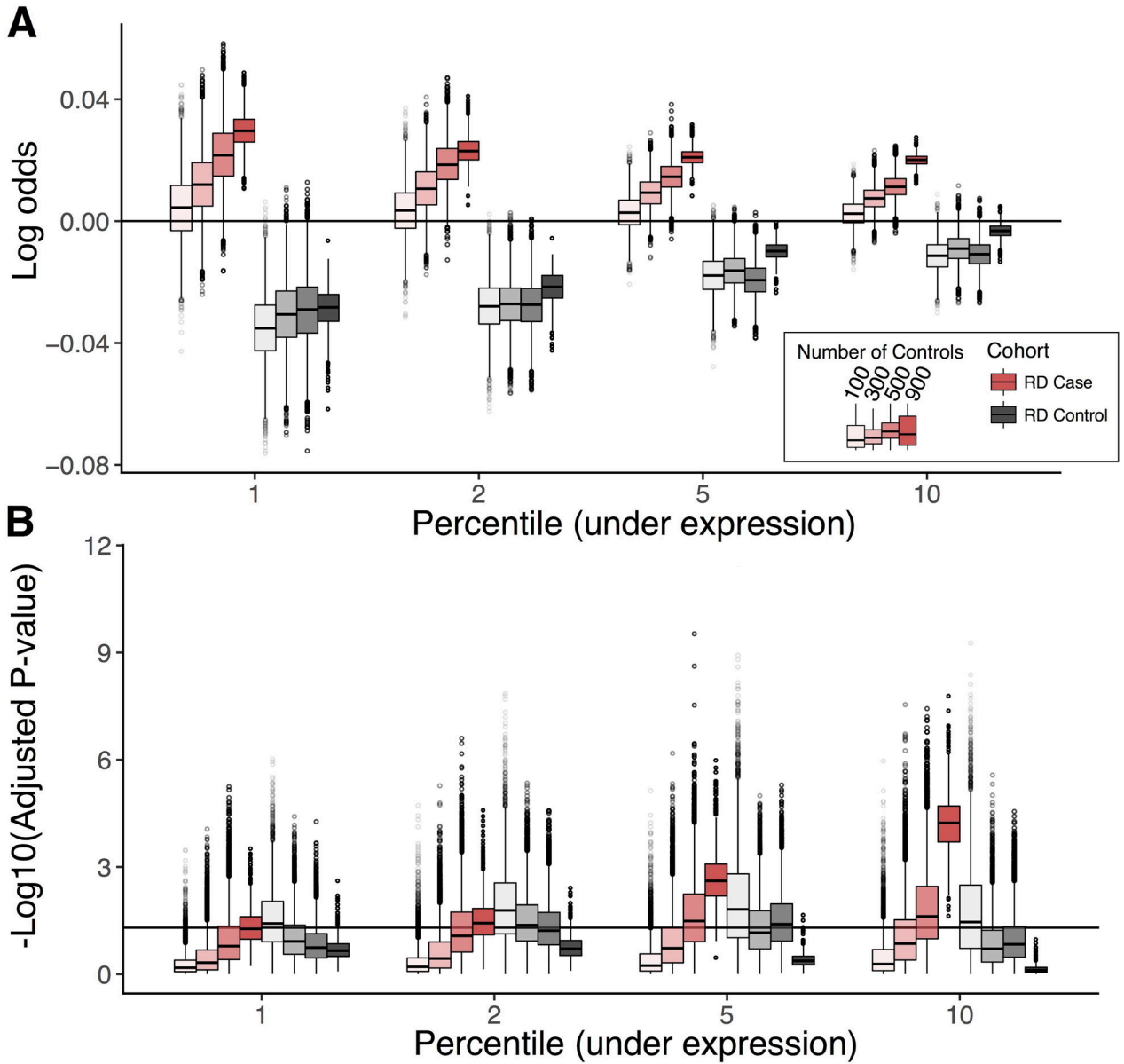
decreasing order, across all genes in the dataset ($n = 14,988$) after correcting data using significant SVs with regression splines, compared to correcting data using significant SVs without regression splines. Mean change in R^2 is 0.036 (s.d. = 0.025).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

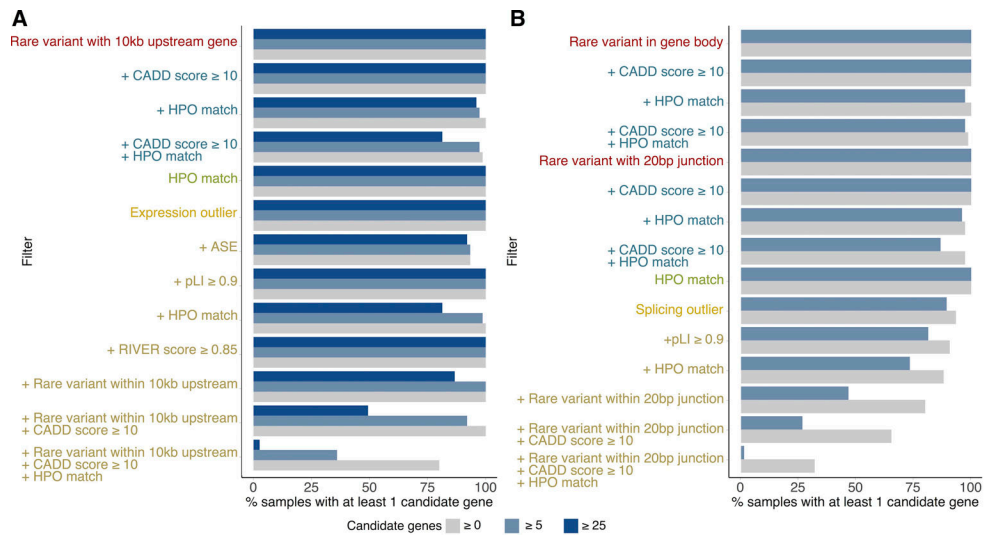


Extended Data Fig. 4. Impact of the number of controls on loss-of-function intolerance enrichment.

a, Enrichment of case (red, $n = 64$) under-expression outliers in LoF sensitive genes as we increase the number of controls (7,600 random subsets for each sample size indicated in legend). This enrichment was not observed for rare disease family member controls (gray, $n = 34$). **b**, Benjamini & Hochberg adjusted $-\log_{10} P$ value associated with the enrichment at different number of controls (two-sided t -test, $n = 64$ cases). Horizontal line indicates 0.05 significance cutoff. The P values are decreasing as we increase the number of controls.

