HUMAN GENOMICS

Effect 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,[6,7,8] Konrad J. Karczewski,[4,5] Julian B. Maller,[4,5] Kimberly R. Kukurba,[6,7] David S. DeLuca,[4] Menachem Fromer,[4,5,9] Pedro G. Ferreira,[10,11,12] Kevin S. Smith,[6,7] Rui Zhang,[6] Fengmei Zhao,[4,5] Eric Banks,[4] Ryan Poplin,[4] Douglas M. Ruderfer,[9,13] Shaun M. Purcell,[4,5,9,13] Taru Tukiainen,[4,5] Eric V. Minikel,[4,5] Peter D. Stenson,[14] David N. Cooper,[14] 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,[15] Peter Donnelly,[1,16] Kristin Ardlie,[4] Michael Sammeth,[15,17] Emmanouil T. Dermitzakis,[10,11,12] Mark I. McCarthy,[1,18] Stephen B. Montgomery,[6,7] Tuuli Lappalainen,[6,10,11,12,19,20]*† Daniel G. MacArthur[4,5,21]*†

Accurate prediction of the functional effect of genetic variation is critical for clinical genome interpretation. We systematically characterized the transcriptome effects of protein-truncating variants, a class of variants expected to have profound effects on gene function, using data from the Genotype-Tissue Expression (GTEx) and Geuvadis projects. We quantitated tissue-specific and positional effects on nonsense-mediated transcript decay and present an improved predictive model for this decay. We directly measured the effect of variants both proximal and distal to splice junctions. Furthermore, we found 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, although PTVs are often described as loss-of-function (LOF) variants, in most cases their precise molecular effect 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 cataloged predicted PTVs and their transcriptomic effect 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 1634 samples from multiple tissues in the Genotype-Tissue Expression (GTEx) study [supplementary materials section S1 (SM S1)] (7, 8). Each GTEx individual has RNA-seq data from 1 to 30 tissues, with 9 tissues having >80 samples. We defined PTVs (4) (table S1) as single-nucleotide

