HHS Public Access

Author manuscript

Science. Author manuscript; available in PMC 2016 May 08.

Published in final edited form as:

Science. 2015 May 8; 348(6235): 666–669. doi:10.1126/science.1261877.

Impact of predicted protein-truncating genetic variants on the human transcriptome

Manuel A. Rivas[1,†], **Matti Pirinen**[2,*], **Donald F. Conrad**[3,*], **Monkol Lek**[4,5,*], **Emily K. Tsang**[6,7], **Konrad J. Karczewski**[4,5], **Julian B. Maller**[4,5], **Kimberly R. Kukurba**[6,7], **David DeLuca**[4], **Menachem Fromer**[8], **Pedro G. Ferreira**[9,10,11], **Kevin S. Smith**[6,7], **Rui Zhang**[6], **Fengmei Zhao**[4,5], **Eric Banks**[4], **Ryan Poplin**[4], **Douglas Ruderfer**[8], **Shaun M. Purcell**[4,5,8], **Taru Tukiainen**[4,5], **Eric V. Minikel**[4,5], **Peter D. Stenson**[12], **David N. Cooper**[12], **Katharine H. Huang**[4], **Timothy J. Sullivan**[4], **Jared Nedzel**[4], **the GTEx Consortium**, **the Geuvadis Consortium**, **Carlos D. Bustamante**[6], **Jin Billy Li**[6], **Mark J. Daly**[4,5], **Roderic Guigo**[13], **Peter Donnelly**[1,14], **Kristin Ardlie**[4], **Michael Sammeth**[13], **Emmanouil Dermitzakis**[9,10,11], **Mark I. McCarthy**[1,15], **Stephen B. Montgomery**[6,7], **Tuuli Lappalainen**[6,9,10,11,16,17,†], and **Daniel G. MacArthur**[4,5,18,†]

[1]Wellcome Trust Centre for Human Genetics, Nuffield Department of Clinical Medicine, University of Oxford, Oxford, UK [2]FIMM, University of Helsinki, Finland [3]Washington University in St. Louis, St. Louis, USA [4]Broad Institute of MIT and Harvard, Cambridge, MA, USA [5]Analytical and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA [6]Department of Genetics, Stanford University, CA, USA [7]Department of Pathology, Stanford University, CA, USA [8]Department of Psychiatry, Mt. Sinai Hospital, NY, USA [9]University of Geneva, Geneva, Switzerland [10]Institute for Genetics and Genomics in Geneva (iGE3), University of Geneva, Geneva, Switzerland [11]Swiss Institute of Bioinformatics, Geneva, Switzerland [12]Institute of Medical Genetics, School of Medicine, Cardiff University, Heath Park, Cardiff, United Kingdom [13]Centre for Genomic Regulation (CRG), Barcelona, Catalonia, Spain [14]Department of Statistics, University of Oxford, Oxford, UK [15]Oxford Center for Diabetes Endocrinology and Metabolism, University of Oxford, Oxford, UK [16]New York Genome Center, NY, USA [17]Department of Systems Biology, Columbia University, NY, USA [18]Department of Medicine, Harvard Medical School, Boston, MA, USA

Abstract

Accurate prediction of the functional impact of genetic variation is critical for clinical genome interpretation. We systematically characterized the transcriptome effects of protein-truncating variants (PTVs), a class of variants expected to have profound impacts on gene function, using

[†]To whom correspondence should be addressed rivas@well.ox.ac.uk, tlappalainen@nygenome.org, macarthur@atgu.mgh.harvard.edu.

[*]Contributed equally to this work

Supplementary Materials:

Materials and Methods
Figures S1-S42
Tables S1-S7
References (*28–56*)
Data Files S1

data from the Genotype-Tissue Expression (GTEx) and Geuvadis projects. We quantitate tissue-specific and positional effects on nonsense-mediated transcript decay, and present an improved predictive model for this decay. We directly measure the impact of variants both proximal and distal to splice junctions. Furthermore, we find that robustness to heterozygous gene inactivation is not due to dosage compensation. Our results illustrate the value of transcriptome data in the functional interpretation of genetic variants.

