nature medicine

LETTERS

https://doi.org/10.1038/s41591-019-0457-8

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

Laure Frésard[1*], Craig Smail[2], Nicole M. Ferraro[2], Nicole A. Teran[3], Xin Li[1], Kevin S. Smith[1], Devon Bonner[4], Kristin D. Kernohan[5], Shruti Marwaha[4,6], Zachary Zappala[3], Brunilda Balliu[1], Joe R. Davis[3], Boxiang Liu[7], Cameron J. Prybol[3], Jennefer N. Kohler[4], Diane B. Zastrow[4], Chloe M. Reuter[4], Dianna G. Fisk[8], Megan E. Grove[8], Jean M. Davidson[4], Taila Hartley[9], Ruchi Joshi[8], Benjamin J. Strober[10], Sowmithri Utiramerur[8], Undiagnosed Diseases Network[11], Care4Rare Canada Consortium[11], Lars Lind[12], Erik Ingelsson[6,13], Alexis Battle[10,14], Gill Bejerano[15,16,17,18], Jonathan A. Bernstein[16], Euan A. Ashley[3,4,13], Kym M. Boycott[9], Jason D. Merker[1,8,19], Matthew T. Wheeler[4,6] and Stephen B. Montgomery[1,3*]

It is estimated that 350 million individuals worldwide suffer from rare diseases, which are predominantly caused by mutation in a single gene[1]. The current molecular diagnostic rate is estimated at 50%, with whole-exome sequencing (WES) among the most successful approaches[2–5]. For patients in whom WES is uninformative, RNA sequencing (RNA-seq) has shown diagnostic utility in specific tissues and diseases[6–8]. This includes muscle biopsies from patients with undiagnosed rare muscle disorders[6,9], and cultured fibroblasts from patients with mitochondrial disorders[7]. However, for many individuals, 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 individuals with undiagnosed rare diseases spanning 16 diverse disease categories. We developed a robust approach to compare data from these individuals with large sets of RNA-seq data for controls ($n = 1{,}594$ unrelated controls and $n = 49$ family members) 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 additional 16.7% with improved candidate gene resolution.

We obtained RNA-seq data from samples from 143 individuals, 94 affected by rare diseases and 49 unaffected family members (Supplementary Table 1), and WES or whole-genome sequencing for 112 of them. In total, WES did not to 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 and 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) ($n = 620$) cohorts[12] (Supplementary Table 3). By comparison with 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 and Supplementary Table 4). When broadly considering disease genes from the Online Mendelian Inheritance in Man (OMIM) database[13], we observed that 70.6% were expressed in blood and that 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 Exome Aggregation Consortium (ExAC)[14], we further observed that genes expressed across multiple tissues were

[1]Department of Pathology, School of Medicine, Stanford University, Stanford, CA, USA. [2]Biomedical Informatics Program, Stanford University, Stanford, CA, USA. [3]Department of Genetics, School of Medicine, Stanford University, Stanford, CA, USA. [4]Stanford Center for Undiagnosed Diseases, Stanford University, Stanford, CA, USA. [5]Newborn Screening Ontario (NSO), Children's Hospital of Eastern Ontario, Ottawa, Ontario, Canada. [6]Stanford Cardiovascular Institute, School of Medicine, Stanford University, Stanford, CA, USA. [7]Department of Biology, School of Humanities and Sciences, Stanford University, Stanford, CA, USA. [8]Stanford Medicine Clinical Genomics Program, School of Medicine, Stanford University, Stanford, CA, USA. [9]Children's Hospital of Eastern Ontario Research Institute, University of Ottawa, Ottawa, Ontario, Canada. [10]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA. [11]A list of members and affiliations appears at the end of the paper. [12]Department of Medical Sciences, Cardiovascular Epidemiology, Uppsala University, Uppsala, Sweden. [13]Department of Medicine, Division of Cardiovascular Medicine, School of Medicine, Stanford University, Stanford, CA, USA. [14]Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA. [15]Department of Computer Science, Stanford University, Stanford, CA, USA. [16]Department of Pediatrics, School of Medicine, Stanford University, Stanford, CA, USA. [17]Department of Developmental Biology, School of Medicine, Stanford University, Stanford, CA, USA. [18]Department of Biomedical Data Science, School of Medicine, Stanford University, Stanford, CA, USA. [19]Present address: Departments of Pathology and Laboratory Medicine & Genetics, Lineberger Comprehensive Cancer Center, University of North Carolina School Medicine, Chapel Hill, NC, USA. *e-mail: lfresard@stanford.edu; smontgom@stanford.edu

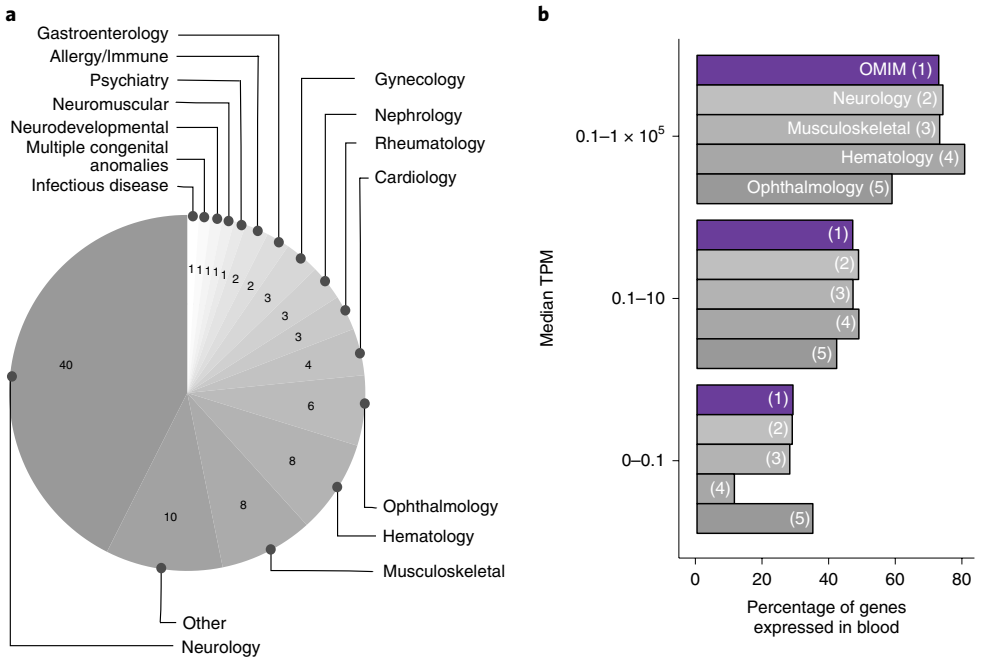


Fig. 1 | Using blood RNA-seq to study rare-disease genes. a, Disease categories of sequenced affected patients. The majority of individuals have diseases in the neurology ($n=40$), musculoskeletal and orthopedics ($n=8$), hematology ($n=8$) and ophthalmology ($n=8$) categories. **b**, Percentages of disease genes (from curated lists) expressed in blood. We used the median TPM across 909 DGN samples, 65 PIVUS samples and our 143 samples.

more depleted in missense or loss-of-function (LoF) mutations[14] (two-sided Wilcoxon test, $P \leq 2 \times 10^{-16}$; Extended Data Fig. 1c and 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 with 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 individuals with disease versus controls after correcting the data for batch effects (see Methods and Extended Data Figs. 2 and 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[18–20]. As we increased the number of controls, the enrichment became stronger, demonstrating the impact of large control data sets (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), were likely to have a regulatory variant impacting gene expression (Filter 2: RNA-informed variant effect on regulation (RIVER) score ≥ 0.85 (ref. [21])), showed allele-specific expression (ASE) (Filter 3), were linked to the phenotype (Filter 4: Human Phenotype Ontology[22] (HPO) match), had a rare variant with minor allele frequency (MAF) $\leq 0.1\%$ within 10 kb upstream of the gene (Filter 5) and had a rare variant that was probably deleterious (Filter 7: Combined Annotation Dependent Depletion (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 individuals with disease having at least 1 candidate gene (Filter 11; Fig. 2c and Extended Data Fig. 5a). Overall, we were able to reduce the number of expression outliers to less than ten on average for all individuals (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 and 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 junctions in each gene, was higher in individuals with disease and, similar to 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 substantially 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, by limiting to genes relevant to the phenotype (Filter 2) and with a deleterious rare variant within 20 base pairs (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 individuals with disease had at least 1 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 reduced candidates to less than ten per individual on average (Fig. 3f).

RNA-seq provides the ability to measure ASE. ASE can inform the presence of a large-effect heterozygous regulatory, splicing or nonsense variant, or epi-mutation, aiding the identification of candidate rare-disease genes and variants[7,27–29]. Of all possible heterozygous sites (~10^5 and ~10^6 per sample for exome and genome, respectively), 10^4 variants had sufficient coverage for analysis