[1]Wellcome Trust Centre for Human Genetics, Nuffield Department of Clinical Medicine, University of Oxford, Oxford, UK. [2]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland. [3]Washington University in St. Louis, St. Louis, MO, 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, Stanford, CA, USA. [7]Department of Pathology, Stanford University, Stanford, CA, USA. [8]Biomedical Informatics Program, Stanford University, Stanford, CA, USA. [9]Department of Psychiatry, Mt. Sinai Hospital, NY, USA. [10]Department of Genetic Medicine and Development, University of Geneva, Geneva, Switzerland. [11]Institute for Genetics and Genomics in Geneva (iGE3), University of Geneva, Geneva, Switzerland. [12]Swiss Institute of Bioinformatics, Geneva, Switzerland. [13]Division of Psychiatric Genomics, Department of Psychiatry, Icahn School of Medicine at Mount Sinai, NY, USA. [14]Institute of Medical Genetics, School of Medicine, Cardiff University, Heath Park, Cardiff, UK. [15]Center for Genomic Regulation (CRG), Universitat Pompeu Fabra (UPF), Barcelona, Catalonia, Spain. [16]Department of Statistics, University of Oxford, Oxford, UK. [17]National Institute for Scientific Computing (LNCC), Petropolis, Rio de Janeiro, Brazil. [18]Oxford Center for Diabetes Endocrinology and Metabolism, University of Oxford, Oxford, UK. [19]New York Genome Center, New York, NY, USA. [20]Department of Systems Biology, Columbia University, New York, NY, USA. [21]Department of Medicine, Harvard Medical School, Boston, MA, USA.
*Corresponding author. E-mail: rivas@well.ox.ac.uk (M.A.R.); tlappalainen@nygenome.org (T.L.); macarthur@atgu.mgh.harvard.edu (D.G.M.) †These authors contributed equally to this work. ‡These authors contributed equally to this work.
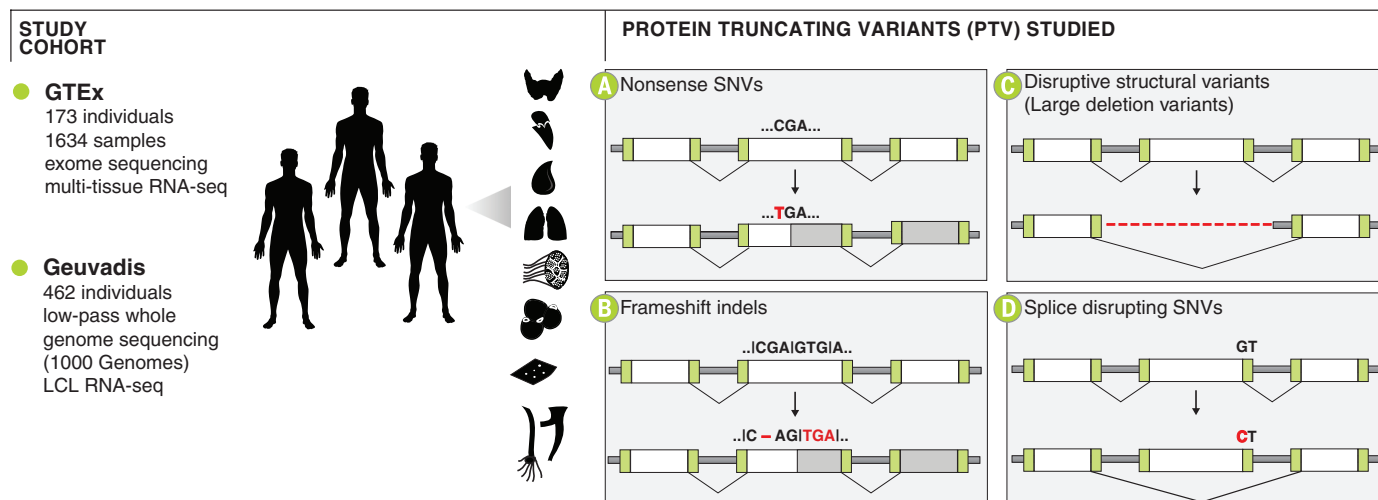
Fig. 1. Schematic overview of the study. We prepared an integrated DNA and RNA sequencing data set by combining the pilot phase of the GTEx project of 173 individuals with up to 30 tissues per individual (total of 1634 samples) and the Geuvadis project of LCL DNA and RNA sequencing in 462 individuals. From these data, we analyzed the effect of predicted protein-truncating genetic variants on the human transcriptome, including (**A**) nonsense SNVs, (**B**) frameshift indels, (**C**) large deletion variants, and (**D**) splice-disrupting SNVs.

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) (SM S2, Fig. 1, and figs. S1 and 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 4584 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, allele-specific expression (ASE) detecting different expression levels of two haplotypes of an individual, and split mappings across annotated exon junctions to quantify splicing (SM S3

and S4). Transcripts containing common PTVs are more weakly expressed and more tissue-specific than transcripts that do not contain common PTVs (SM S5 and figs. S3 to S7), 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 (SM S6) (4, 5, 10) for a total of 1814 PTVs (SM S6, figs. S8 to S12, and table S3) and validated the allelic ratios obtained from RNA-seq data (figs. S13 to S18 and table S4) (11). We also generated a method