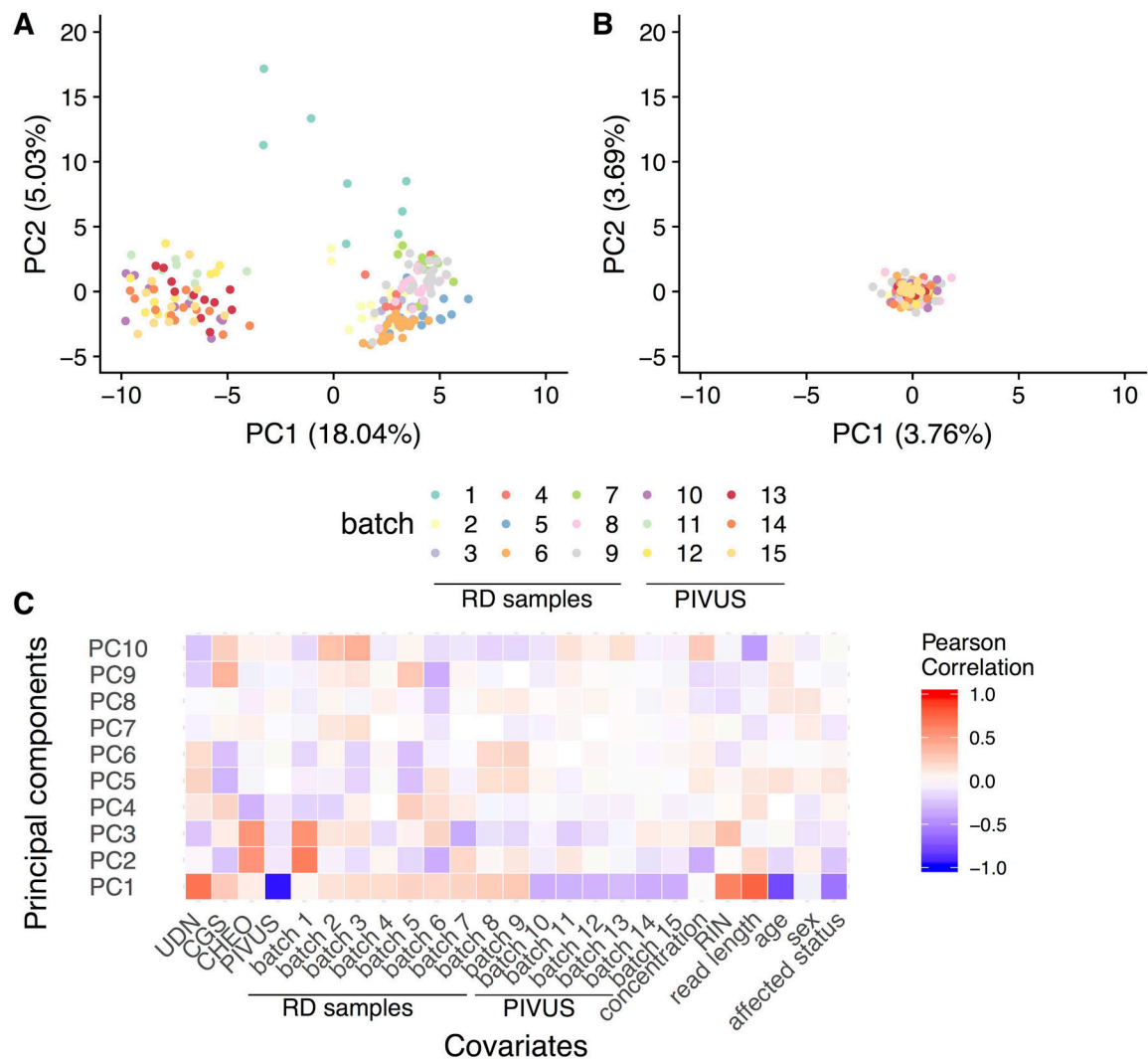
When switching cases for controls (gray) we observed significant negative log odds when using the a smaller number of controls, but this trend disappeared when using the full set of 900 controls. For **a** and **b**, Boxplots represent median value, with lower and upper hinges corresponding to the 25th and 75th percentiles, and lower and upper whiskers extend from

the hinge to the smallest and largest value at most $1.5\times$ interquartile range of the hinge, respectively.



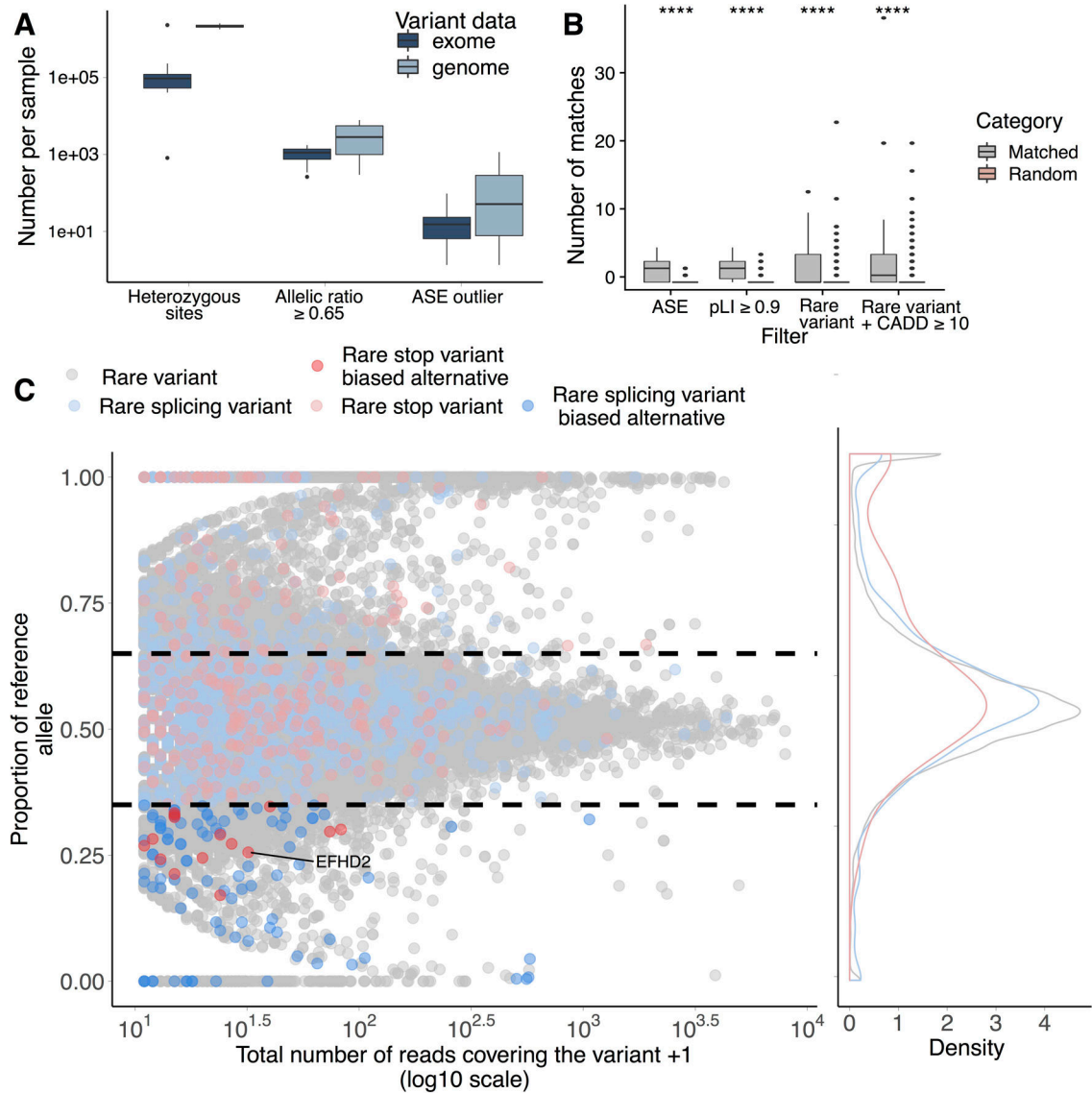
Extended Data Fig. 5. Percentage of samples left when filtering outliers.