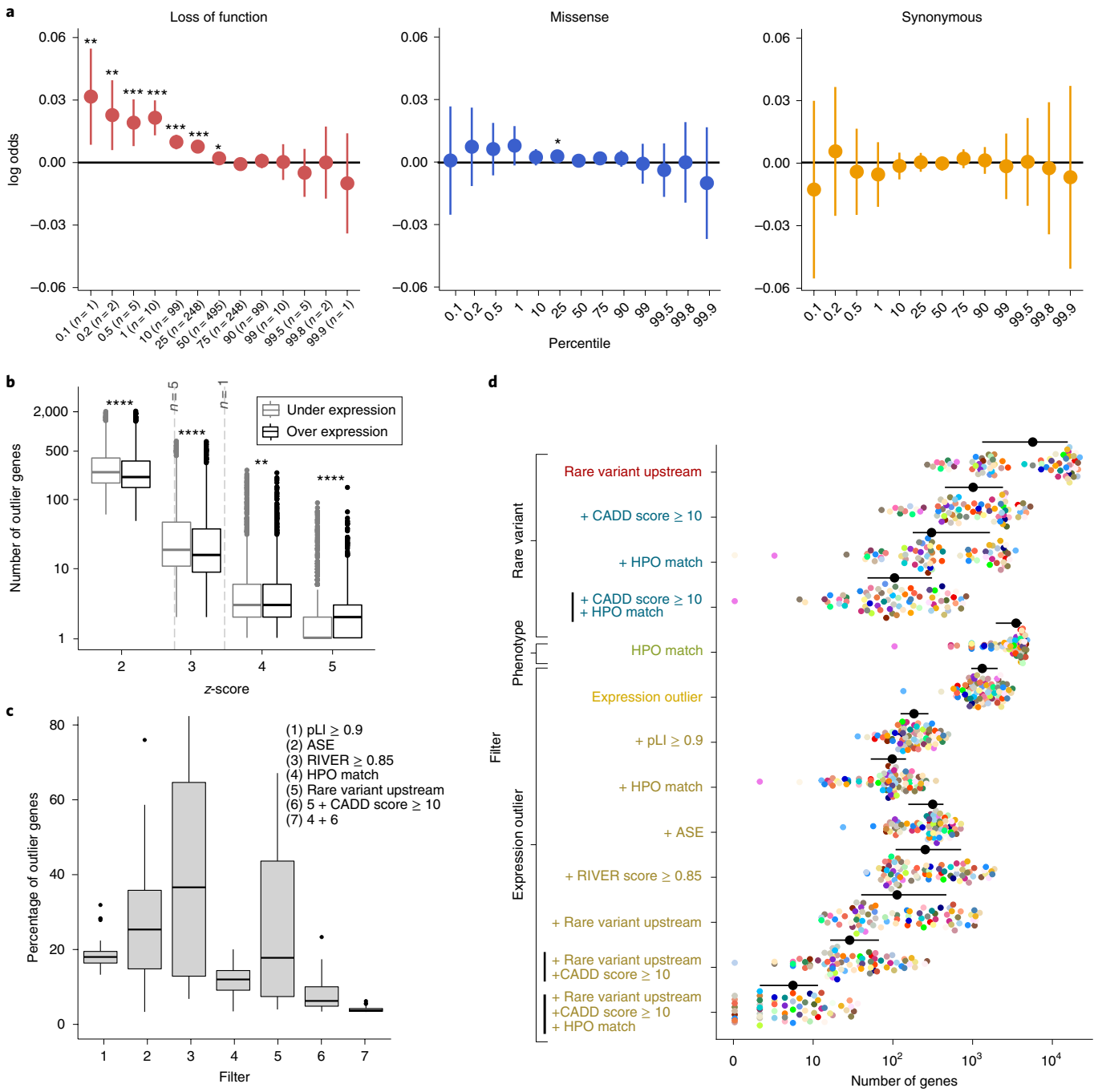


Fig. 2 | Expression outliers in rare-disease samples. a, Enrichment for outlier genes in controls or individuals with disease in intolerance to LoF (red), missense (blue) and synonymous (yellow) mutations at different percentiles of gene expression. This is represented as the log odds ratio with 95% Wald confidence intervals using 931 samples. *P* values were calculated on the basis of the *z*-statistic. There was 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 rank 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 levels were: $^{****}P \leq 1\times10^{-4}$; $^{***}P \leq 1\times10^{-3}$; $^{**}P \leq 1\times10^{-2}$; $^{*}P \leq 5\times10^{-2}$. **c**, Proportion of under-expression outlier genes remaining after filters ($n=75$ individuals with genetic information). Adding genetic information (rare variant within 10 kb upstream of the gene body) allows filtering down to 50% of the original set of outliers. Keeping only genes for which HPO information of the affected individual matches helps narrow down to less than 10% of candidates. For **b** and **c**, boxplots represent the median value, with lower and upper hinges corresponding to the 25th and 75th percentiles, and lower and upper whiskers extending from the hinge to the smallest and largest value that is at most 1.5× interquartile range of the hinge, respectively. **d**, Number of candidate genes at each filter for all individuals with genetic information ($n=75$). Average number of candidate genes and s.d. are represented in black for each filter. Vertical bars were used to highlight filters that were used together at a particular instance. They were used only with more than one filter was applied.

(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 gene outlier ASE sites within the individuals with disease whom we studied when compared with all other rare-disease individuals and GTEx samples

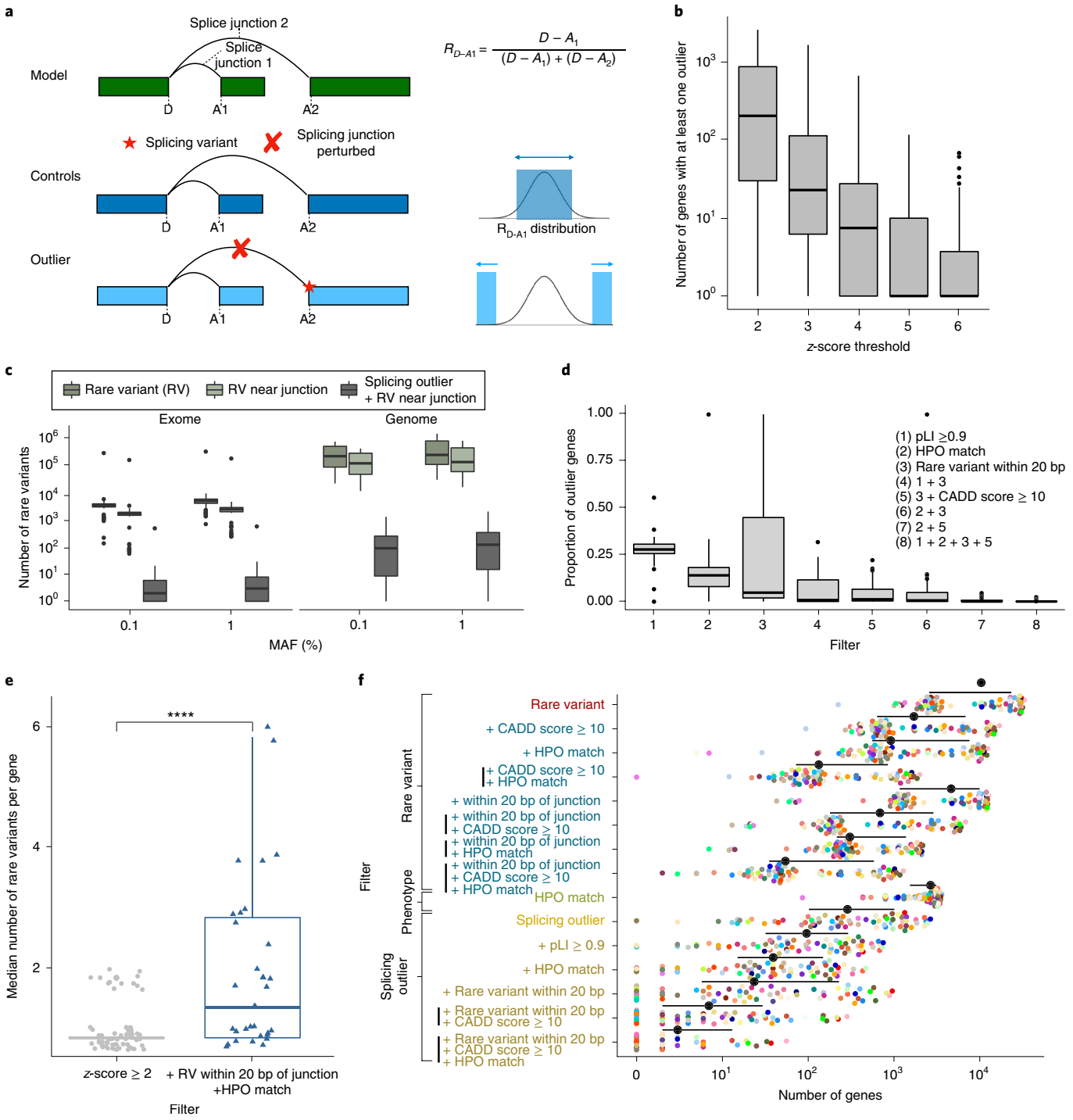


Fig. 3 | Splicing outlier detection. a, Splicing outlier definition. The gene model is in green, and rectangles represent three exons. In this model, we show junction information for one donor (D) and two acceptors (A1 and A2). For each sample for this gene, we have coverage information for the two existing 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 1 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 ≤ 0.1%. **d**, Impact of different filters on splicing outlier discoveries ($n = 75$ individuals with disease 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 ≥ 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 individuals with disease with genetic information ($n = 75$). Mean value and s.d. are represented in black for each filter. Vertical bars were used to highlight filters that were used together at a particular instance. They were used only with more than one filter was applied. For **b**, **d** and **e**, boxplots represent median value, with lower and upper hinges corresponding to the 25th and 75th percentiles, and lower and upper whiskers extending from the hinge to the smallest and largest value at most 1.5× interquartile range of the hinge, respectively.

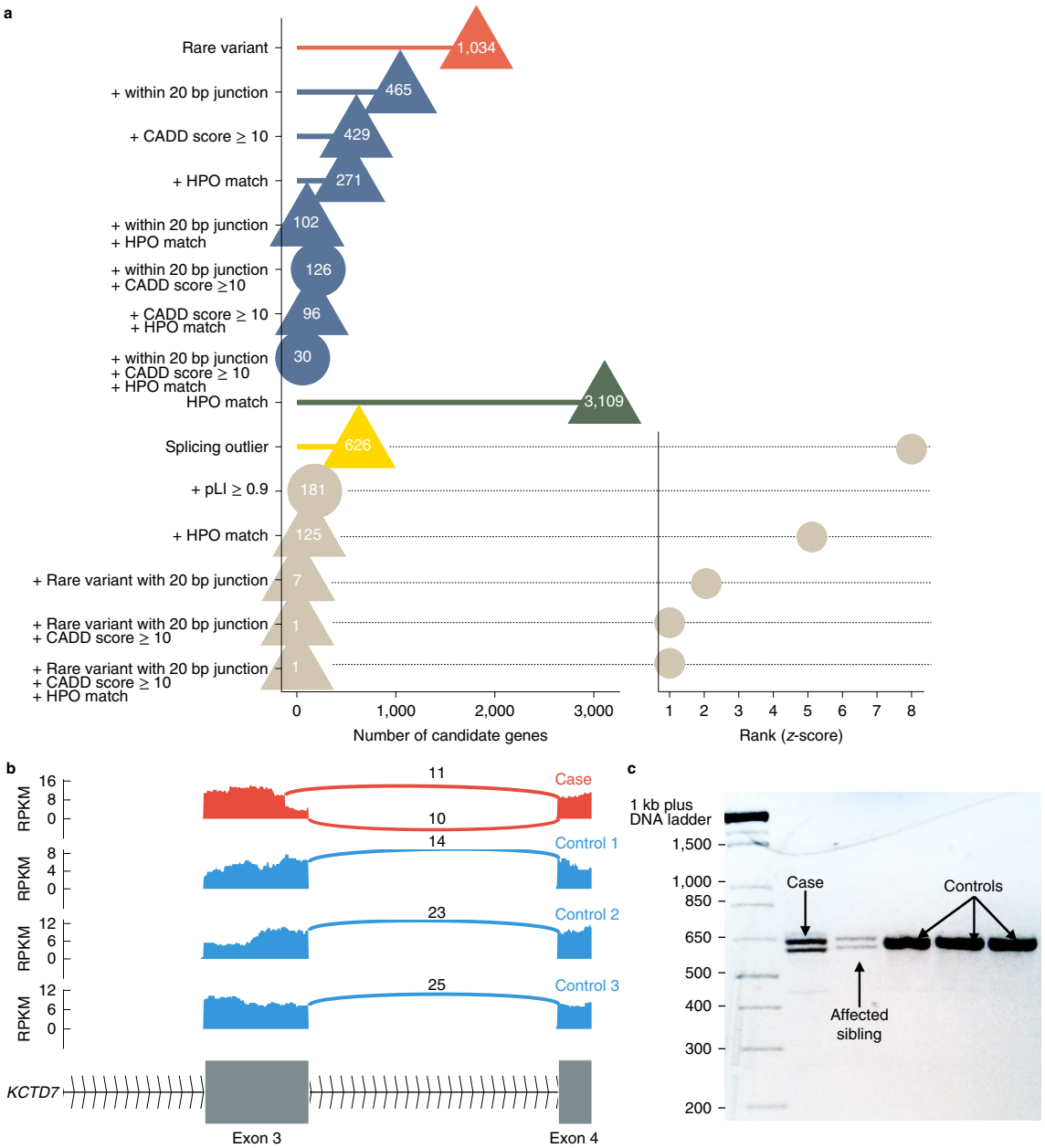


Fig. 4 | Identification of disease gene through splicing outlier detection. a, Splicing outliers in the individual with mutant KCTD7. The number of candidate genes was obtained throughout different filters using genetic data, splicing outlier data and phenotypic data. Shape indicates if the causal gene is in the list of genes left after filtering for the displayed criteria (triangle, yes; circle, no). 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), one candidate gene was left, *KCTD7*. **b**, Sashimi plot of the proband and three controls of the splicing gain region in *KCTD7*. For the proband only (red track), we observed a 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 3 unaffected controls (no independent replicate). Both for the proband and her affected brother, we observed two fragments of different sizes, 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. Reads per kilobase of transcript, per million mapped reads (RPKM) represents the level of expression of *KCTD7* for the samples displayed.