to assess the ASE effect of frameshift indels (SM S6 and figs. S8 to S12), 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 (SM S7 and figs. S19 to S26) (15). We observe a higher proportion of strong or moderate allelic imbalance in rare and singleton nonsense SNVs compared with 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 50-bp (base pair) rule (SM S7) (16) have a larger proportion of ASE than SNVs that escape NMD (69.5% versus 31.9%,
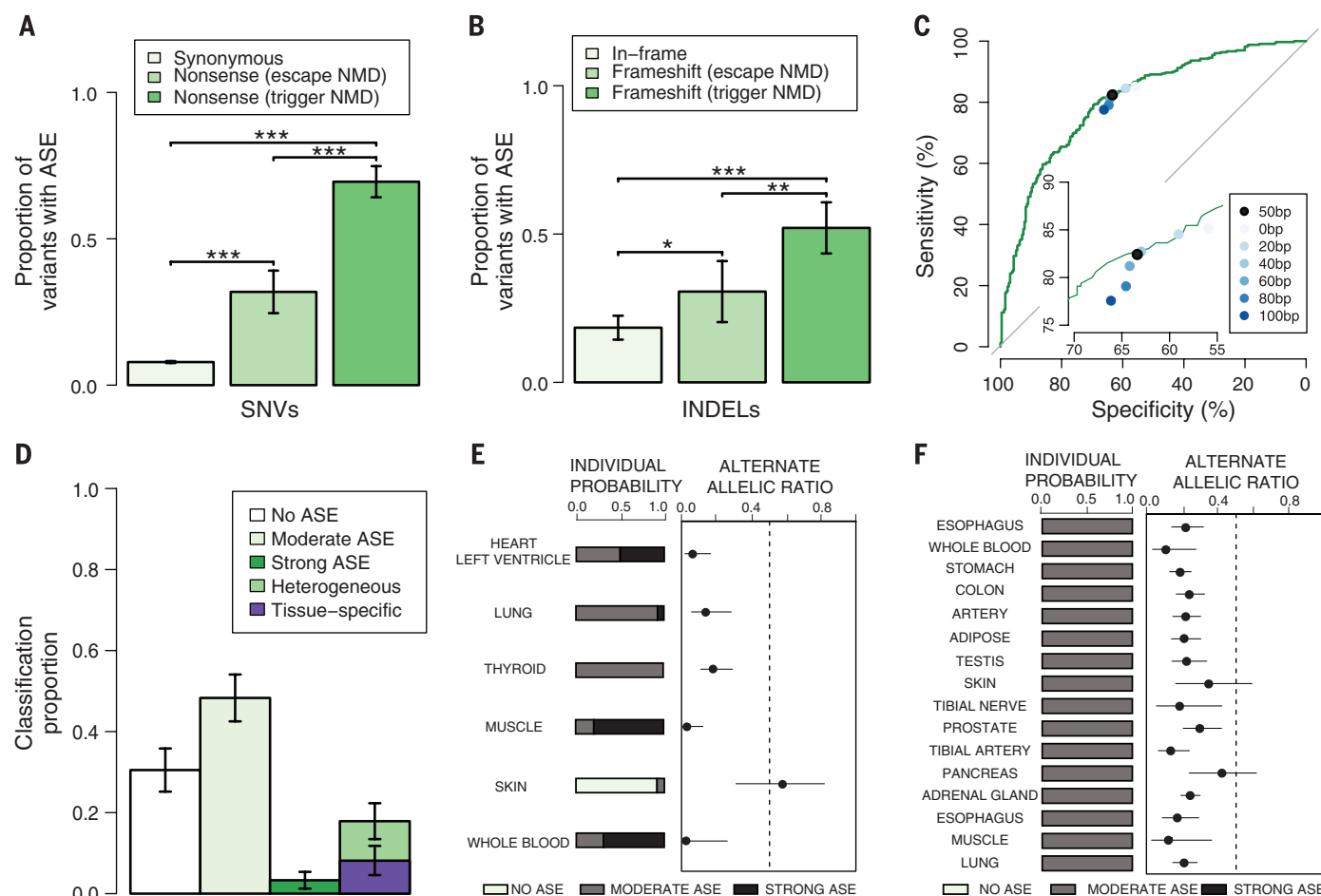