Genetic variants predicted to shorten the coding sequence of genes – termed protein-truncating variants (PTVs) – are typically expected to have large effects on gene function. These variants are enriched for disease-causing mutations (1, 2), but some may be protective against disease (3). However, PTVs are abundant in the genomes of healthy individuals (4), indicating that they often do not have major phenotypic consequences. In addition, while PTVs are often described as loss-of-function (LOF) variants, in most cases their precise molecular impact has not been characterized, and in other cases show gain-of-function effects (1). Clinical interpretation of PTVs will thus require direct characterization of their biochemical effects.

We catalogue predicted PTVs and their transcriptomic impact in 462 healthy individuals with DNA and mRNA sequencing (RNA-seq) from lymphoblastoid cell lines (LCLs) in the Geuvadis study (5, 6), and 173 individuals with exome sequencing and RNA-seq from a total of 1,634 samples from multiple tissues in the Genotype-Tissue Expression (GTEx) study (S1, 7, 8). Each GTEx individual has RNA-seq data from 1–30 tissues, with 9 tissues having >80 samples. We defined PTVs (4, Table S1) as single nucleotide variants (SNVs) predicted to introduce a premature stop codon or to disrupt a splice site, small insertions or deletions (indels) predicted to disrupt a transcript's reading frame, and larger deletions that remove the full protein coding sequence (CDS) (S2, Figs. 1, S1, S2). We identified 13,182 candidate PTVs using Phase 1 data of the 1000 Genomes Project (9) of the 421 individuals included in the Geuvadis RNA-seq project, as well as 4,584 candidate PTVs in the GTEx data, for a combined total of 16,286 candidate variants (Table S2).

We measured total gene expression levels in reads per kilobase of exon per million mapped reads (RPKM), allele-specific expression (ASE) detecting different expression levels of two haplotypes of an individual, and split mappings across annotated exon junctions to quantify splicing (S3, S4). Transcripts containing common PTVs are more weakly expressed and more tissue-specific than transcripts that do not contain common PTVs (S5, Figs. S3–7), consistent with previous work (4).

PTVs that generate premature stop codons may trigger nonsense-mediated decay (NMD). Such variants are often recessive and may protect against detrimental phenotypic effects but also may cause disease via haploinsufficiency (1). Variants that escape NMD may create a truncated protein with dominant-negative or gain-of-function effects (1). We compared transcript levels between the PTV and the non-PTV alleles within the same individual (S6, 4, 5, 10) for a total of 1,814 PTVs (S6, Figs. S8–12, Table S3) and validated the allelic ratios obtained from RNA-seq data (Figs. S13–18, Table S4, 11). We also generated a method to assess the ASE effect of frameshift indels (S6, Figs. S8–12) which were not previously examined (5, 10) due to the technical challenges of mapping bias (12–14).

Allelic count data were analyzed with a Bayesian statistical method to address whether a variant exhibits ASE in a given tissue and whether this signal is shared across multiple tissues of the same individual (S7, Figs. S19–26, 15). We observe a higher proportion of strong or moderate allelic imbalance in rare and singleton nonsense SNVs compared to common nonsense variants (54.3%, 55.4%, and 35.7%, respectively), suggesting that rare PTVs are more likely to trigger NMD (Fig. S19).

Rare nonsense SNVs predicted to trigger NMD according to the 50bp rule (S7, 16) have a larger proportion of ASE than SNVs that escape NMD (69.5% vs 31.9% respectively), and both classes demonstrate ASE more often than synonymous variants (7.9%, $P < 0.001$ across all comparisons, two-proportion z-test, Fig. 2A). A higher proportion of ASE is also observed for frameshift indels predicted to trigger NMD (52.1%) compared to those predicted to escape NMD (30.6%) and at higher levels than that predicted for in-frame indels (18.4%, Fig. 2B). Testing alternative simple distance rules showed that the 50bp rule has the highest predictive value (Fig. 2C).