(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 (that is, 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 them, one variant is in *EFHD2*, a gene coding for Ca^{2+} adapter protein involved in B cell apoptosis, NF-kB-mediated 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 of 80 independent individuals with disease (7.5%, 4 expression outliers, 2 splicing outliers), and identify candidate genes potentially linked to the disease phenotype (gene matching HPO terms for the symptoms of the proband) in 5 of 30 cases with candidate gene information (16.7%) (Extended Data Fig. 8a and Supplementary Table 1). We did not find highly relevant candidate genes for 69 individuals (86%). Notably, candidates were identified for five 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 the first case, of two brothers aged 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 literature[34]. The c.830+2dupT variant has been described previously[34]. 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 were 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 10 kb (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 20 bp of annotated junctions with a CADD score ≥ 10 (ref. [35]). The causal gene is missing from the most stringent filter because there are no rare deleterious variants within 20 bp of known junctions. Adding splicing outlier information from RNA-seq data left us with 1 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 to be responsible for the creation of a splicing junction in this gene (p.V152V; Fig. 4b). PCR with reverse transcription of RNA extracted from fibroblasts from exon 2–4 regions of the gene confirmed a difference in fragment size in the probands (Fig. 4c and Source Data). In addition, this variant exhibited monoallelic expression toward 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 study[8]. 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 8 genes were left (Extended Data Fig. 10a), *ASAH1* being the strongest outlier, and for which we subsequently identified with Sanger sequencing a splice-loss mutation leading to the creation of a transcript, skipping exon 6 (Extended Data Fig. 10b). While we obtained genetic data for many individuals in this study, 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 individuals with disease 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 sets 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 to be 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.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41591-019-0457-8.

Received: 30 August 2018; Accepted: 15 April 2019;
Published online: 3 June 2019

References

1. Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh, A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* **43**, D789–D798 (2015).
2. Boycott, K. M. et al. International cooperation to enable the diagnosis of all rare genetic diseases. *Am. J. Hum. Genet.* **100**, 695–705 (2017).
3. Gilissen, C., Hoischen, A., Brunner, H. G. & Veltman, J. A. Unlocking Mendelian disease using exome sequencing. *Genome Biol.* **12**, 228 (2011).
4. Yang, Y. et al. Clinical whole-exome sequencing for the diagnosis of Mendelian disorders. *N. Engl. J. Med.* **369**, 1502–1511 (2013).
5. Ewans, L. J. et al. Whole-exome sequencing reanalysis at 12 months boosts diagnosis and is cost-effective when applied early in Mendelian disorders. *Genet. Med.* **20**, 1564–1574 (2018).
6. Cummings, B. B. et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Sci. Transl. Med.* **9**, eaal5209 (2017).
7. Kremer, L. S. et al. Genetic diagnosis of Mendelian disorders via RNA sequencing. *Nat. Commun.* **8**, 15824 (2017).
8. Kernohan, K. D. et al. Whole-transcriptome sequencing in blood provides a diagnosis of spinal muscular atrophy with progressive myoclonic epilepsy. *Hum. Mutat.* **38**, 611–614 (2017).
9. Hamanaka, K. et al. RNA sequencing solved the most common but unrecognized NEB pathogenic variant in Japanese nemaline myopathy. *Genet. Med.* https://doi.org/10.1038/s41436-018-0360-6. (2018).
10. Battle, A. et al. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res.* **24**, 14–24 (2014).

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. *Arterioscler. Thromb. Vasc. Biol.* **25**, 2368–2375 (2005).
12. GTEx Consortium Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
13. Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. & McKusick, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33**, D514–D517 (2005).
14. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
15. Zeng, Y. et al. Aberrant gene expression in humans. *PLOS Genet.* **11**, e1004942 (2015).
16. Zhao, J. et al. A burden of rare variants associated with extremes of gene expression in human peripheral blood. *Am. J. Hum. Genet.* **98**, 299–309 (2016).
17. Pala, M. et al. Population- and individual-specific regulatory variation in Sardinia. *Nat. Genet.* **49**, 700–707 (2017).
18. Cao, D. & Parker, R. Computational modeling and experimental analysis of nonsense-mediated decay in yeast. *Cell* **113**, 533–545 (2003).
19. Lykke-Andersen, S. & Jensen, T. H. Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes. *Nat. Rev. Mol. Cell Biol.* **16**, 665–677 (2015).
20. Nickless, A., Bailis, J. M. & You, Z. Control of gene expression through the nonsense-mediated RNA decay pathway. *Cell Biosci.* **7**, 26 (2017).
21. Li, X. et al. The impact of rare variation on gene expression across tissues. *Nature* **550**, 239–243 (2017).
22. Köhler, S. et al. The Human Phenotype Ontology in 2017. *Nucleic Acids Res.* **45**, D865–D876 (2017).
23. Estivill, X. Genetic variation and alternative splicing. *Nat. Biotechnol.* **33**, 357–359 (2015).
24. Xiong, H. Y. et al. The human splicing code reveals new insights into the genetic determinants of disease. *Science* **347**, 1254806–1254806 (2015).
25. Walter, K. et al. The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015).
26. Soens, Z. T. 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. *Hum. Mutat.* **38**, 1521–1533 (2017).
27. Albers, C. A. 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**, S1–S2 (2012).
28. Reinius, B. & Sandberg, R. Random monoallelic expression of autosomal genes: stochastic transcription and allele-level regulation. *Nat. Rev. Genet.* **16**, 653–664 (2015).
29. Barbosa, M. et al. Identification of rare de novo epigenetic variations in congenital disorders. *Nat. Commun.* **9**, 2064 (2018).
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).
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).
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 Commun. Signal.* **9**, 2 (2011).
33. Thylur, R. P., Gowda, R., Mishra, S. & Jun, C.-D. Swiprosin-1: its expression and diverse biological functions. *J. Cell. Biochem.* **119**, 150–156 (2018).
34. Heimer, G. et al. MECR mutations cause childhood-onset dystonia and optic atrophy, a mitochondrial fatty acid synthesis disorder. *Am. J. Hum. Genet.* **99**, 1229–1244 (2016).
35. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
36. Eldomery, M. K. et al. Lessons learned from additional research analyses of unsolved clinical exome cases. *Genome Med.* **9**, 26 (2017).
37. Wright, C. F. 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.* **20**, 1216–1223 (2018).
38. Dewey, F. E. et al. Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. *Science* **354**, aaf6814 (2016).
39. Eilbeck, K., Quinlan, A. & Yandell, M. Settling the score: variant prioritization and Mendelian disease. *Nat. Rev. Genet.* **18**, 599–612 (2017).
40. Rao, A. R. & Nelson, S. F. Calculating the statistical significance of rare variants causal for Mendelian and complex disorders. *BMC Med. Genomics* **11**, 53 (2018).

Acknowledgements

The authors would like to thank the patients and their families for their participation in this study. S.B.M. is supported by NIH grants nos. R01HG008150 (NoVa) and U01HG009080 (GSPAC) and the Glenn Center for Aging at Stanford. L.F. was supported by the Stanford Center for Computational, Evolutionary, and Human Genomics Fellowship. C.S. is supported by a BD2K Training Grant (T32 LM012409). N.M.F. is supported by a National Science Foundation Graduate Research Fellowship. N.A.T. is supported by the Stanford Genome Training Program (2T32HG000044-21). B.L. was supported by the Stanford Computational, Evolutionary, and Human Genomics Fellowship and the National Key R&D Program of China (2016YFD0400800). K.M.B. is supported by a CIHR Foundation grant (FDN-154279). Z.Z. was supported by the CEHG Fellowship, the National Science Foundation GRFP (DGE-114747) and the Stanford Genome Training Program (NIH/NHGRI T32HG000044). B.B. was supported by the Stanford Genome Training Program and Dean's Postdoctoral Fellowship. J.R.D. was supported by a Lucille P. Markey Biomedical Research 688 Stanford Graduate Fellowship. J.R.D. acknowledges the Stanford Genome Training Program 689 (NIH/NHGRI T32HG000044). C.J.P. is supported by NIST/JIMB grant no. 70NANB15H268. A.B. is supported by NIH grant no. 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, the Ontario Research Fund and the 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.

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. and R.J. contributed to sample preparation and case 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. and S.U. helped with processing the variant data. L.F., C.S., N.M.F., N.A.T., Z.Z., X.L., B.B., J.R.D. and 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. and 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

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 of 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.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41591-019-0457-8.