Fig. 2. Allele-specific expression analysis. (**A**) Proportion of rare SNVs with allele-specific expression (ASE) for synonymous variants (n = 25,233) and nonsense variants predicted to escape (n = 158) or trigger (n = 287) NMD. (**B**) Proportion of rare indels with ASE for in-frame (n = 355) and frameshift 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 (A). (**C**) Receiver operating characteristic 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**) Multitissue ASE classification for rare nonsense variants predicted to trigger NMD (n = 287). (**E**) Example of ASE data across six tissues for a heterozygous carrier of the nonsense variant rs149244943 in gene *PHKB* (phosphorylase kinase, beta) classified as having heterogeneous ASE effects across the six tissues. We confirmed that this effect is not driven by a common tissue-specific expression quantitative trait locus. (**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% CIs are shown.
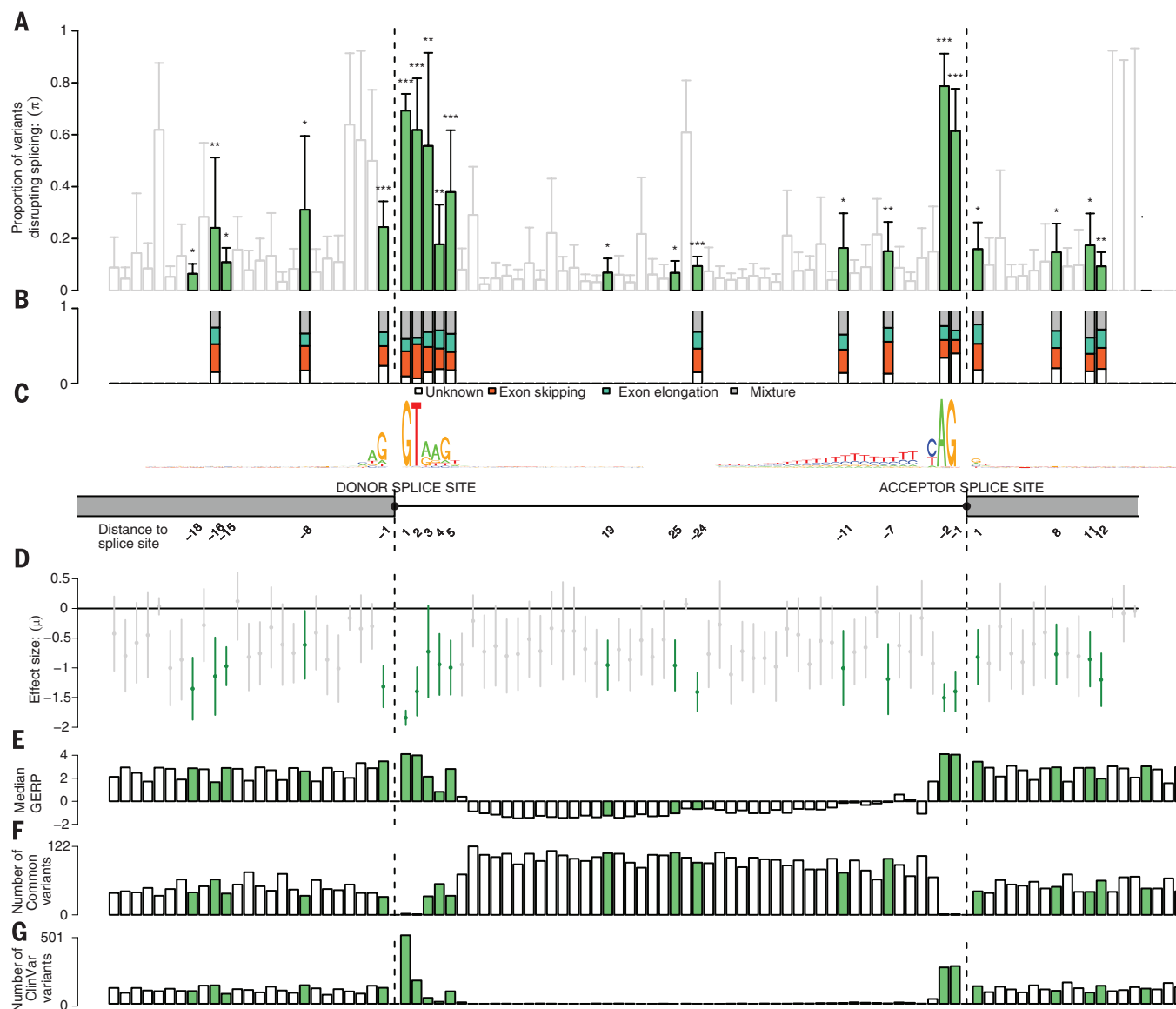
Fig. 3. Splicing disruption. (**A**) Proportion of variants disrupting splicing at each distance ±25 bp from donor and acceptor site (*$P < 0.05$, **$P < 0.01$; ***$P < 0.001$; green for $P < 0.05$; upper limit of 95% CI is shown; P value evaluated using the estimated proportion of variants supporting the alternative distribution times the effect size of the alternative distribution). (**B**) Classification of splice disruption events: exon skipping (low exon quantification value, no effect on intron quantification), exon elongation (high intron quantification value, no effect 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 genomic evolutionary rate profiling score (GERP) of all variants. (**F**) Distribution of common variants identified in an independent exome sequencing study of 4500 Swedish individuals. (**G**) Distribution of reported disease-causing variants in ClinVar.