We next generated an improved predictive model for no ASE versus strong/moderate ASE for all nonsense SNVs (S7). Our model predicts NMD better than the 50bp rule, with an Area Under the Curve (AUC) = 80.8% (95% CI 77.3–84.4%) compared to 50bp rule AUC = 72.9% (69.3–76.5%) (Figs. 2C, S21, S22). Our results provide a quantitative estimate of the value of NMD predictions, and illustrate that the 50bp rule (16) remains a valuable heuristic. Nonetheless, our model improves NMD prediction, allows a more flexible analysis of the probability that a variant will trigger NMD from variant data (Fig. S21) and provides data for understanding the molecular mechanisms of NMD (Fig. S22).

The GTEx study design allows us to study variation in NMD across tissues. We applied a Bayesian hierarchical model (S7, 15) to rare nonsense variants predicted to trigger NMD, according to the 50bp rule, with ASE data from at least two tissues. We estimate that 30.5% of these nonsense variants have no ASE in any tissue, and 48.3% and 3.3% have moderate or strong ASE across all tissues, respectively. Finally, 17.9% have heterogeneous effects across tissues, and 8.1% of ASE effects are specific to a single tissue (Figs. 2D-F, S23–26). The tissue-specificity of NMD implies that the same protein-truncating variant may have different effects across tissues, which could contribute to tissue-specific effects of disease-causing mutations (17).

We examined if heterozygous carriers of PTVs exhibit compensatory up-regulation of the functional allele, which could contribute to tolerance of PTVs and partially explain the widespread haplosufficiency of human genes (18). Dosage compensation has been reported to correlate with gene expression levels (19) and occur in over 80% of deleted genes in *Drosophila melanogaster* (20). To minimize the impact of genotyping error we focused only on biallelic whole-gene deletions with strong experimental support and manual curation (S2, Figs. S27–29). We first analyzed the few examples of common whole-gene deletion polymorphisms (S8). For 5/6 of these genes an additive model relating gene expression to gene copy number provided a better fit than a dominant model, providing no evidence for dosage compensation (Table S6). Additionally, heterozygous carriers of rare deletions also had consistently decreased expression of the respective gene compared to the population

median (P =1.37 × 10^{-5}, one-sided binomial test of 11 rare PTV deletions in 25 genes, Figs. S27–28). Similar results were obtained for 53 nonsense PTVs with strong ASE signals (P = 2.90 × 10^{-9}, one-sided binomial test; Figs. S30–31). These results suggest that full dosage compensation is rare for human genes.

Disruption of splicing can result in changes in protein structure either via in-frame changes in exon structure or by introducing a premature stop codon (21). Splicing variant annotation tools typically focus only on the two bases at either end of a spliced intron, “essential splice sites” (22) despite the fact that more distant sites are also known to affect splicing (21, 23, 24).

Variation around splice junctions tends to be rare (MAF ≤ 0.01). We standardized the population distribution of each splice-junction quantification per tissue, and grouped variants by their distance from their respective donor and acceptor sites. We then analyzed if individuals carrying variants in these positions differ from the population in the quantification of the splice junction and the proximal exon and intron (Fig. 3A-D).

In the Geuvadis data set, up to 79% of variants in the four essential splice site loci cause splice disruptions (P < 0.01; Fig. 3A; GTEx results Figs. S33–37). We also find evidence of splice disruption from variants outside these regions, especially at position 1–5bp of intronic donor splice sites, 1bp into the adjacent exon, and also more distally −including the −24 position from the acceptor site, which likely reflects the branch-point position required for pre-mRNA splicing (25).

These patterns are consistent with other estimates of functional effects (Fig. 3E), depletion of common variants in exome sequencing data sets (Fig. 3F, 26), and a higher prevalence of disease-causing mutations (Fig. 3G). Analyses of common variants did not capture these patterns of enrichment (Table S7, Figs. S38–40, S9). Our posterior probability estimates for sites with significant alternative distributions (P < 0.05) provide a resource for analyses (Figs. S41–42, S9–10).