Supplementary information is available for this paper at https://doi.org/10.1038/s41591-019-0457-8.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to L.F. or S.B.M.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Undiagnosed Diseases Network

David R. Adams[20], Aaron Aday[20], Mercedes E. Alejandro[20], Patrick Allard[20], Euan A. Ashley[20], Mahshid S. Azamian[20], Carlos A. Bacino[20], Eva Baker[20], Ashok Balasubramanyam[20], Hayk Barseghyan[20], Gabriel F. Batzli[20], Alan H. Beggs[20], Babak Behnam[20], Hugo J. Bellen[20], Jonathan A. Bernstein[20], Gerard T. Berry[20], Anna Bican[20], David P. Bick[20], Camille L. Birch[20], Devon Bonner[20], Braden E. Boone[20], Bret L. Bostwick[20], Lauren C. Briere[20], Elly Brokamp[20], Donna M. Brown[20], Matthew Brush[20], Elizabeth A. Burke[20], Lindsay C. Burrage[20], Manish J. Butte[20], Shan Chen[20], Gary D. Clark[20], Terra R. Coakley[20], Joy D. Cogan[20], Heather A. Colley[20], Cynthia M. Cooper[20], Heidi Cope[20], William J. Craigen[20], Precilla D'Souza[20], Mariska Davids[20], Jean M. Davidson[20], Jyoti G. Dayal[20], Esteban C. Dell'Angelica[20], Shweta U. Dhar[20], Katrina M. Dipple[20], Laurel A. Donnell-Fink[20], Naghmeh Dorrani[20], Daniel C. Dorset[20], Emilie D. Douine[20], David D. Draper[20], Annika M. Dries[20], Laura Duncan[20], David J. Eckstein[20], Lisa T. Emrick[20], Christine M. Eng[20], Gregory M. Enns[20], Ascia Eskin[20], Cecilia Esteves[20], Tyra Estwick[20], Liliana Fernandez[20], Carlos Ferreira[20], Elizabeth L. Fieg[20], Paul G. Fisher[20], Brent L. Fogel[20], Noah D. Friedman[20], William A. Gahl[20], Emily Glanton[20], Rena A. Godfrey[20], Alica M. Goldman[20], David B. Goldstein[20], Sarah E. Gould[20], Jean-Philippe F. Gourdine[20], Catherine A. Groden[20], Andrea L. Gropman[20], Melissa Haendel[20], Rizwan Hamid[20], Neil A. Hanchard[20], Frances High[20], Ingrid A. Holm[20], Jason Hom[20], Ellen M. Howerton[20], Yong Huang[20], Fariha Jamal[20], Yong-hui Jiang[20], Jean M. Johnston[20], Angela L. Jones[20], Lefkothea Karaviti[20], David M. Koeller[20], Isaac S. Kohane[20], Jennefer N. Kohler[20], Donna M. Krasnewich[20], Susan Korrick[20], Mary Koziura[20], Joel B. Krier[20], Jennifer E. Kyle[20], Seema R. Lalani[20], C. Christopher Lau[20], Jozef Lazar[20], Kimberly LeBlanc[20], Brendan H. Lee[20], Hane Lee[20], Shawn E. Levy[20], Richard A. Lewis[20], Sharyn A. Lincoln[20], Sandra K. Loo[20], Joseph Loscalzo[20], Richard L. Maas[20], Ellen F. Macnamara[20], Calum A. MacRae[20], Valerie V. Maduro[20], Marta M. Majcherska[20], May Christine V. Malicdan[20], Laura A. Mamounas[20], Teri A. Manolio[20], Thomas C. Markello[20], Ronit Marom[20], Martin G. Martin[20], Julian A. Martínez-Agosto[20], Shruti Marwaha[20], Thomas May[20], Allyn McConkie-Rosell[20], Colleen E. McCormack[20], Alexa T. McCray[20], Jason D. Merker[20], Thomas O. Metz[20], Matthew Might[20], Paolo M. Moretti[20], Marie Morimoto[20], John J. Mulvihill[20], David R. Murdock[20], Jennifer L. Murphy[20], Donna M. Muzny[20], Michele E. Nehrebecky[20], Stan F. Nelson[20], J. Scott Newberry[20], John H. Newman[20], Sarah K. Nicholas[20], Donna Novacic[20], Jordan S. Orange[20], James P. Orengo[20], J. Carl Pallais[20], Christina GS. Palmer[20], Jeanette C. Papp[20], Neil H. Parker[20], Loren DM. Pena[20], John A. Phillips III[20], Jennifer E. Posey[20], John H. Postlethwait[20], Lorraine Potocki[20], Barbara N. Pusey[20], Genecee Renteria[20], Chloe M. Reuter[20], Lynette Rives[20], Amy K. Robertson[20], Lance H. Rodan[20], Jill A. Rosenfeld[20], Jacinda B. Sampson[20], Susan L. Samson[20], Kelly Schoch[20], Daryl A. Scott[20], Lisa Shakachite[20], Prashant Sharma[20], Vandana Shashi[20], Rebecca Signer[20], Edwin K. Silverman[20], Janet S. Sinsheimer[20], Kevin S. Smith[20], Rebecca C. Spillmann[20], Joan M. Stoler[20], Nicholas Stong[20], Jennifer A. Sullivan[20], David A. Sweetser[20], Queenie K.-G. Tan[20], Cynthia J. Tifft[20], Camilo Toro[20], Alyssa A. Tran[20], Tiina K. Urv[20], Eric Vilain[20], Tiphanie P. Vogel[20], Daryl M. Waggott[20], Colleen E. Wahl[20], Nicole M. Walley[20], Chris A. Walsh[20], Melissa Walker[20], Jijun Wan[20], Michael F. Wangler[20], Patricia A. Ward[20], Katrina M. Waters[20], Bobbie-Jo M. Webb-Robertson[20], Monte Westerfield[20], Matthew T. Wheeler[20], Anastasia L. Wise[20], Lynne A. Wolfe[20], Elizabeth A. Worthey[20], Shinya Yamamoto[20], John Yang[20], Yaping Yang[20], Amanda J. Yoon[20], Guoyun Yu[20], Diane B. Zastrow[20], Chunli Zhao[20] and Allison Zheng[20]

Care4Rare Canada Consortium

Kym Boycott[9], Alex MacKenzie[9], Jacek Majewski[21], Michael Brudno[22], Dennis Bulman[9] and David Dyment[9]

[20]NIH Undiagnosed Diseases Network, National Institutes of Health, Bethesda, MD, USA. [21]McGill University, Montreal, Quebec, Canada. [22]University of Toronto, Toronto, Ontario, Canada.

Methods

We sequenced 143 whole-blood samples, 94 extracted from affected individuals and 49 from unaffected family members. The 94 individuals represent a total of 80 independent diseases. Samples were collected from three different institutions, the Children's Hospital of Eastern Ontario (CHEO), the Stanford Clinical Genomics Program and the Undiagnosed Disease Network (UDN). Ethical and research approval was obtained by CHEO Research Ethics Board (REB) (REB Protocol Number 11/04E), National Human Genome Research Institute (NHGRI) Institutional Review Board (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 RNA integrity number (RIN) quality check. Globin mRNA was removed using GLOBINclear before 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, 20 pooled samples were run simultaneously on an Illumina NextSeq 500 (high-output cartridge). Pooled samples were run in 9 distinct sequencing runs: 2 runs generated 75-bp paired-end reads and 7 runs generated 150-bp paired-end reads. Output bcl files were converted to fastq files and demultiplexed using bcl2fastq v.2.15.0.4 from Illumina. Overall, we obtained around 50 million reads per sample (median 52 million ±20 million).

Reads were trimmed and adapters were removed using cutadapt v.1.11 (https://github.com/marcelm/cutadapt). Reads were then aligned to the reference human genome (hg19) with STAR v.2.4.0j (https://github.com/alexdobin/STAR/releases/tag/STAR_2.4.0j). We used gencode v.19 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 v.1.131 (http://broadinstitute.github.io/picard/). Gene-level and transcript-level quantifications were generated with RSEM v.1.2.21 (ref. [41]) (https://github.com/deweylab/RSEM/releases/tag/v1.2.21). Junction files generated by STAR were filtered: to consider a junction, a minimum of ten reads uniquely spanning were required. For faster processing of samples, we used GNU parallel[42].

Independent control cohorts for expression, splicing and ASE analyses. We used whole-blood transcriptome data of 909 samples from the DGN cohort[10] as well as 65 samples (age 70) from the PIVUS cohort[11] to serve as independent healthy controls for expression analysis and splicing, respectively. DGN samples are single-end 50-bp reads and PIVUS samples are 75-bp paired-end reads. Sequences were aligned, quantified and filtered following the same protocol used for individuals with rare disease and controls. We determined outlier ASE events at the gene level per individual by comparing our data with 620 individuals in GTEx v.7 (ref. [12]) across 48 tissues. ASE in GTEx was processed as in ref. [21], 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 2-sided Wilcoxon rank sum test on 620 individuals from GTEx v.7 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 with regard to the disease (Supplementary Table 4). We obtained OMIM gene lists ($n=3{,}766$ genes) from https://omim.org/downloads/. Gene expression of disease genes in our samples was restricted to protein-coding genes.

Genetic data. Variant data were 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 read alignment was performed using BWA-mem v.0.7.12 (ref. [43]) and variant calling was made using Genome Anaysis Tool Kit (GATK) v.3.3 (ref. [44]). For samples from the Stanford Clinical Genomics Program (CGS), variant calling was performed using GATK v.3.4. 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)
- $\frac{\text{Allele depth}}{\text{Total depth}} > 0.8$ for homozygous calls and >0.2 for each allele for heterozygous calls
- Exclude variants with Hardy–Weinberg equilibrium $P < 1\times10^{-6}$
- Exclude variants with call rate <0.80 (missing >20%)

We obtained genetic information for 112 samples (of 143; 54 from whole-genome sequencing, 58 from WES) (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 (v.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 MAF ≤ 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 CADD[35] scores using Vcfanno (v.0.2.7)[47]. We used CADD v.1.3 and gnomAD release v.2.0.2.

Ancestry inference. VCF files were processed for ancestry inference using BCFtools v.1.8 as follows. 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 level normalization. We filtered out genes for which less than 50% of samples from each origin (that is, rare-disease individuals and unaffected family members sequenced in-house, external controls) had TPM > 0.5 and/or variance equal to 0. This resulted in 14,988 genes being retained in the data set. We performed surrogate variable analysis (SVA) using the 'two-step' method on a centered and scaled matrix of $\log_{10}$-transformed ($\log_{10}(\text{TPM}+1)$) RNA-seq count data output by RSEM[41]. We did not provide any known covariates to SVA. To control for non-linearity in uncorrected gene expression data, we added regression splines for the top 2 surrogate variables (SVs) significantly associated with batch and institution ($P < 1\times10^{-30}$ 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 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 1,052 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 Fig. 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 outlier 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 data set 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 $Pr(Y_i = AllCases/X_i)$, where X_i is the gene constraint Z-score for gene i. We assessed the effect of X using logistic regression: $logit(p(X)) = \beta_0 + \beta_1 \times$. 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 $Y_i = 1$ (that is, all outliers being case outliers). A separate model was fit for each mutation class. We reported results as the log odds (±1.96 × s.e.m.) associated with each feature for each percentile. P values were calculated based on the z-statistic.

RIVER analysis. RIVER is a hierarchical Bayesian model to infer rare variants of their regulatory effects. Compared with other variant scoring methods, RIVER has the advantage of using both genomic information and transcriptome information[21]. We used GTEx v.7 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, and uses RNA status (in this case outlier status based on total gene expression levels) as the target/response variable. Rare variants here are defined as those with

MAF < 0.01 in 1000 Genomes Project phase III, with 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 within 10 kb of genes (from 10 kb before transcription start site to 10 kb after transcription end site). Overall, there was a median of two rare variants per gene for GTEx subjects and rare-disease subjects. For this analysis, we considered protein-coding and long intergenic noncoding RNAs (lincRNA) genes only. We used the following genomic annotations: Ensembl VEP[48], CADD[35], DANN[49], conservation score (Gerp[50], PhyloP[51], PhastCons[52]), CpG content, GC content, chromHMM[53] and Encode chromatin-openness track. We selected these features based on their earlier evidence of association with regulatory effects[21]. Features were aggregated over each gene and individual pair, using either max(), min() for quantitative features, or any() for categorical features. Expression outliers (response variables) were defined as those with $|Z\text{-score}| > 2$. *Z*-scores were calculated based on total gene expression level reads per kilobase of transcript, per million mapped reads (RPKM) from RNA-seq. In addition, for GTEx training data, gene expression levels were corrected by PEER[54] to remove technical artifacts and major common-variant expression quantitative trait loci (eQTL) effects were also removed. *Z*-scores for GTEx are median over all available tissues[21].