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 with 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 50-bp 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 (SM S7). Our model predicts NMD better than the 50-bp rule, with an area under the curve (AUC) of 80.8% [95% confidence interval (CI) 77.3 to 84.4%] compared to a 50-bp rule AUC of 72.9% (69.3 to 76.5% CI) (Fig. 2C and figs. S21 and S22). Our results provide a quantitative estimate of the value of NMD predictions and illustrate that the 50-bp 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 under-

standing 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 (SM S7) (15) to rare nonsense variants predicted to trigger NMD, according to the 50-bp 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 (Fig. 2, D to F, and figs. S23 to S26). The tissue specificity of NMD implies that the same PTV may have different effects across tissues, which could contribute to tissue-specific effects of disease-causing mutations (*17*).

We examined whether 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 effect of genotyping error, we focused only on biallelic whole-gene deletions with strong experimental support and manual curation (SM S2 and figs. S27 to S29). We first analyzed the few examples of common whole-gene deletion polymorphisms (SM 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 with the population median ($P = 1.37 \times 10^{-5}$, one-sided binomial test of 11 rare PTV deletions in 25 genes) (figs. S27 and S28). Similar results were obtained for 53 nonsense PTVs with strong ASE signals ($P = 2.90 \times 10^{-9}$, one-sided binomial test) (figs. S30 and S31). 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 (minor allele frequencies ≤ 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 whether individuals carrying variants in these positions differ from the population in the quantification of the splice junction and the proximal exon and intron (Fig. 3, A to 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 to S37). We also find evidence of splice disruption from variants outside these regions, especially at positions 1 to 5 bp downstream of intronic donor splice sites, 1 bp 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 to S40, and SM S9). Our posterior probability estimates for sites with significant alternative distributions ($P < 0.05$) provide a resource for analyses (figs. S41 and 42 and SM S9 and S10).

By drawing on data from a wide range of adult tissues across 635 individuals, we provide a systematic assessment of the effect of predicted PTVs on the human transcriptome. Furthermore, this study indicates that nonsense-mediated decay has heterogeneous effects across tissues and also shows 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.

REFERENCES AND NOTES

1. J. A. Holbrook, G. Neu-Yilik, M. W. Hentze, A. E. Kulozik, *Nat. Genet.* **36**, 801–808 (2004).
2. P. D. Stenson *et al.*, *Hum. Genet.* **133**, 1–9 (2014).
3. J. C. Cohen, E. Boerwinkle, T. H. Mosley Jr., H. H. Hobbs, *N. Engl. J. Med.* **354**, 1264–1272 (2006).
4. D. G. MacArthur *et al.*, *Science* **335**, 823–828 (2012).
5. T. Lappalainen *et al.*, *Nature* **501**, 506–511 (2013).
6. P. A. C. 't Hoen *et al.*, *Nat. Biotechnol.* **31**, 1015–1022 (2013).
7. J. Lonsdale *et al.*, *Nat. Genet.* **45**, 580–585 (2013).
8. The GTEx Consortium, *Science* **348**, 648–660 (2014).
9. The 1000 Genomes Consortium, *Nature* **491**, 56–65 (2012).
10. K. R. Kukurba *et al.*, *PLOS Genet.* **10**, e1004304 (2014).
11. R. Zhang *et al.*, *Nat. Methods* **11**, 51–54 (2014).
12. S. B. Montgomery *et al.*, *Genome Res.* **23**, 749–761 (2013).
13. D. A. Skelly, M. Johansson, J. Madeoy, J. Wakefield, J. M. Akey, *Genome Res.* **21**, 1728–1737 (2011).
14. K. R. Stevenson, J. D. Coolon, P. J. Wittkopp, *BMC Genomics* **14**, 536 (2013).
15. M. Pirinen *et al.*, *Bioinformatics* **2015**, btv074 (2015).
16. E. Nagy, L. E. Maquat, *Trends Biochem. Sci.* **23**, 198–199 (1998).
17. J. F. Bateman, S. Freddi, G. Nattrass, R. Savarirayan, *Hum. Mol. Genet.* **12**, 217–225 (2003).
18. N. Huang, I. Lee, E. M. Marcotte, M. E. Hurles, *PLOS Genet.* **6**, e1001154 (2010).
19. A. A. McNally, L. Y. Yampolsky, *Genome Biol. Evol.* **2**, 44–52 (2010).
20. J. Zhou, B. Lemos, E. B. Dopman, D. L. Hartl, *Genome Biol. Evol.* **3**, 1014–1024 (2011).
21. N. A. Faustino, T. A. Cooper, *Genes Dev.* **17**, 419–437 (2003).
22. D. J. McCarthy *et al.*, *Genome Medicine* **6**, 26 (2014).
23. C. B. Burge, T. Tuschl, P. A. Sharp, *Cold Spring Harbor Monogr. Arch.* **37**, 525–560 (1999).
24. H. Y. Xiong *et al.*, *Science* **347**, 1254806 (2015).
25. A. Corvelo, M. Hallegger, C. W. Smith, E. Eyras, *PLOS Comput. Biol.* **6**, e1001016 (2010).
26. S. M. Purcell *et al.*, *Nature* **506**, 185–190 (2014).
27. H. Kacser, J. A. Burns, *Genetics* **97**, 639–666 (1981).