By drawing on data from a wide range of adult tissues across 635 individuals we provide a systematic assessment of the impact of predicted protein-truncating variants on the human transcriptome. Furthermore, this study indicates that nonsense-mediated decay has heterogeneous effects across tissues and how to better detect splice-disrupting variants outside the “essential” sites at the splice junction.

We find no evidence for widespread dosage compensation maintaining normal expression levels of genes affected by heterozygous PTVs. This, together with the fact that most human genes are haplosufficient (18), suggests that homeostatic mechanisms at the cellular level, possibly as proposed in the theory of dominance (27), maintain biological function in the face of heterozygous, or even homozygous (4), inactivation of human genes.

The resource made available with this study provides a starting point for cataloging variants affecting protein function, but larger data sets will be required to increase our power to predict molecular consequences of variants from sequence data alone. These results highlight the benefits of direct RNA sequencing of either patient tissue or genetically

engineered cell lines for interpretation of genetic variation, and suggest that personal transcriptomics will become an important complement to genome analysis.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to thank all the members of the GTEx and Geuvadis consortia and L. Solomon for assistance with the figures. This work was supported by the National Institutes of Health [NIGMS R01GM104371 to DGM; NIMH R01MH101814 and R01MH090941 to MIM, U01HG007593 to JBL and SBM, and R01MH101810 to DFC]; Academy of Finland [257654 to MP]; a Hewlett-Packard Stanford Graduate Fellowship and a doctoral fellowship from the Natural Science and Engineering Research Council of Canada to EKT; a National Defense Science & Engineering Graduate Fellowship (NDSEG) from the United States Department of Defense (DoD) to KRK; European Research Council, Swiss National Science Foundation, and Louis-Jeantet Foundation to ETD; Wellcome Trust [095552/Z/11/Z and 090532/Z/09/Z to PD, and 098381 to MIM]; and a Clarendon Scholarship, NDM Studentship, and Green Templeton College Award from University of Oxford to MAR. The Genotype-Tissue Expression (GTEx) project was supported by the Common Fund of the Office of the Director of the National Institutes of Health. Additional funds were enrolled at Biospecimen Source sites funded by NCI\SAIC-Frederick, Inc. (SAIC-F) subcontracts to the National Disease Research Interchange (10XS170), Roswell Park Cancer Institute (10XS171), and Science Care, Inc (X10S172). The Laboratory, Data Analysis, and Coordinating Center (LDACC) was funded through a contract (HHSN268201000029C) to the Broad Institute, Inc. Biorepository operations were funded through an SAIC-F subcontract to Van Andel Institute (10ST1035). Additional data repository and project management were provided by SAIC-F (HHSN261200800001E). The Brain Bank was supported by a supplement to University of Miami grant DA006227. Statistical Methods development grants were made to the University of Geneva (MH090941), the University of Chicago (MH090951 & MH090937), the University of North Carolina – Chapel Hill (MH090936) and to Harvard University (MH090948). The primary and processed data used to generate the analyses presented here are available in the following locations: all primary sequence and clinical data files, and any other protected data, are deposited in and available from dbGaP (http://www.ncbi.nlm.nih.gov/gap) (phs000424.v3.p1, except for whole exome sequencing data in phs000424.v5.p1 and mmPCR-seq data and processed ASE data in phs000424.v6.p1); derived analysis files are available on the GTEx Portal (www.gtexportal.org). Biospecimens remaining from the study may be requested for research studies. The sample request form, biospecimen access policy, and material transfer agreement (MTA) are available on the GTEx Portal (http://www.gtexportal.org/home/samplesPage). The Geuvadis data is available in ArrayExpress accession E-GEUV-1. Further details and links to data and software are available in http://www.well.ox.ac.uk/~rivas/ptv2015/. Dr. Bustamante is a paid member of the Scientific Advisory Boards of Personalis, InVitae, and Ancestry.com; he is founder and chair of the SAB of Identify Genomics, LLC; he also owns stock options in Personalis, Invitae and Identify Genomics, LLC.