Filters have various impacts on the number of samples with at least one candidate gene. By combining several layers of filters we are drastically reducing the number of candidate genes but also the number of samples for which we have candidates. We recommend to relax filter stringency after looking at sets of genes that match the most stringent criterion. **a**, Expression outliers. After filtering for outlier genes matching HPO terms, with a deleterious rare variant within 10 kb, we observed less than 2.6% of samples with over 25 candidate genes. **b**, Splicing outliers. When keeping only genes with HPO match, and a deleterious rare variant with 20 bp of the outlier junction, we observed less than 1.3% of samples with more than five candidate genes.



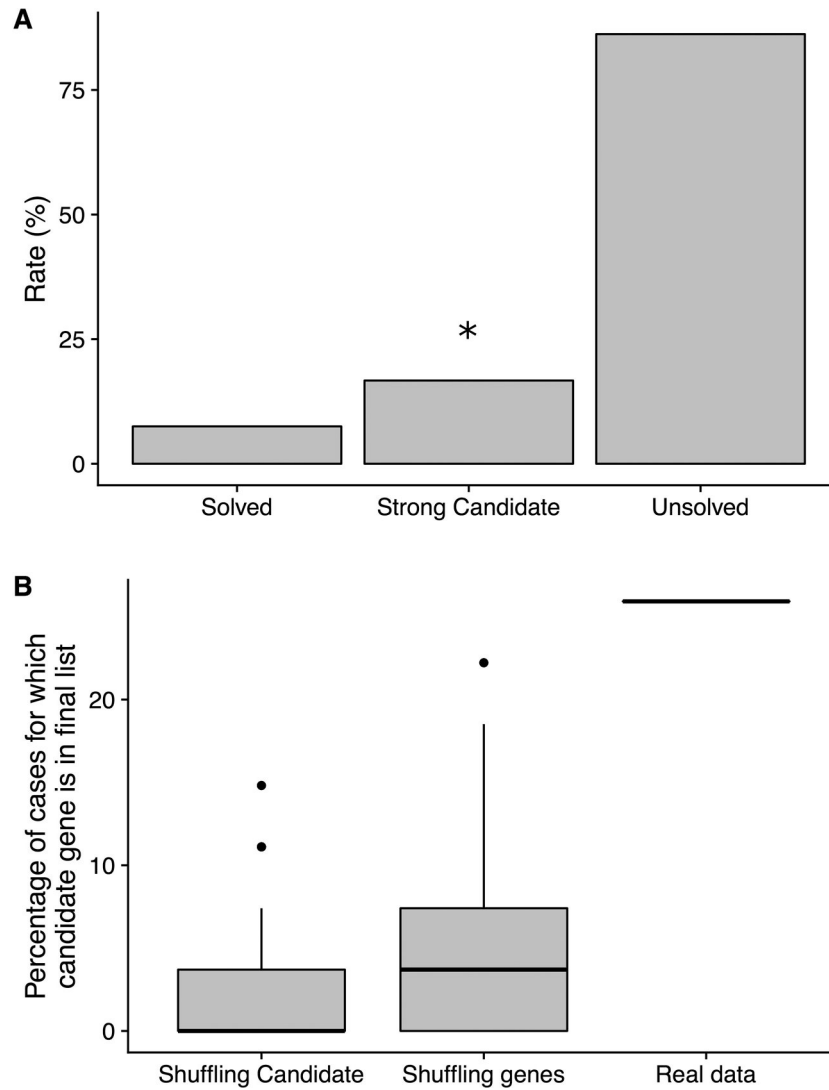
Extended Data Fig. 6. Correction for batch effects - Splicing data.

Analyses performed on 65 PIVUS samples and 143 rare disease samples. **a**, Plot of first two principal components (PCs) run on uncorrected splicing ratio data. Samples are coloured by batch. We observed that PC1 was separating PIVUS controls samples (left) from rare disease samples (right). **b**, Plot of first two PCs on splicing ratios after regressing out PCs explained up to 95% of the variance in the data. Batches were no longer separated on the first PCs. **c**, Correlation between known covariates 10 first PCs. We observed that PC1 is highly correlated with the batch, whereas PCs 2 and 3 separated samples from one institution (batch 1, CHEO) from others. We also observed that PC1 is highly correlated with RIN, highlighting differences in quality across samples.



Extended Data Fig. 7. Allele specific expression across rare disease samples.

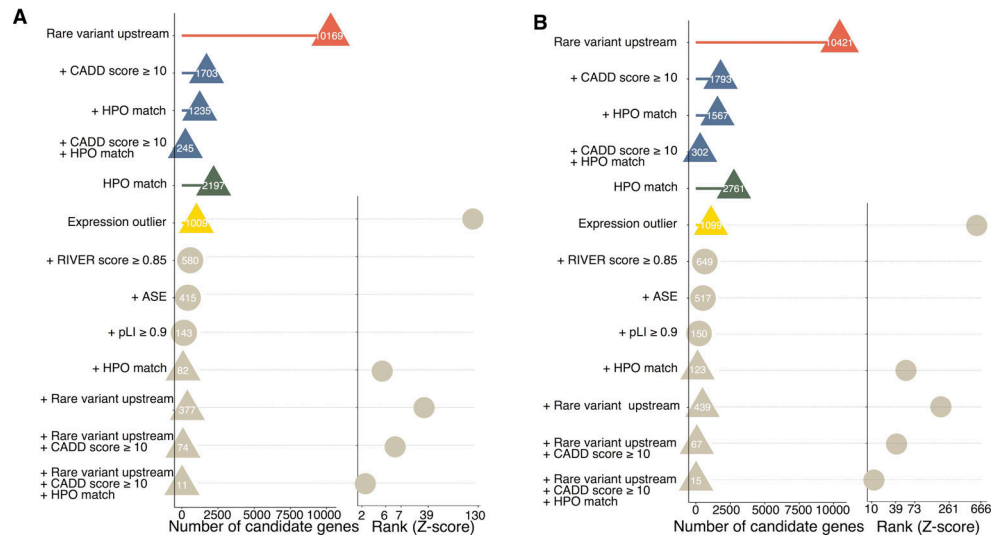
a, Prevalence of ASE events in rare diseases samples ($n = 112$). Results are displayed separately for exome and genome sequencing. **b**, Difference in proportion of genes matching HPO terms for top 20 ASE outliers per case in comparison to random genes (100 random gene sets for each sample, $n = 109$ samples). Analysis performed for all genes, genes with $pLI \geq 0.9$, genes with a rare variant (RV) and genes with a RV with CADD score ≥ 10 . The top 20 ASE outlier genes are enriched for overlap with HPO-associated genes per case, regardless of the filters applied to the extreme ASE genes and background genes (**** P value 1×10^{-4} , two-sided Wilcoxon test). For **a** and **b**, Boxplots represent median value, with lower and upper hinges corresponding to the 25th and 75th percentiles, and lower and upper whiskers extend from the hinge to the smallest and largest value at most $1.5 \times$ interquartile range of the hinge respectively. **c**, Rare deleterious variants are biased toward the alternative allele across all samples. A stop-gain variant was highly expressed in *EFHD2* for one sample where there were matching symptoms.