Junction coverage ratios. Reference junctions were derived from Gencode v.19 annotation files 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 sample junction 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 of the cases samples ($n = 94$). To remove possible noise, and to allow missing values imputation, we removed junctions for which there were 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. Principal component analysis was then performed using prcomp R package. We regressed out principal components accounting for 95% of the variation in our imputed dataset (176 principal components). We then put back original missing values in the data set and derived *Z*-scores used in the outlier analysis. We looked at the correlation pattern between the first ten principal components and known covariates from our data set. In brief, principal component 1 mainly separated the source of the data (UDN, CGS, CHEO or PIVUS). Principal component 2 highlighted differences between the first batch and the other batches. Overall, we observed some level of correlation between all known covariates and the principal components 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 2-sided Wilcoxon rank sum test on all samples with splicing outliers and genetic information ($n = 74$).

ASE. ASEReadCounter[55] v.3.8-0-ge9d806836 from Genome Anaysis Tool Kit[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 with an imbalanced allelic ratio (≥ 0.65 or ≤ 0.35). We eliminated total mono-allelic expression from the analysis (that is, allelic ratio = 0 or 1).

To detect ASE outliers, we restricted our analysis to sites and genes common to our samples and GTEx data set 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 individual, 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's listed HPO terms, as well as the parent and child terms. To determine whether the overlap of ASE outliers with HPO-associated genes was 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 with 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 individual for whom we had 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 Phenotype 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' the individuals for which we found candidates from RNA-seq data for which the causal mutation was found and validated. To evaluate the number of individuals for which we had strong candidates, we took a subset of 30 individuals 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 labeled as 'strong candidate'. Individuals for whom no strong candidate genes were found after analyzing RNA-seq data were labeled 'no candidate'.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

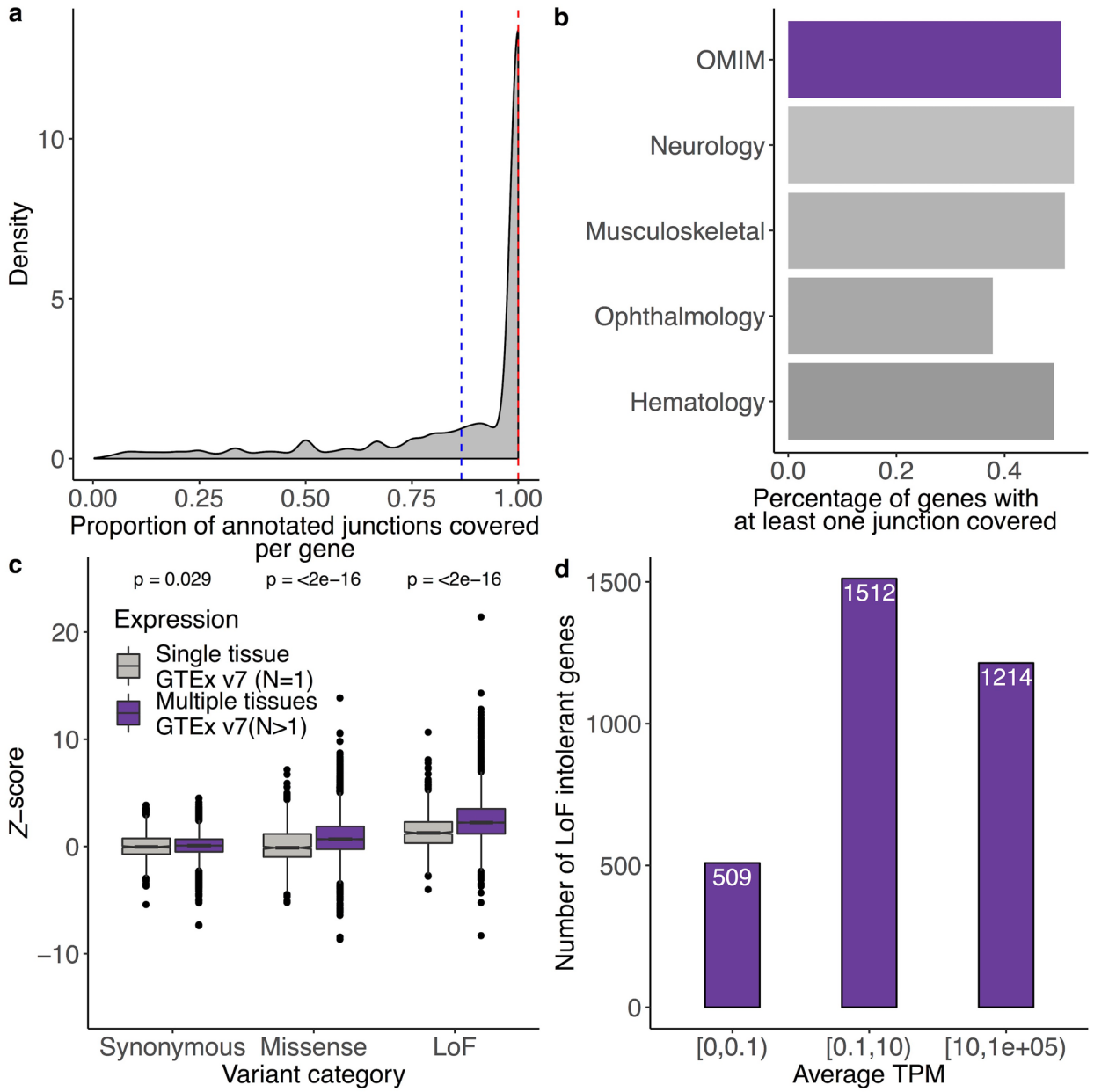
UDN data are accessible through the UDN Gateway and through dbGaP entry at phs001232.v1.p1. DGN RNA-seq data are available by application through the NIMH Center for Collaborative Genomic Studies on Mental Disorders. Instructions for requesting access to data can be found at https://www.nimhgenetics.org/access_data_biomaterial.php, and inquiries should reference the 'Depression Genes and Networks study (D. Levinson, PI)'. The GTEx Analysis v.7 release allele-specific expression data are available from dbGaP (dbGaP Accession phs000424.v7.p2). PIVUS RNA-seq data are accessible on the European Genome-Phenome Archive (EGAS00001003583). The Care4Rare data are available through Genomics4RD.

Code availability

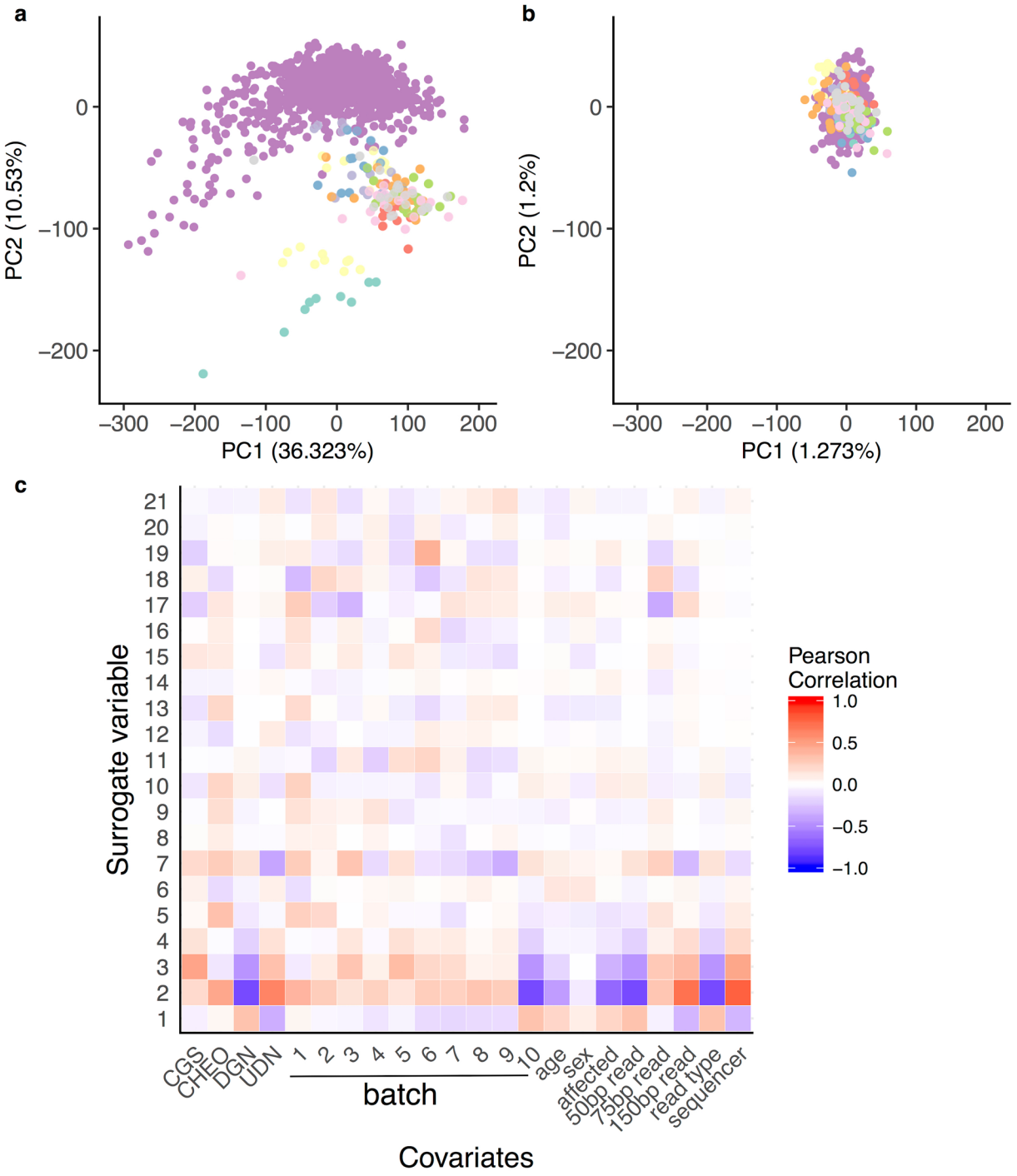
Code for running the analysis and producing the figures throughout the manuscript is available at https://github.com/lfresard/blood_rnaseq_rare_disease_paper. Our pipeline to highlight candidate variants is available at https://github.com/lfresard/blood_rnaseq_rare_disease_paper/blob/master/pipeline.md