References and Notes

1. Holbrook JA, Neu-Yilik G, Hentze MW, Kulozik AE. Nature Genetics. 2004; 36:801. [PubMed: 15284851]
2. Stenson PD, et al. Human Genetics. 2014; 133:1. [PubMed: 24077912]
3. Cohen JC, Boerwinkle E, Mosley TH Jr, Hobbs HH. New England Journal of Medicine. 2006; 354:1264. [PubMed: 16554528]
4. MacArthur DG, et al. Science. 2012; 335:823. [PubMed: 22344438]
5. Lappalainen T, et al. Nature. 2013; 501:506. [PubMed: 24037378]
6. AC't Hoen P, et al. Nature Biotechnology. 2013
7. Lonsdale J, et al. Nature Genetics. 2013; 45:580. [PubMed: 23715323]
8. The GTEx Consortium. Submitted. 2014
9. The 1000 Genomes Consortium. Nature. 2012; 491:56. [PubMed: 23128226]
10. Kukurba KR, et al. PLoS Genetics. 2014; 10:e1004304. [PubMed: 24786518]
11. Zhang R, et al. Nature Methods. 2014; 11:51. [PubMed: 24270603]
12. Montgomery SB, et al. Genome Research. 2013; 23:749. [PubMed: 23478400]

13. Skelly DA, Johansson M, Madeoy J, Wakefield J, Akey JM. Genome Research. 2011; 21:1728. [PubMed: 21873452]
14. Stevenson KR, Coolon JD, Wittkopp PJ. BMC Genomics. 2013; 14:536. [PubMed: 23919664]
15. Pirinen M, et al. bioRxiv. 2014
16. Nagy E, Maquat LE. Trends in Biochemical Sciences. 1998; 23:198. [PubMed: 9644970]
17. Bateman JF, Freddi S, Nattrass G, Savarirayan R. Human Molecular Genetics. 2003; 12:217. [PubMed: 12554676]
18. Huang N, et al. PLoS Genetics. 2010; 6:10.
19. McAnally AA, Yampolsky LY. Genome Biology and Evolution. 2010; 2:44–52. [PubMed: 20333221]
20. Zhou J, Lemos B, Dopman EB, Hartl DL. Genome Biology and Evolution. 2011; 3:1014. [PubMed: 21979154]
21. Faustino NA, Cooper TA. Genes & Development. 2003; 17:419. [PubMed: 12600935]
22. McCarthy DJ, et al. Genome Medicine. 2014; 6:26. [PubMed: 24944579]
23. Burge CB, Tuschl T, Sharp PA. Cold Spring Harbor Monograph Archive. 1999; 37:525.
24. Xiong HY, et al. Science. 2014
25. Corvelo A, Hallegger M, Smith CW, Eyras E. PLoS Computational Biology. 2010; 6:e1001016. [PubMed: 21124863]
26. Purcell SM, et al. Nature. 2014
27. Kacser H, Burns JA. Genetics. 1981; 97:639–666. [PubMed: 7297851]