Extended Data Fig. 8. Diagnostic rate after analysis of 80 distinct cases.

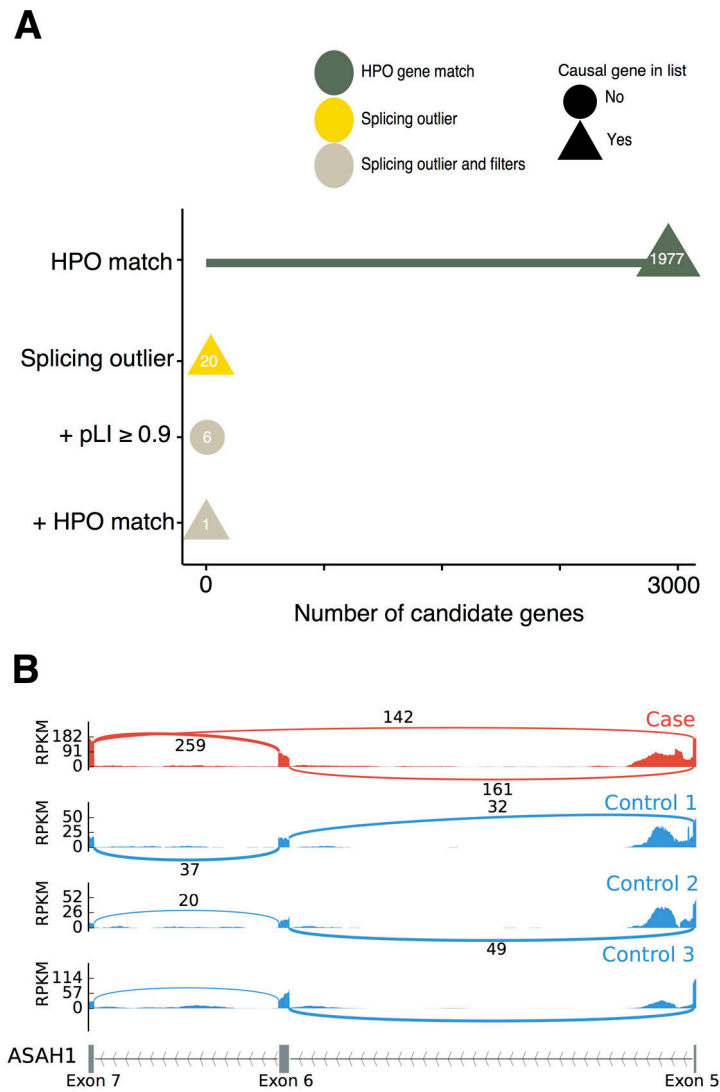
a, Overview of cases. Solved: causal gene found and further validated. Strong candidate: Strong candidate after RNA-seq analysis (out of a subset of 30 affected individuals for which we have prior candidate genes information from literature). Unsolved: Other cases for which further investigation is needed. **b**, Percentage of cases for which prior candidate gene is in final set of filtered genes (outlier with deleterious rare variant in a gene linked to symptoms). Analysis was performed only on a subset of 30 cases for which we have prior candidate gene information and for which we have genetic information. Shuffling candidates corresponds to the percentage of cases for which we observe a prior candidate genes in the most stringent gene list when shuffling gene lists across individuals (10,000 permutations). On average, no match is found. Shuffling genes correspond to the percentage of prior candidate genes we observed within the final set of DNA-only filters when sampling from this list a matched number of genes corresponding to the expression filters. Average matched percentage is 4.1% after 10,000 permutations. Real data corresponds to the percentage of cases for which we found a candidate gene in the most stringent RNA-based filter set. We

find a match for 7 affected samples out of 30, that is, 25.9 % of cases. There is significantly more match in real data in comparison to permuted data (two-sided Wilcoxon rank sum test, P value $< 10^{-5}$). Boxplots represent median value, with lower and upper hinges corresponding to the 25th and 75th percentiles, and lower and upper whiskers extend from the hinge to the smallest and largest value at most $1.5\times$ interquartile range of the hinge, respectively.



Extended Data Fig. 9. Identification of disease gene through expression outlier detection.

MECR case. a. Proband results. After our most stringent filter, there are 11 candidate genes left and *MECR* is rank 2nd by *Z*-score. **b.** Proband's brother. After filtering, only 15 out of 1,099 expression outliers are left and *MECR* is ranked 10th in that list.



Extended Data Fig. 10. Solved case without genetic data: ASAHI case.

a. After filtering our detected splicing outliers for genes related to the phenotype (through HPO IDs), only eight genes were left, with *ASAHI* being the most extreme outlier (Z -score = 3.9) and for which we previously confirmed the association with SMA-PME phenotype in the case. **b.** Sashimi plot of the case and 2 controls of the *ASAHI* gene. For the case (red track), we observed an alternative transcript skipping exon 6 (supported by 142 reads). This pattern was never observed in controls.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to thank the patients and their families for their participation in this study. SBM is supported by NIH grants R01HG008150 (NoVa), U01HG009080 (GSPAC) and the Glenn Center for Aging at Stanford. LF was supported by the Stanford Center for Computational, Evolutionary, and Human Genomics Fellowship. CS is