References

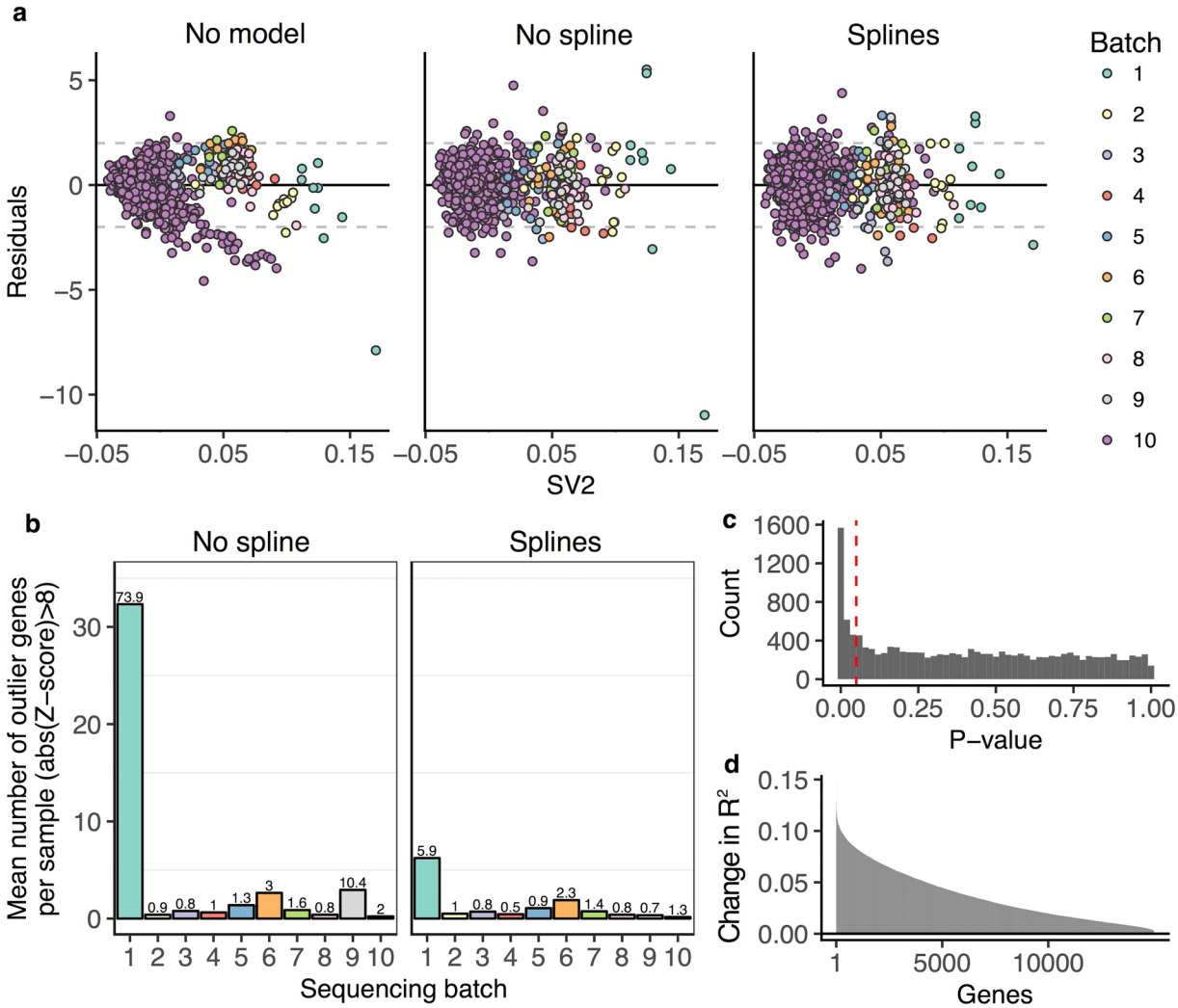
41. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
42. Tange, O. GNU Parallel - The Command-Line Power Tool. *The USENIX Magazine* **36**, 42–47 (2011).
43. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
44. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
45. Ganna, A. et al. Quantifying the impact of rare and ultra-rare coding variation across the phenotypic spectrum. *Am. J. Hum. Genet.* **102**, 1204–1211 (2018).
46. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
47. Pedersen, B. S., Layer, R. M. & Quinlan, A. R. Vcfanno: fast, flexible annotation of genetic variants. *Genome Biol.* **17**, 118 (2016).
48. McLaren, W. et al. The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (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).
50. Davydov, E. V. et al. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.* **6**, e1001025 (2010).
51. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).
52. Siepel, A. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
53. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Meth.* **9**, 215–216 (2012).
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. *Nat. Protoc.* **7**, 500–507 (2012).
55. Castel, S. E., Levy-Moonshine, A., Mohammadi, P., Banks, E. & Lappalainen, T. Tools and best practices for data processing in allelic expression analysis. *Genome Biol.* **16**, 195 (2015).
56. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).



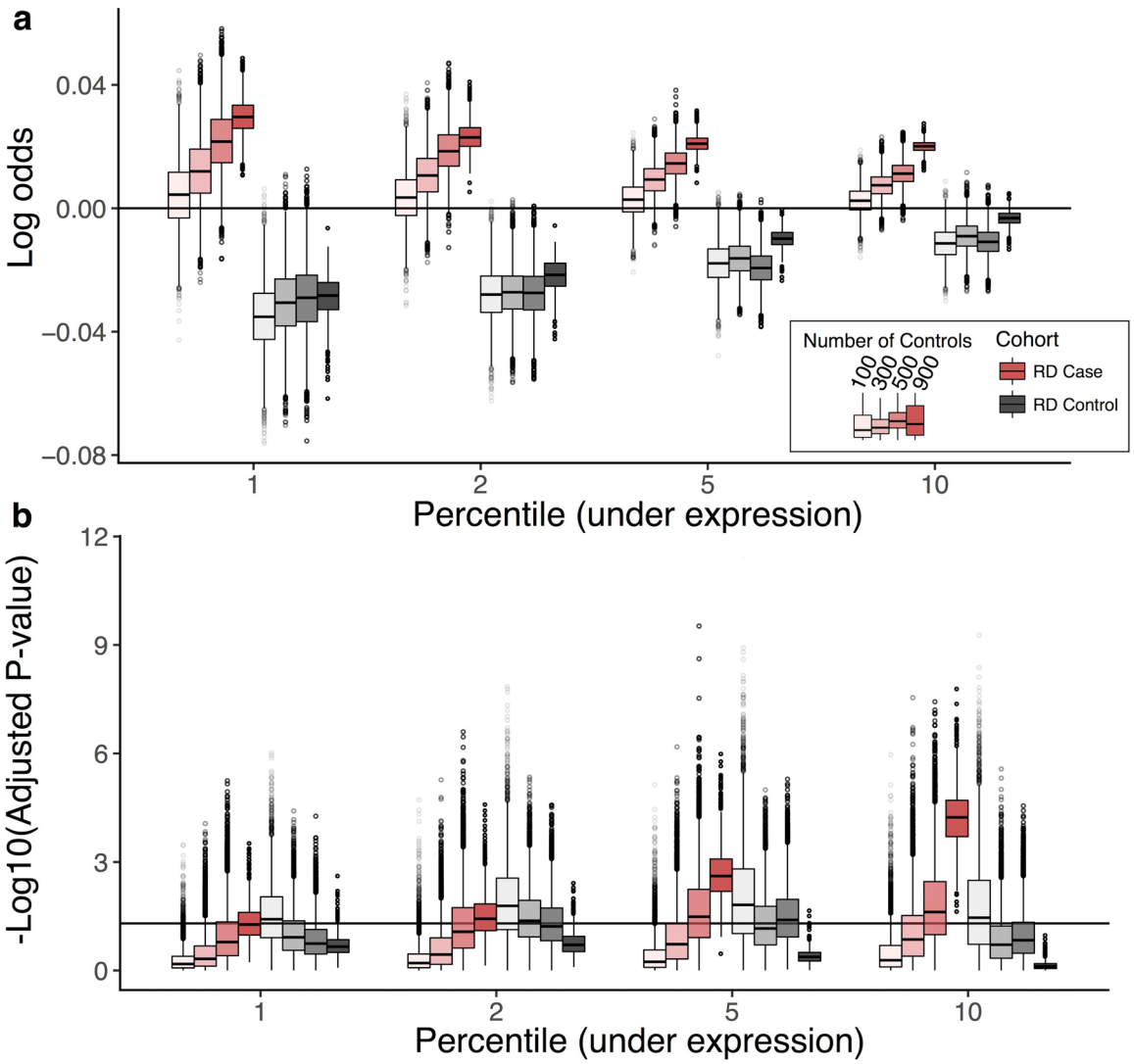
Extended Data Fig. 1 | Gene expression patterns across whole blood samples. We used a total of 1,061 whole blood samples from our controls cohorts and rare disease samples. **a**, Density plot representing the proportion of annotated junctions covered per gene. Those are a subset of genes for which at 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 $\leq 2 \times 10^{-16}$). 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× 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.



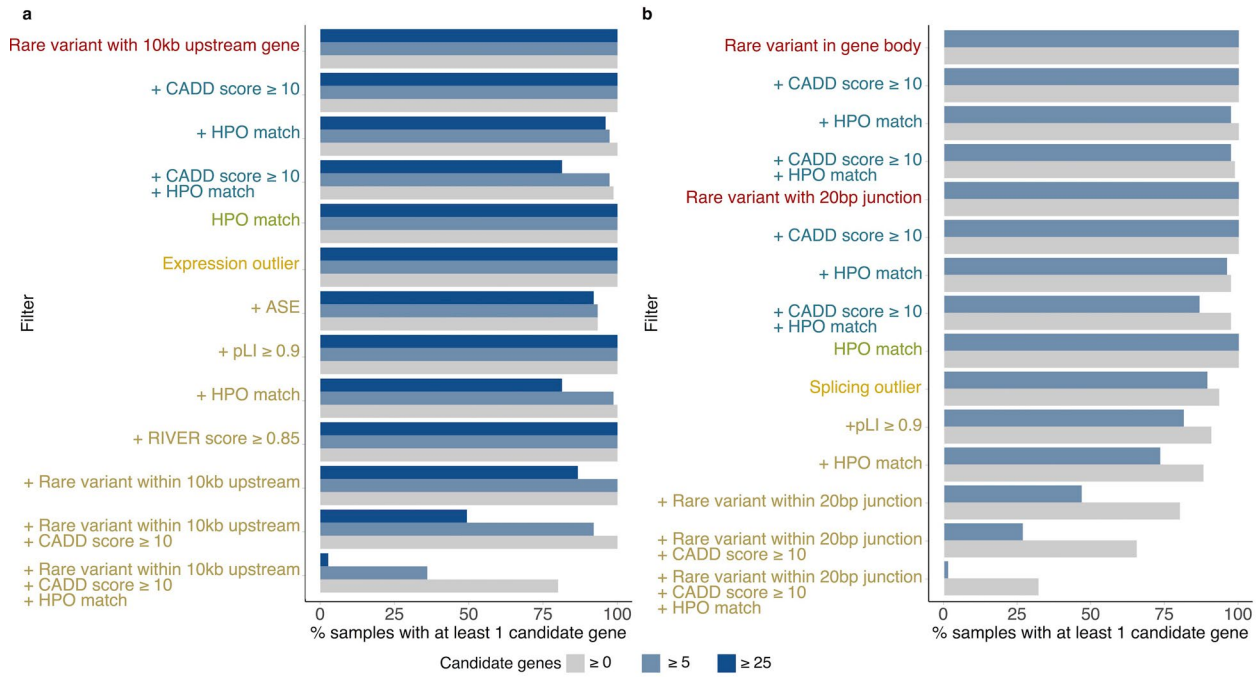
Extended Data Fig. 2 | Correction for batch effects: Expression data. Analyses performed on $n = 909$ DGN samples and 143 rare diseases (cases and family controls). **a**, Plot of first two principal components run on uncorrected gene expression data. Samples are coloured by batch. Largest cluster (green 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.