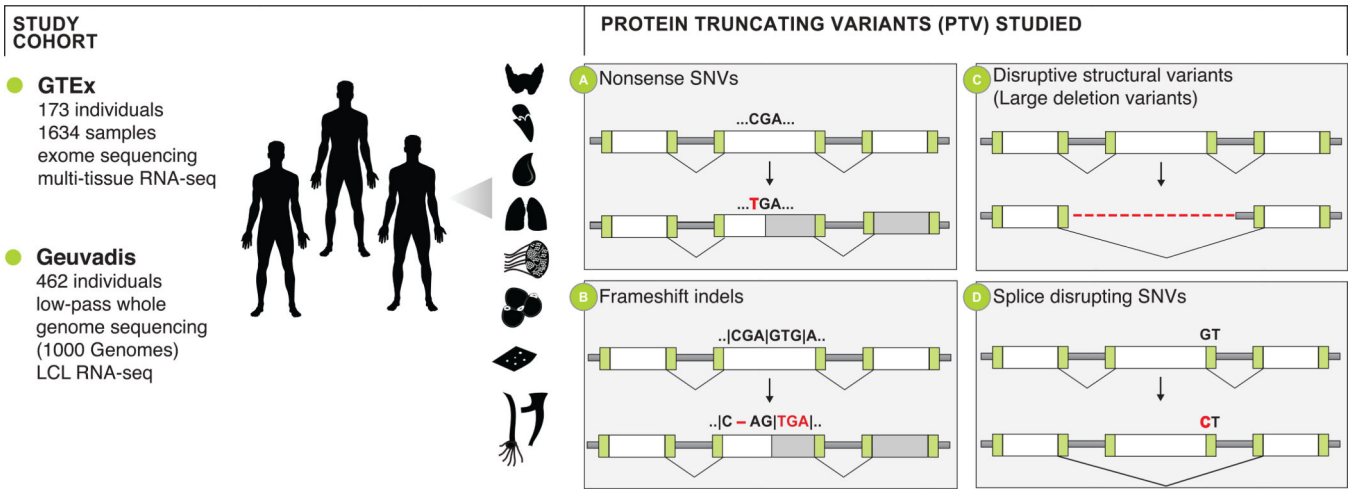

Fig. 1.

Schematic overview of the study. We prepared an integrated DNA and RNA sequencing data set by combining the pilot phase of the GTEx project of 173 individuals with up to 30 tissues per individual (total = 1634 samples) and the Geuvadis project of lymphoblastoid cell line (LCL) DNA and RNA sequencing in 462 individuals. From these data, we analyzed the effect of predicted proteintruncating genetic variants on the human transcriptome, including: a) nonsense single nucleotide variants (SNVs); b) frameshift indels; c) large deletion variants; and d) splice-disrupting SNVs.