supported by BD2K Training Grant (T32 LM012409). NMF is supported by a National Science Foundation Graduate Research Fellowship. NAT is supported by the Stanford Genome Training Program (2T32HG000044–21). BL was supported by the Stanford Computational, Evolutionary and Human Genomics fellowship and the National Key R&D Program of China (2016YFD0400800). KMB is supported by a CIHR Foundation grant (FDN-154279). ZZ was supported by the CEHG Fellowship; the National Science Foundation (NSF) GRFP (DGE-114747); and by the Stanford Genome Training Program (SGTP; NIH/NHGRI T32HG000044). BB was supported by the Stanford Genome Training Program and Dean's Postdoctoral Fellowship. JRD was supported by a Lucille P. Markey Biomedical Research 688 Stanford Graduate Fellowship. JRD acknowledges the Stanford Genome Training Program 689 (SGTP; NIH/NHGRI T32HG000044). CJP is supported by the NIST/JIMB grant 70NANB15H268. AB is supported by NIH grant R01HG008150 (NoVa) and the Searle Scholar Fund. Clinical sample collection was supported, in part, by the Care4Rare Canada Consortium funded by Genome Canada, the Canadian Institutes of Health Research, the Ontario Genomics Institute, Ontario Research Fund, and Children's Hospital of Eastern Ontario Foundation. Research reported in this manuscript was in part supported by the NIH Common Fund, through the Office of Strategic Coordination/Office of the NIH Director under Award Number U01HG007708. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF & Hamosh A OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Research* 43, D789–D798 (2015). [PubMed: 25428349]
2. Boycott KM et al. International Cooperation to Enable the Diagnosis of All Rare Genetic Diseases. *The American Journal of Human Genetics* 100, 695–705 (2017). [PubMed: 28475856]
3. Gilissen C, Hoischen A, Brunner HG & Veltman JA Unlocking Mendelian disease using exome sequencing. *Genome Biology* 12, 228 (2011). [PubMed: 21920049]
4. Yang Y et al. Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders. *New England Journal of Medicine* 369, 1502–1511 (2013). [PubMed: 24088041]
5. Ewans LJ et al. Whole-exome sequencing reanalysis at 12 months boosts diagnosis and is cost-effective when applied early in Mendelian disorders. *Genet. Med* (2018). doi:10.1038/gim.2018.39
6. Cummings BB et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Sci Transl Med* 9, (2017).
7. Kremer LS et al. Genetic diagnosis of Mendelian disorders via RNA sequencing. *Nat Commun* 8, 15824 (2017). [PubMed: 28604674]
8. Kernohan KD et al. Whole-transcriptome sequencing in blood provides a diagnosis of spinal muscular atrophy with progressive myoclonic epilepsy. *Hum. Mutat* 38, 611–614 (2017). [PubMed: 28251733]
9. Hamanaka K et al. RNA sequencing solved the most common but unrecognized NEB pathogenic variant in Japanese nemaline myopathy. *Genetics in Medicine* (2018). doi:10.1038/s41436-018-0360-6
10. Battle A et al. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Research* 24, 14–24 (2014). [PubMed: 24092820]
11. Lind L A Comparison of Three Different Methods to Evaluate Endothelium-Dependent Vasodilation in the Elderly: The Prospective Investigation of the Vasculature in Uppsala Seniors (PIVUS) Study. *Arteriosclerosis, Thrombosis, and Vascular Biology* 25, 2368–2375 (2005).
12. GTEx Consortium et al. Genetic effects on gene expression across human tissues. *Nature* 550, 204–213 (2017). [PubMed: 29022597]
13. Hamosh A, Scott AF, Amberger JS, Bocchini CA & McKusick VA Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* 33, D514–517 (2005). [PubMed: 15608251]
14. Lek M et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536, 285–291 (2016). [PubMed: 27535533]
15. Zeng Y et al. Aberrant Gene Expression in Humans. *PLOS Genetics* 11, e1004942 (2015). [PubMed: 25617623]

16. Zhao J et al. A Burden of Rare Variants Associated with Extremes of Gene Expression in Human Peripheral Blood. *The American Journal of Human Genetics* 98, 299–309 (2016). [PubMed: 26849112]
17. Pala M et al. Population- and individual-specific regulatory variation in Sardinia. *Nat. Genet* 49, 700–707 (2017). [PubMed: 28394350]
18. Cao D & Parker R Computational Modeling and Experimental Analysis of Nonsense-Mediated Decay in Yeast. *Cell* 113, 533–545 (2003). [PubMed: 12757713]
19. Lykke-Andersen S & Jensen TH Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes. *Nature Reviews Molecular Cell Biology* 16, 665–677 (2015). [PubMed: 26397022]
20. Nickless A, Bailis JM & You Z Control of gene expression through the nonsense-mediated RNA decay pathway. *Cell & Bioscience* 7, (2017).
21. Li X et al. The impact of rare variation on gene expression across tissues. *Nature* 550, 239–243 (2017). [PubMed: 29022581]
22. Köhler S et al. The Human Phenotype Ontology in 2017. *Nucleic Acids Research* 45, D865–D876 (2017). [PubMed: 27899602]
23. Estivill X Genetic variation and alternative splicing. *Nature Biotechnology* 33, 357–359 (2015).
24. Xiong HY et al. The human splicing code reveals new insights into the genetic determinants of disease. *Science* 347, 1254806–1254806 (2015). [PubMed: 25525159]
25. Walter K et al. The UK10K project identifies rare variants in health and disease. *Nature* 526, 82–90 (2015). [PubMed: 26367797]
26. Soens ZT et al. Leveraging splice-affecting variant predictors and a minigene validation system to identify Mendelian disease-causing variants among exon-captured variants of uncertain significance. *Human Mutation* 38, 1521–1533 (2017). [PubMed: 28714225]
27. Albers CA 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* 44, 435–439, S1–2 (2012). [PubMed: 22366785]
28. Reinius B & Sandberg R Random monoallelic expression of autosomal genes: stochastic transcription and allele-level regulation. *Nat. Rev. Genet* 16, 653–664 (2015). [PubMed: 26442639]
29. Barbosa M et al. Identification of rare de novo epigenetic variations in congenital disorders. *Nat Commun* 9, 2064 (2018). [PubMed: 29802345]
30. Avramidou A et al. The novel adaptor protein Swiprosin-1 enhances BCR signals and contributes to BCR-induced apoptosis. *Cell Death Differ.* 14, 1936–1947 (2007). [PubMed: 17673920]
31. Kroczek C et al. Swiprosin-1/EFhd2 controls B cell receptor signaling through the assembly of the B cell receptor, Syk, and phospholipase C gamma2 in membrane rafts. *J. Immunol* 184, 3665–3676 (2010). [PubMed: 20194721]
32. Dütting S, Brachs S & Mielenz D Fraternal twins: Swiprosin-1/EFhd2 and Swiprosin-2/EFhd1, two homologous EF-hand containing calcium binding adaptor proteins with distinct functions. *Cell Communication and Signaling* 9, 2 (2011). [PubMed: 21244694]
33. Thylur RP, Gowda R, Mishra S & Jun C-D Swiprosin-1: Its Expression and Diverse Biological Functions. *J. Cell. Biochem* 119, 150–156 (2018). [PubMed: 28590012]
34. Heimer G et al. MECP Mutations Cause Childhood-Onset Dystonia and Optic Atrophy, a Mitochondrial Fatty Acid Synthesis Disorder. *The American Journal of Human Genetics* 99, 1229–1244 (2016). [PubMed: 27817865]
35. Kircher M et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics* 46, 310–315 (2014). [PubMed: 24487276]
36. Eldomery MK et al. Lessons learned from additional research analyses of unsolved clinical exome cases. *Genome Med* 9, 26 (2017). [PubMed: 28327206]
37. Wright CF et al. Making new genetic diagnoses with old data: iterative reanalysis and reporting from genome-wide data in 1,133 families with developmental disorders. *Genet. Med* (2018). doi: 10.1038/gim.2017.246