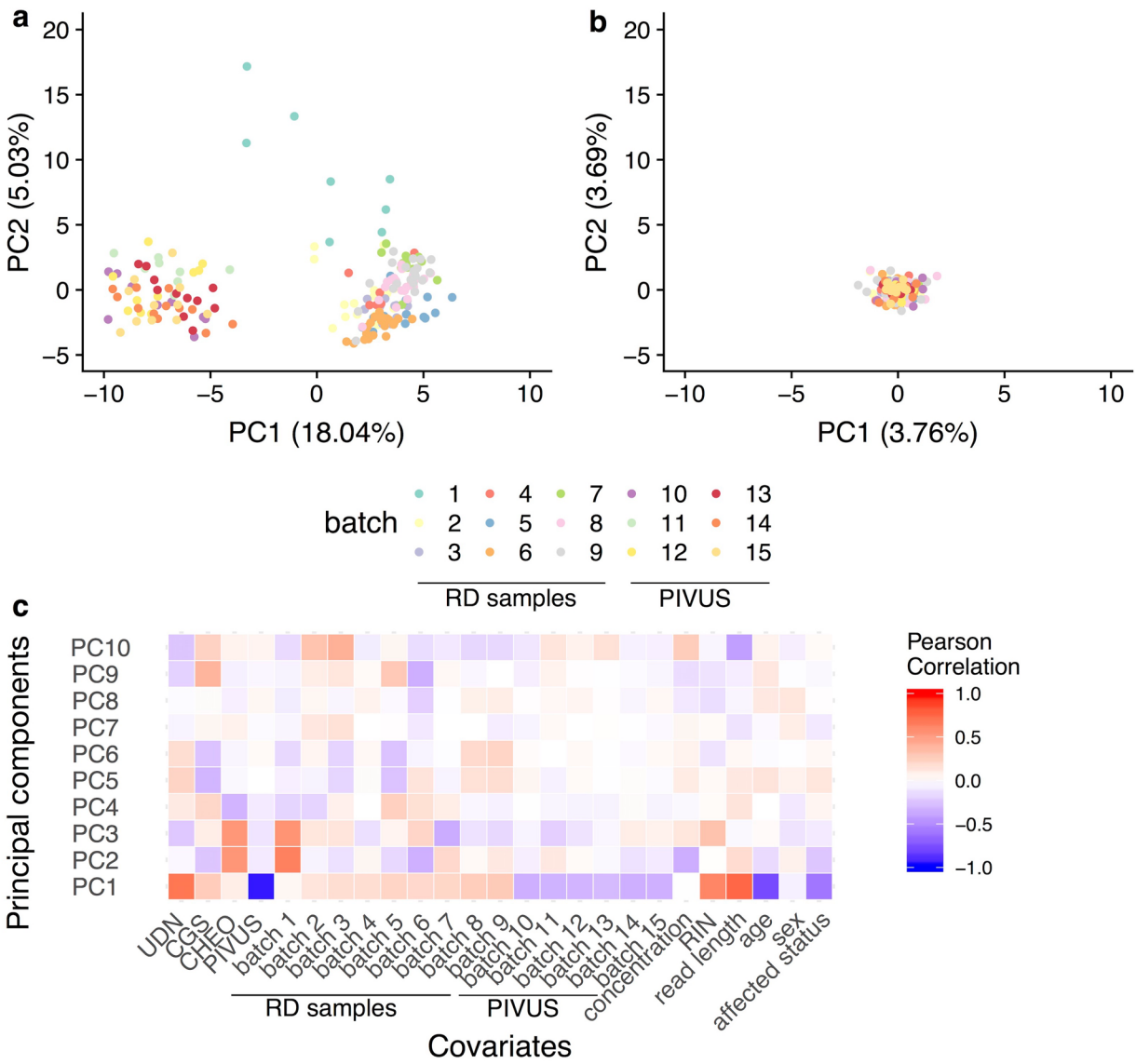
Extended Data Fig. 3 | Use of regression splines in expression data normalization. a, Normalized gene expression residuals from 1,052 samples in an example gene without correction (left panel), after regressing out significant surrogate variables (SVs) (middle panel) and significant SVs plus regression 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).



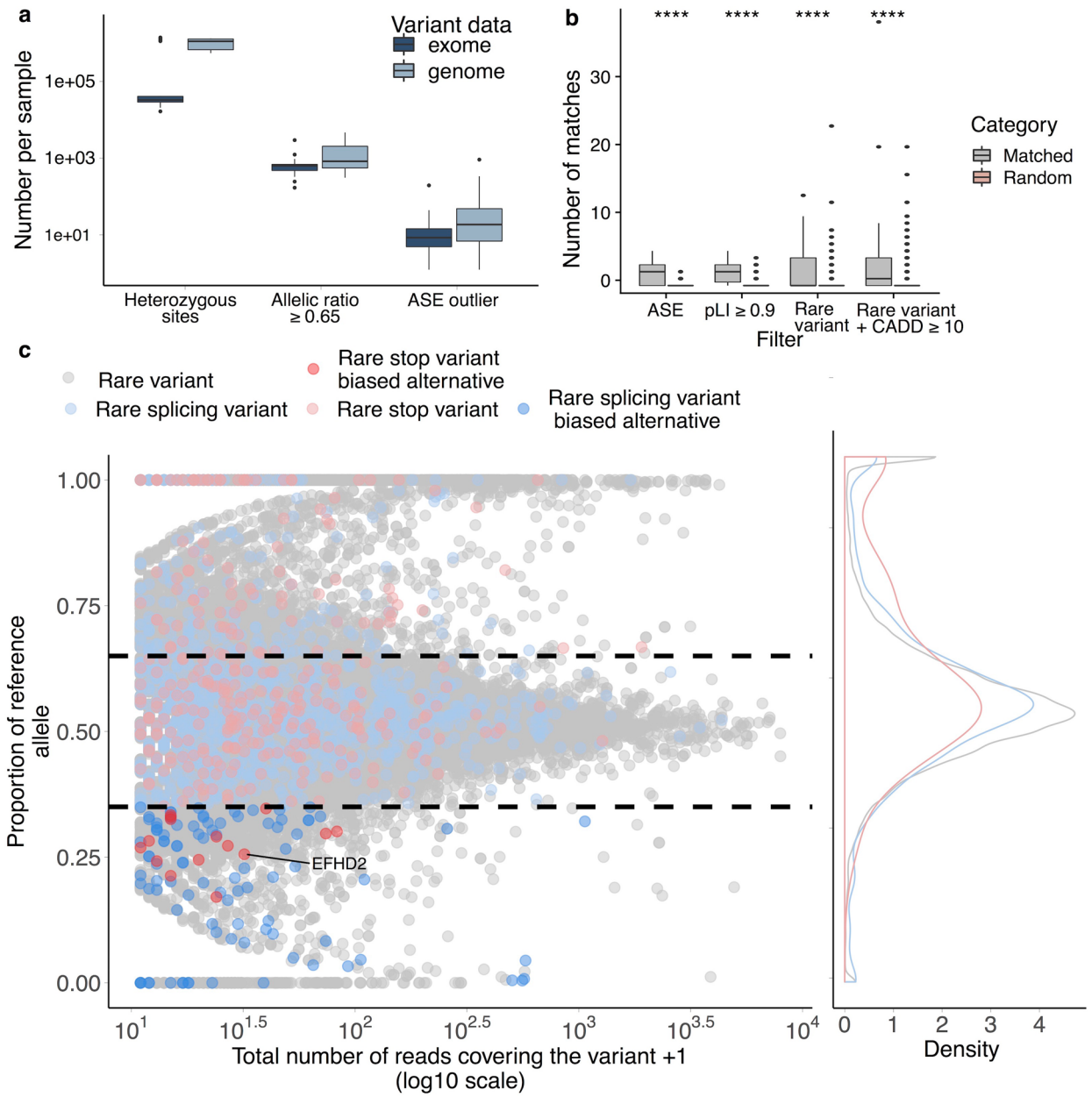
Extended Data Fig. 4 | Impact of the number of controls on loss-of-function intolerance enrichment. a, Enrichment of case (red, $n = 64$) under-expression outliers in LoF sensitive genes as we increase the number of controls (7,600 random subsets for each sample size indicated in legend). This enrichment 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× interquartile range of the hinge, respectively.