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/348/6235/666/suppl/DC1
Materials and Methods
Figs. S1 to S42
Tables S1 to S7
Data File S1
References (*28–56*)

Science

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

Manuel A. Rivas, Matti Pirinen, Donald F. Conrad, Monkol Lek, Emily K. Tsang, Konrad J. Karczewski, Julian B. Maller, Kimberly R. Kukurba, David S. DeLuca, Menachem Fromer, Pedro G. Ferreira, Kevin S. Smith, Rui Zhang, Fengmei Zhao, Eric Banks, Ryan Poplin, Douglas M. Ruderfer, Shaun M. Purcell, Taru Tukiainen, Eric V. Minikel, Peter D. Stenson, David N. Cooper, Katharine H. Huang, Timothy J. Sullivan, Jared Nedzel, The GTEx Consortium, The Geuvadis Consortium, Carlos D. Bustamante, Jin Billy Li, Mark J. Daly, Roderic Guigo, Peter Donnelly, Kristin Ardlie, Michael Sammeth, Emmanouil T. Dermitzakis, Mark I. McCarthy, Stephen B. Montgomery, Tuuli Lappalainen and Daniel G. MacArthur

Expression, genetic variation, and tissues

Human genomes show extensive genetic variation across individuals, but we have only just started documenting the effects of this variation on the regulation of gene expression. Furthermore, only a few tissues have been examined per genetic variant. In order to examine how genetic expression varies among tissues within individuals, the Genotype-Tissue Expression (GTEx) Consortium collected 1641 postmortem samples covering 54 body sites from 175 individuals. They identified quantitative genetic traits that affect gene expression and determined which of these exhibit tissue-specific expression patterns. Melé *et al.* measured how transcription varies among tissues, and Rivas *et al.* looked at how truncated protein variants affect expression across tissues.
Science, this issue p. 648, p. 660, p. 666; see also p. 640

ARTICLE TOOLS	http://science.sciencemag.org/content/348/6235/666
SUPPLEMENTARY MATERIALS	http://science.sciencemag.org/content/suppl/2015/05/06/348.6235.666.DC1
RELATED CONTENT	http://science.sciencemag.org/content/sci/348/6235/618.full http://science.sciencemag.org/content/sci/348/6235/640.full http://science.sciencemag.org/content/sci/348/6235/648.full http://science.sciencemag.org/content/sci/348/6235/660.full
REFERENCES	This article cites 55 articles, 9 of which you can access for free http://science.sciencemag.org/content/348/6235/666#BIBL
PERMISSIONS	http://www.sciencemag.org/help/reprints-and-permissions