38. Dewey FE et al. Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. *Science* 354, (2016).
39. Eilbeck K, Quinlan A & Yandell M Settling the score: variant prioritization and Mendelian disease. *Nature Reviews Genetics* 18, 599–612 (2017).
40. Rao AR & Nelson SF Calculating the statistical significance of rare variants causal for Mendelian and complex disorders. *BMC Med Genomics* 11, 53 (2018). [PubMed: 29898714]
41. Li B & Dewey CN RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12, 323 (2011). [PubMed: 21816040]
42. Tange O GNU Parallel - The Command-Line Power Tool, ;login. *The USENIX Magazine* (2011).
43. Li H & Durbin R Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760 (2009). [PubMed: 19451168]
44. DePristo MA et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* 43, 491–498 (2011). [PubMed: 21478889]
45. Ganna A et al. Quantifying the Impact of Rare and Ultra-rare Coding Variation across the Phenotypic Spectrum. *The American Journal of Human Genetics* 102, 1204–1211 (2018). [PubMed: 29861106]
46. Quinlan AR & Hall IM BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842 (2010). [PubMed: 20110278]
47. Pedersen BS, Layer RM & Quinlan AR Vcfanno: fast, flexible annotation of genetic variants. *Genome Biology* 17, (2016).
48. McLaren W et al. The Ensembl Variant Effect Predictor. *Genome Biology* 17, (2016).
49. Quang D, Chen Y & Xie X DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* 31, 761–763 (2015). [PubMed: 25338716]
50. Davydov EV et al. Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. *PLoS Computational Biology* 6, e1001025 (2010). [PubMed: 21152010]
51. Pollard KS, Hubisz MJ, Rosenbloom KR & Siepel A Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research* 20, 110–121 (2010). [PubMed: 19858363]
52. Siepel A Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research* 15, 1034–1050 (2005). [PubMed: 16024819]
53. Ernst J & Kellis M ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods* 9, 215–216 (2012). [PubMed: 22373907]
54. Stegle O, Parts L, Piipari M, Winn J & Durbin R Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature Protocols* 7, 500–507 (2012). [PubMed: 22343431]
55. Castel SE, Levy-Moonshine A, Mohammadi P, Banks E & Lappalainen T Tools and best practices for data processing in allelic expression analysis. *Genome Biology* 16, (2015).
56. McKenna A et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* 20, 1297–1303 (2010). [PubMed: 20644199]

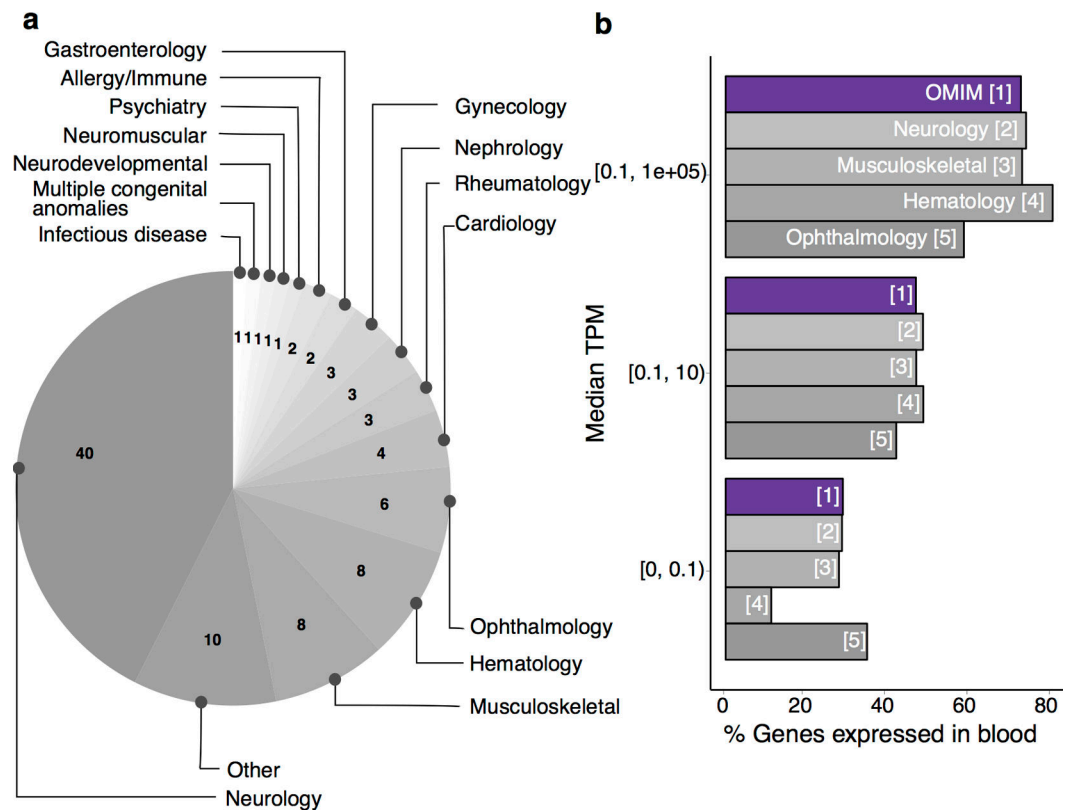


Figure 1: Using blood RNA-seq to study rare disease genes.

a: Disease categories of sequenced affected patients. The majority of cases belong to neurology (n=40), musculoskeletal and orthopedics(n=8), hematology (n=8) and ophthalmology (n=8) categories. **b:** Percentage of disease genes (from curated lists) expressed in blood. We used the median TPM across 909 DGN samples, 65 PIVUS samples and our 143 samples.