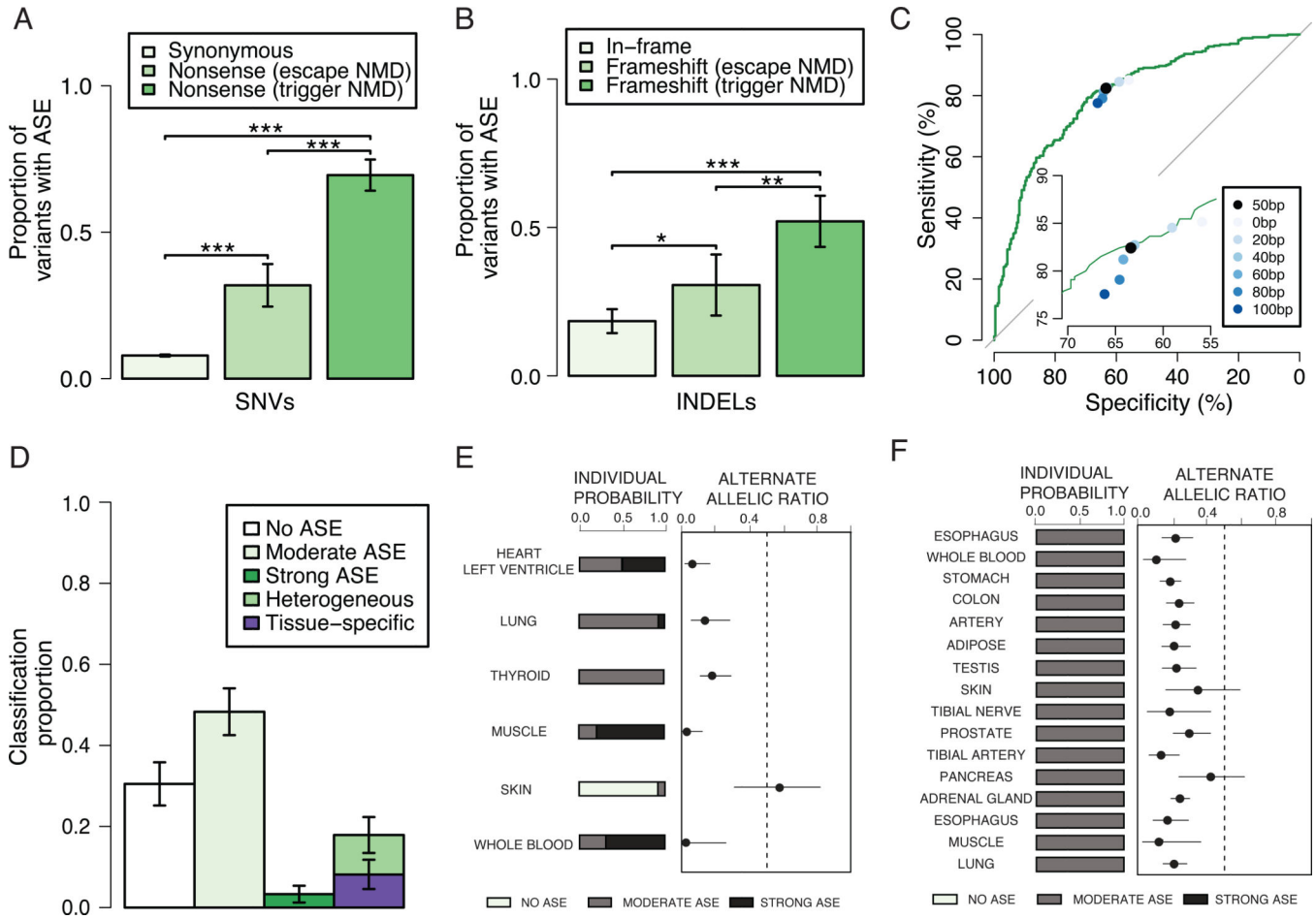

Fig. 2.

Allele-specific expression analysis. **A**. Proportion of rare SNVs with allele-specific expression (ASE) for synonymous variants (n = 25,233) and nonsense variants predicted to escape (n = 158) or trigger (n = 287) nonsense-mediated decay (NMD). **B**. Proportion of rare indels with ASE for inframe (n = 355) and frameshift indel variants predicted to escape (n = 77) or trigger (n = 129) NMD. Due to different quality filters, the proportions are not directly comparable to those in panel A. **C**. ROC curve for predicting NMD with binary classification defined as no ASE (= escape) and moderate, strong, or heterogeneous ASE (= trigger). The filled circles show the specificity and sensitivity for NMD prediction with alternative simple distance rules (inset). **D**. Multi-tissue ASE classification for rare nonsense variants predicted to trigger NMD (n = 287). **E**. Example of ASE data across 6 tissues for a heterozygous carrier of the nonsense variant rs149244943 in gene *PHKB* (phosphorylase kinase, beta) classified as having heterogeneous ASE effects across the seven tissues. We confirmed that this effect is not driven by a common tissue-specific eQTL. **F**. Example of ASE data across 16 tissues for a heterozygous carrier of the nonsense variant rs119455955, a disease mutation for recessive late-infantile neuronal ceroid lipofuscinosis in gene *TPP1* (tripeptidyl peptidase I), classified as having moderate ASE across all tissues. For all plots 95% confidence intervals are shown.