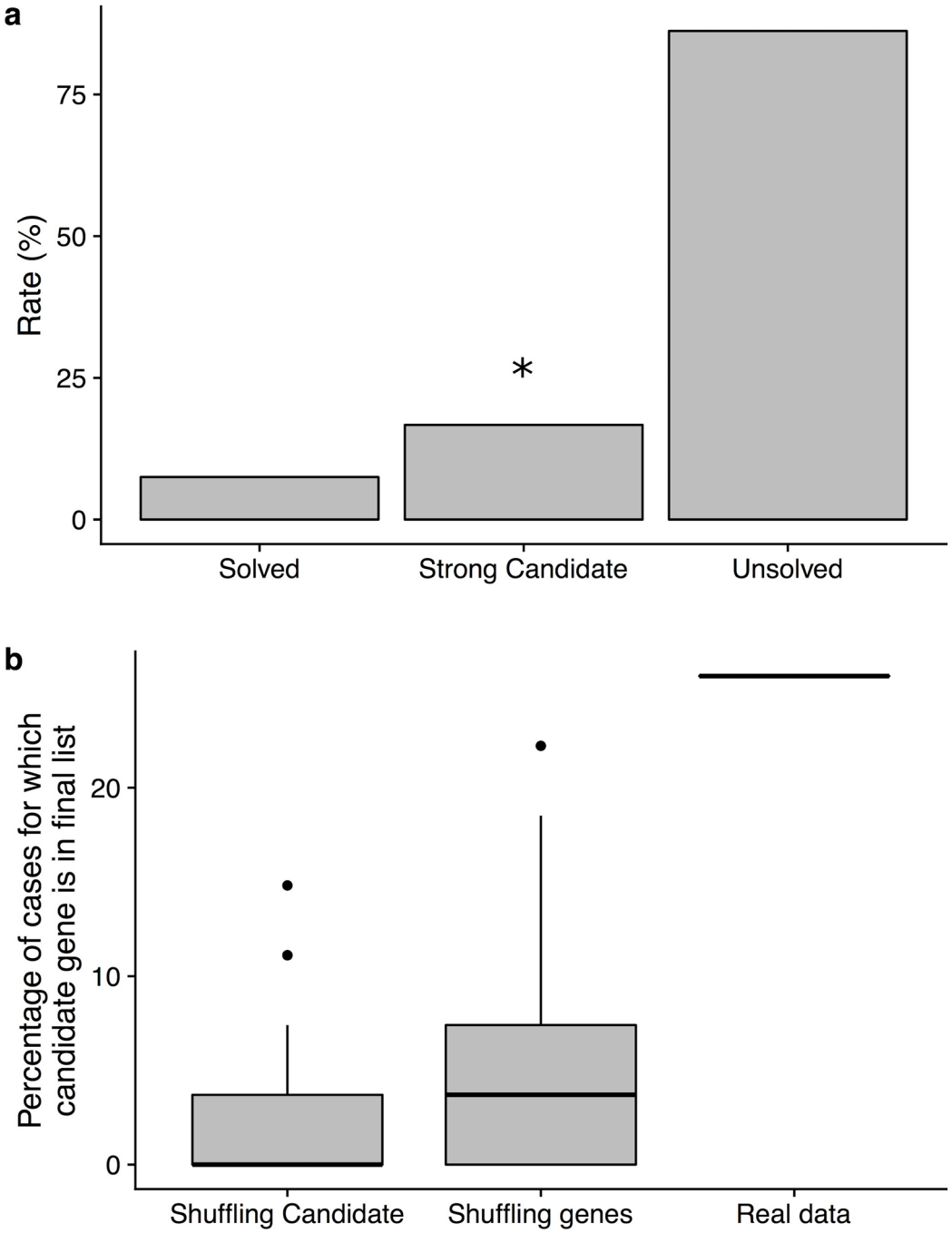
Extended Data Fig. 5 | Percentage of samples left when filtering outliers. Filters have various impacts on the number of samples with at least one candidate gene. By combining several layers of filters we are drastically reducing the number of candidate genes but also the number of samples for which 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.



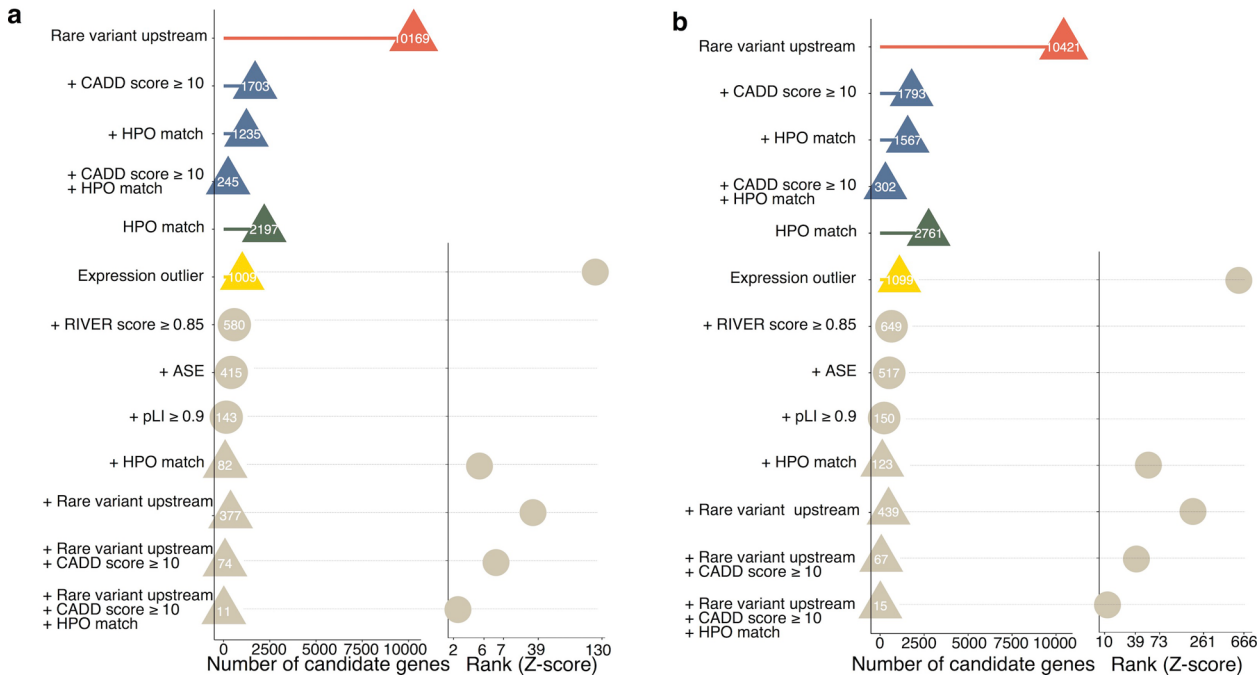
Extended Data Fig. 6 | Correction for batch effects - Splicing data. Analyses performed on 65 PIVUS samples and 143 rare disease samples. **a**, Plot of first two principal components (PCs) run on uncorrected splicing ratio data. Samples are coloured by batch. We observed that PC1 was separating PIVUS 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.



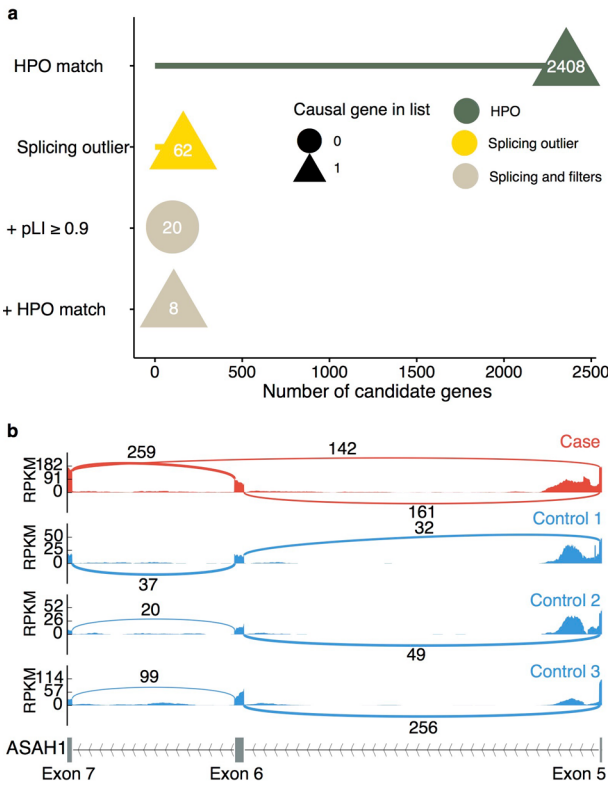
Extended Data Fig. 7 | Allele specific expression across rare disease samples. a, Prevalence of ASE events in rare diseases samples ($n=112$). Results are displayed separately for exome and genome sequencing. **b**, Difference in proportion of genes matching HPO terms for top 20 ASE outliers per case in comparison to random genes (100 random gene sets for each sample, $n=109$ samples). Analysis performed for all genes, genes with pLI ≥ 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 $\leq 1\times10^{-4}$, 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× 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× 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: ASAH1 case. a, After filtering our detected splicing outliers for genes related to the phenotype (through HPO IDs), only eight genes were left, with *ASAH1* 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 *ASAH1* 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.

nature research

Corresponding author(s): Stephen B Montgomery

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	Sequencing data was generated with illumina sequencing machines. No specialized software was developed for data collection.
Data analysis	cutadapt (v1.11); bcl2fastq (v2.15.0.4); STAR (v2.4.0j); Picard Tools MarkDuplicates (v1.131); RSEM (v1.2.21); BWA-mem (v0.7.12); GATK (v3); Vcfanno (v0.2.7); BCFtools (v1.8); BEDtools (v2.26.0-112-gd8c0fe4); Vcfanno (v0.2.7); ASEReadCounter (v3.8-0-ge9d806836). Custom code is available at https://github.com/lfresard/blood_rnaseq_rare_disease_paper

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on 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.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The sample size was based on the availability of undiagnosed cases for which we had blood data. This paper is showing that comparing those cases to large control cohorts can lead to diagnosis. We used controls sets for which we had blood transcriptome data and varied sample size to assess its impact on observed results. We investigated 94 cases across 16 diverse disease categories, using over 1500 controls. This is more samples than any publication on the subject to date.
Data exclusions	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-score| > 4 after SVA correction (54 samples). In general, samples were excluded if not meeting QC requirements specific to analyses. Sample size used are reported for each analysis.
Replication	Cases for which the diagnosis is described as validated were confirmed through independent experiments such as Sanger sequencing or RT-PCR and reflect cases for which the gene was previously known in the literature to be involved in similar phenotypes. All attempts for replication were successful.
Randomization	RNA-sequencing runs were performs based upon data availability. Family members were grouped on the same run. We corrected the data for batch effects (as described in the manuscript) and show the correlation of those batches to the sequencing run before and after correction.
Blinding	No blinding was performed in the context of our study. Phenotypes and family relationship were needed for appropriate diagnostic of the affected samples.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	94 affected individuals together with 49 unaffected family members were recruited for this study. There were 76 women and 67 men between 1 and 80 years old. Among them, there was 43 European descendant, 15 Asian-descendants, 19 Hispanics, 4 persons of multiple origins and 62 for which we don't have the information however, ancestry PCs have been inferred for all.
Recruitment	For the UDN, the recruitment is based on the following criteria from to the consortium (https://commonfund.nih.gov/diseases): - the applicant does not have a diagnosis that explains the objective findings. - the applicant or legal guardian agrees to the storage and sharing of information and biomaterials in an identified fashion amongst the UDN centers and in a de-identified fashion to research sites beyond the network. For the CGS, we requested de-identified specimens and/or subsequently generated data from participants (and, when applicable, their family members) that have been enrolled in the Stanford Clinical Genomics Program's research pilot pro

gram. In this research pilot program, participants were already consented to be involved in research and disease (Stanford IRB-approved protocol #23066) and were referred from healthcare providers at Stanford. These referring physicians have expertise in heritable disorders, and patients selected for the pilot program have conditions with suspected genetic etiology, based on the physician's evaluation of specific clinical or differential diagnosis as well as family history. We requested deidentified specimens and/or subsequently generated data for the purposes of development of RNA seq and other computational/experimental approaches aimed at improving the diagnostic yield of genomic sequencing (Stanford IRB-approved protocol # 38046). Appropriate Biobank/ laboratory staff responsible for processing the biobank specimens provided de-identified specimens/data to our research group.
The CHEO patient recruitment was performed under the Care4Rare Canada consortium protocol (http://care4rare.ca/).

We analyzed all samples received that passed our QC, independently of the disease type or the origin of the sample.