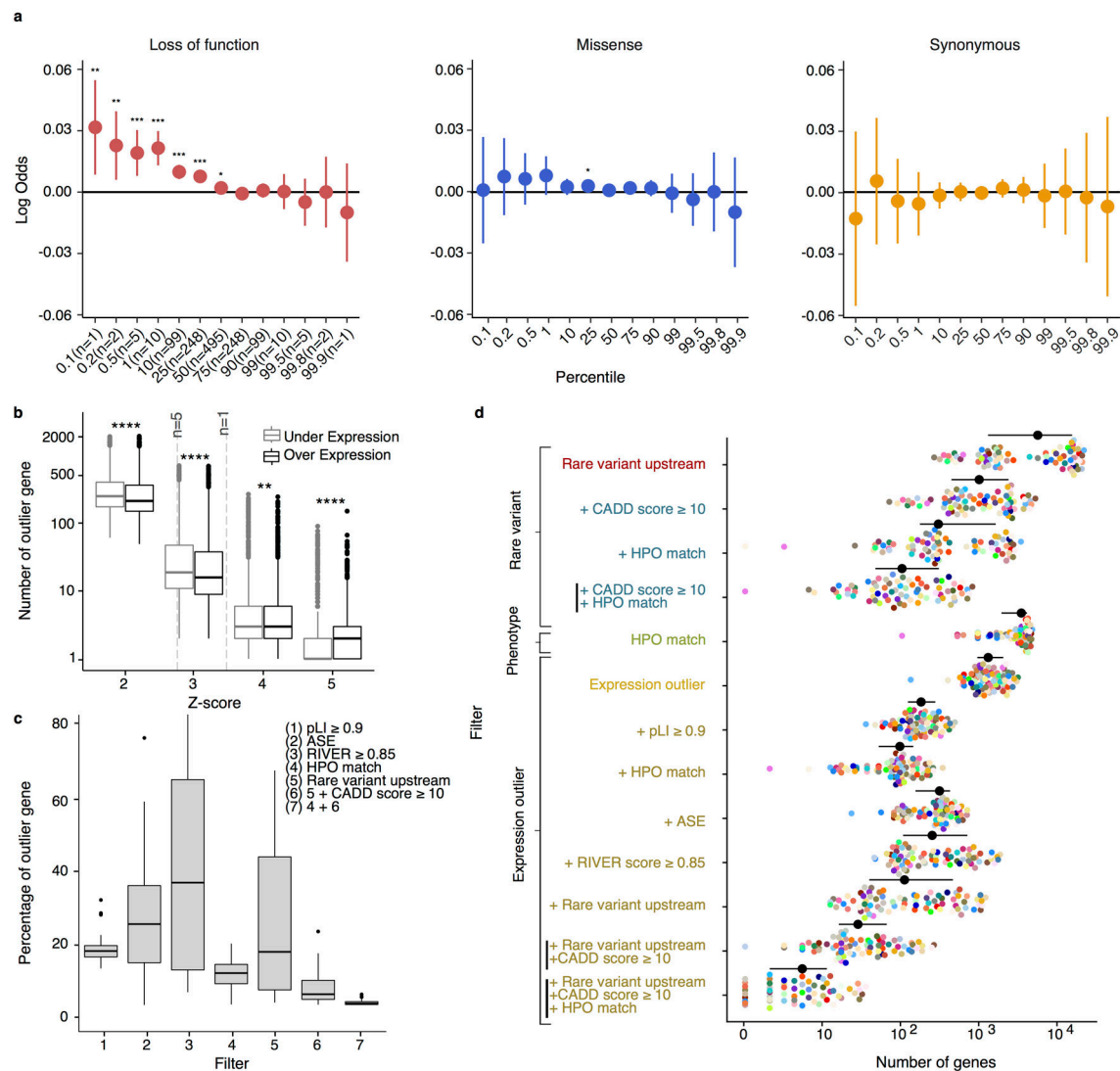


Figure 2: Expression outliers in rare disease samples.

a: Enrichment for case or control outlier genes in intolerance to LoF (red), missense (blue) and synonymous (yellow) mutations at different percentiles of gene expression. Represented here as the log odds ratio with 95% Wald confidence intervals using 931 samples. P-values were calculated based on the z -statistic. No adjustment for multiple testing. **b** Impact of Z-score thresholds on number of outliers. Differences between under and over expression outliers were tested using a two-sided Wilcoxon rang sum test. No adjustment for multiple testing. Vertical dashed lines indicate mean Z-score for n=1 and n=5 percentiles across all genes used in analysis (N = 14,988). For **a** and **b**: significance level: **** p-value 1×10^{-4} ; *** p-value 1×10^{-3} ; ** p-value 1×10^{-2} ; * p-value 5×10^{-2} . **c:** Proportion of under-expression outlier genes remaining after filters (n=75 cases with genetic information). Adding genetic information (rare variant within 10kb upstream of the gene body) allows to filter down to 50% of the original set of outliers. Keeping only genes for which HPO information of the affected individual match helps narrow down to less than 10% of candidates. For **b** and **c**: Boxplots represent median value, with lower and upper hinges

corresponding to the 25th and 75th percentiles, and lower and upper whiskers extend from the hinge to the smallest and largest value at most $1.5 \times$ inter-quartile range of the hinge respectively. **d**: Number of candidate genes at each filter for all cases with genetic information (n=75). Average number of candidate genes and s.d are represented in black for each filter.