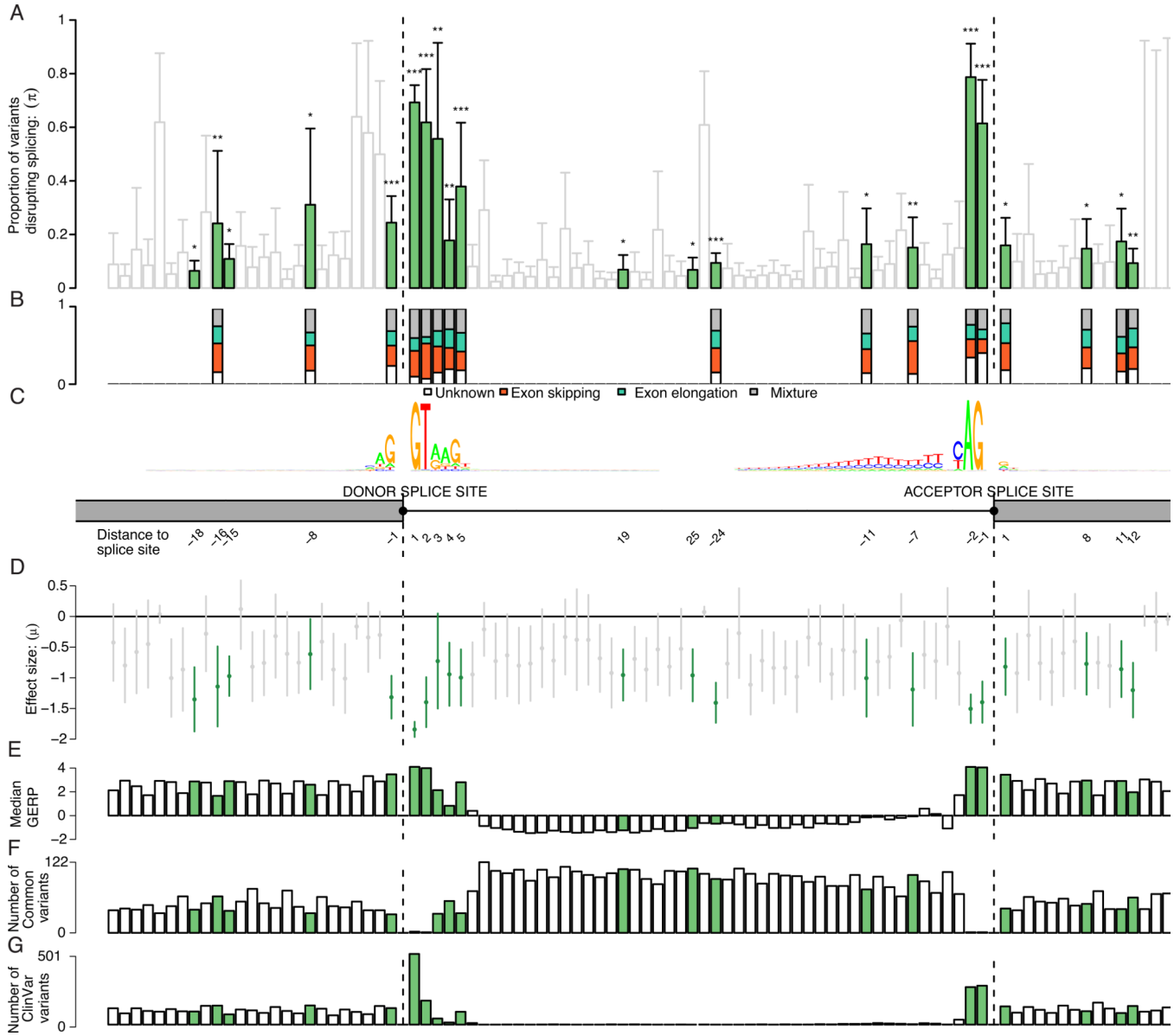

Fig. 3.

Splicing disruption. **A**. Proportion of variants disrupting splicing at each distance +/− 25bp from donor and acceptor site, (* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$; green for $P < 0.05$; upper limit of 95% CI is shown; P value evaluated using the estimated proportion of variants supporting the alternative distribution x the effect size of the alternative distribution). **B**. Classification of splice disruption events: exon skipping (low exon quantification value, no impact on intron quantification), exon elongation (high intron quantification value, no impact on exon quantification), and mixture (high intron and low exon quantification values). **C**. Diagram of donor and acceptor splice junctions and sequence logo of represented sequences. **D**. Effect size estimates (in standard deviations from the population distribution; 95% CI is shown) of the variants on splice junction quantification value. **E**. Median GERP of all variants **F**. Distribution of common variants identified in an independent exome

sequencing study of 4,500 Swedish individuals. **G**. Distribution of reported disease-causing variants in ClinVar.