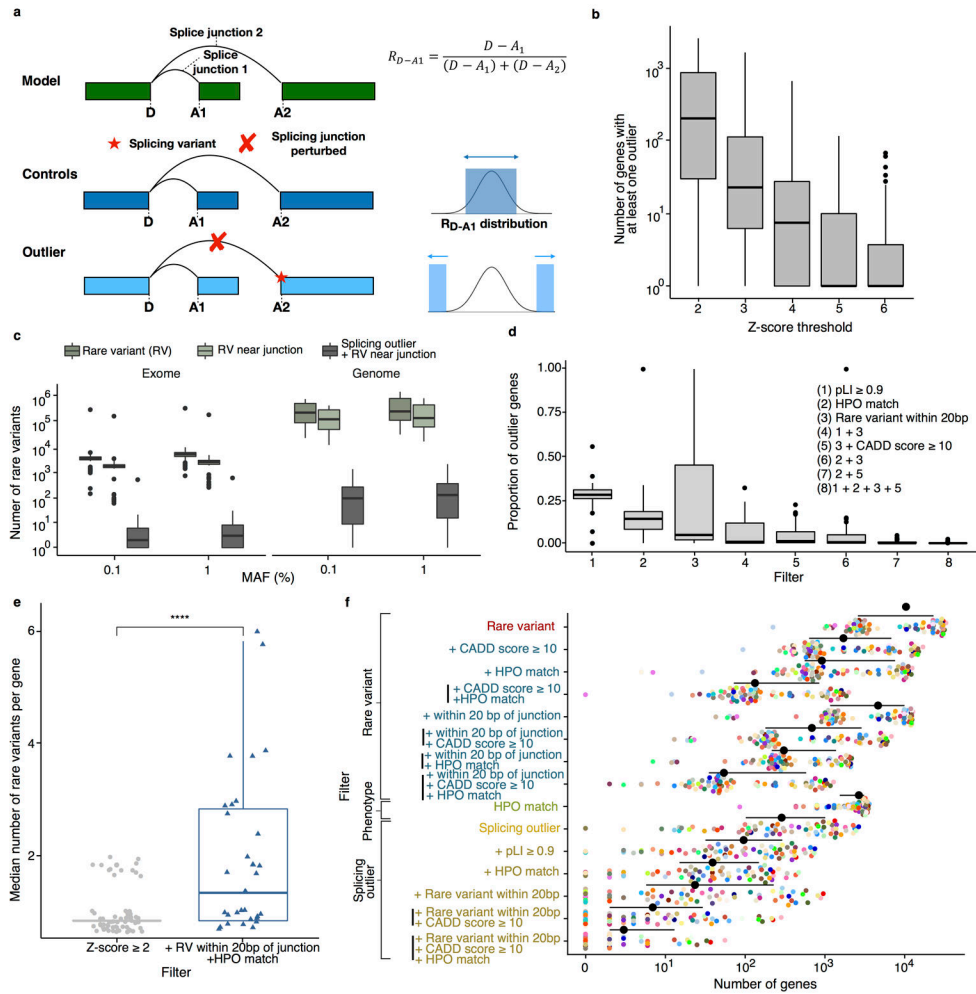


Figure 3: Splicing outlier detection.

a: Splicing outlier definition. Gene model is in green, rectangles represent 3 exons. In this model, we show junction information for one donor (D) and two acceptors (A1 and A2). For each sample for this gene we have coverage information for the two existing splicing junctions (D-A1 and D-A2). We defined the proportion of one splice junction as the number of reads overlapping this junction divided by the total number of reads spanning all junctions from a common donor (or acceptor). **b:** Number of genes with at least one splicing outlier at different Z-score thresholds (n=208 samples from PIVUS and rare disease cohorts). **c:** Number of rare variants in each sample, in total, nearby junction and associated with a splicing outlier (n=111 samples with genetic information). Rare variants were defined as variants with MAF \leq 0.1%. **d:** Impact of different filters on splicing outlier discoveries (n=75 cases with genetic information). **e:** We observed a significant increase (two-sided Wilcoxon test, p-value 9.8×10^{-5}) in the median number of rare variants with CADD score \geq 10 in the gene when filtering outliers with a rare variant within 20 bp of the junction and relevant to the disease phenotype (HPO match) (n=109 samples with splicing outliers and genetic data). **f:** Number of candidate genes at each filter for all cases with genetic information (n=75). Mean value and s.d. are represented in black for each filter. For **b** and **d** and **e**: Boxplots represent median value, with lower and upper hinges corresponding to the 25th and 75th

percentiles, and lower and upper whiskers extend from the hinge to the smallest and largest value at most $1.5 * \text{inter-quartile range}$ of the hinge respectively.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

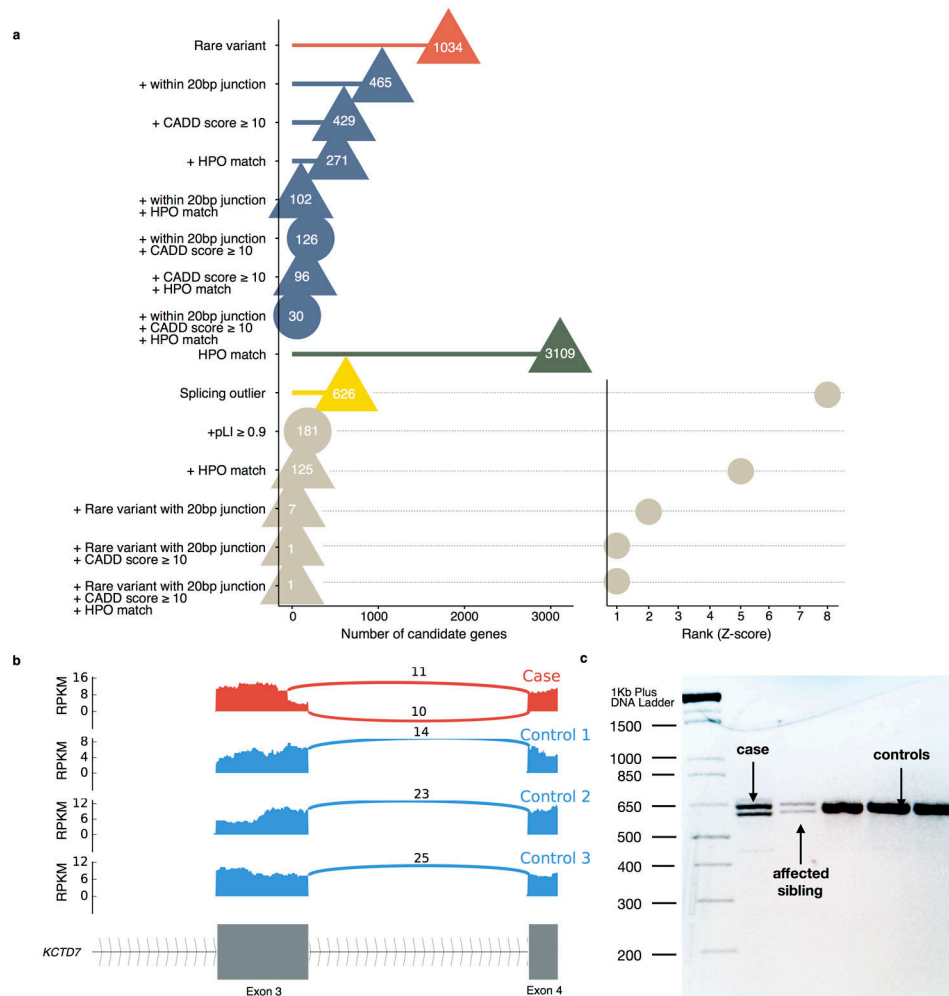


Figure 4: Identification of disease gene through splicing outlier detection.

a: Splicing outliers in the *KCTD7* case. Number of candidate genes obtained throughout different filters using genetic data, splicing outlier data, and phenotypic data. Shape indicates if the causal gene is in the list. After filtering for splicing outliers with a deleterious rare splice variant within 20 bp of a splice junction and limiting the search to genes for which there is a link to phenotype (HPO match), 1 candidate gene was left, *KCTD7*. **b:** Sashimi plot of the case and 3 controls of the splicing gain region in *KCTD7*. For the case only (red track), we observed a new splicing junction ahead of the annotated one in exon 3. **c:** cDNA gel from fibroblast cDNA of exons 2–4 of *KCTD7* for the proband, her affected sibling and three unaffected controls (no independent replicate). Both for the case and her affected brother we observed 2 fragments of different size, corresponding to the alternative splice products induced by the splice-gain mutation. In control samples, only one fragment is observed, corresponding to the original transcript.