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. DeLuna,4 Menachem Fromer,4,5,9 Pedro G. Ferreira,10,11,12 Kevin S. Smith,6,7 Rui Zhang,9 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 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,5,10,11,12,19,20 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 (5). 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

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.

1Wellcome Trust Centre for Human Genetics, Nuffield Department of Clinical Medicine, University of Oxford, Oxford, UK. 2Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland. 3Washington University in St. Louis, St. Louis, MO, USA. 4Broad Institute of MIT and Harvard, Cambridge, MA, USA. 5Analytical and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA. 6Department of Genetics, Stanford University, Stanford, CA, USA. 7Department of Pathology, Stanford University, Stanford, CA, USA. 8Biomedical Informatics Program, Stanford University, Stanford, CA, USA. 9Department of Psychiatry, Mt. Sinai Hospital, NY, USA. 10Department of Genetic Medicine and Development, University of Geneva, Geneva, Switzerland. 11Institute for Genetics and Genomics in Geneva (iGE3), University of Geneva, Geneva, Switzerland. 12Swiss Institute of Bioinformatics, Geneva, Switzerland. 13Division of Psychiatric Genomics, Department of Psychiatry, Icahn School of Medicine at Mount Sinai, NY, USA. 14Institute of Medical Genetics, School of Medicine, Cardiff University, Heath Park, Cardiff, UK. 15Center for Genomic Regulation (CRG), Universitat Pompeu Fabra (UPF), Barcelona, Catalonia, Spain. 16Department of Statistics, University of Oxford, Oxford, UK. 17National Institute for Scientific Computing (LNLNC), Petropolis, Rio de Janeiro, Brazil. 18Oxford Center for Diabetes Endocrinology and Metabolism, University of Oxford, Oxford, UK. 19New York Genome Center, New York, NY, USA. 20Department of Systems Biology, Columbia University, New York, NY, USA. 21Department of Medicine, Harvard Medical School, Boston, MA, USA.

*Corresponding author. E-mail: rivas@well.ox.ac.uk (M.A.R.); macarthu@alum.mit.edu (D.G.M.). ‡These authors contributed equally to this work. §These authors contributed equally to this work.

STUDY COHORT

- **GTEx**
  - 173 individuals
  - 1634 samples
  - exome sequencing
  - multi-tissue RNA-seq

- **Geuvadis**
  - 462 individuals
  - low-pass whole genome sequencing
  - (1000 Genomes)
  - LCL RNA-seq

PROTEIN TRUNCATING VARIANTS (PTV) STUDIED

- **A** Nonsense SNVs
  - ...CGA...
  - ...TGA...

- **B** Frameshift indels
  - ...AGA/GTGIA...
  - ...AC...

- **C** Disruptive structural variants
  - (Large deletion variants)
  - GT

- **D** Splice disrupting SNVs
  - CT

666 8 MAY 2015 • VOL 348 ISSUE 6235 sciencemag.org SCIENCE
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 (7). Variants that escape NMD may create a truncated protein with dominant-negative or gain-of-function effects (7). 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) (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 PHK8 (phosphorylase kinase, beta) classified as having heterogeneous ASE effects across the six tissues. (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 TPPI (tripetidyl peptidase I), classified as having moderate ASE across all tissues. For all plots, 95% CIs are shown.

Downloaded from http://science.sciencemag.org/ on August 6, 2017
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 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 (SM S7) 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 PTM 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 (16). Dosage compensation has been reported to correlate with gene expression levels (19) and occur in over 80% of deleted genes in Dro sophila melanographaster (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 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 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).

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 splice sites at the splice junction (Figs. S30 to S37). We also had consistently decreased expression of a dominant model, providing no evidence for results, figs. S30 and S31). 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**


**ACKNOWLEDGMENTS**

We 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 RO1GM104371 to D.G.M.; NIH MH090043 to E.T.D. and S.B.M.; S.B.M. R.G., T.L., and M.M.; RMHM090494 to E.T.D. and M.M.; U19HG007593 to J.B.L. and S.B.M.; and RMHM101810 to D.F.C.); Academy of Finland (257654 to M.P.); a Hewlett-Packard Stanford Graduate Fellowship and a doctoral fellowship from the Natural Science and Engineering Research Council of Canada to E.K.T.; a National Defense Science and Engineering Graduate Fellowship (NDSEG) from the United States Department of Defense (DoD) to K.R.K.; European Research Council, Swiss National Science Foundation, and Louis-Jeantet Foundation to E.T.D.; Wellcome Trust (095552/22/11/2 and 090532/2/0/2 to P.D. and 093831 to M.M.); and a Clarendon Scholarship, NDM Studentship, and Green Templeton College Award from University of Oxford to M.A.R. The Genotype-Tissue Expression (GTEX) project was supported by the Common Fund of the Office of the Director of NIH. Additional funds were enrolled at Biospecimen Source sites funded by NCIC-SAIC-Frederick, Inc. (SAIC-F) subcontracts to the National Disease Research Interchange (DIX5170), Roswell Park Cancer Institute (DIX5717), and Science Care, Inc. (XDS1572). 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 DA005227. Statistical methods development grants were made to the University of Geneva (MH050941), the University of Chicago (MH090951 and MH090957), the University of North Carolina–Chapel Hill (MH090356), and Harvard University (MH090485). 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 the database of Genotypes and Phenotypes (www.ncbi.nlm.nih.gov/gap) (psh000424.v5.p1, except for whole-exome sequencing data in psh000424.v6.p1 and mmPCR-seq data and processed ASE data in psh000424.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 are available on the GTEx Portal (www.gtexportal.org/home/samplesPage). The Geuvadis data are available in ArrayExpress accession E-GEUV-1. Further details and links to data and software are available at www.well.ox.ac.uk/~rivas/ptv2015. C.D.B. 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 Genetics, LLC; he also owns stock options in Personalis, Invitae, and Identify Genetics, LLC.

**SUPPLEMENTARY MATERIALS**

www.sciencemag.org/content/348/6235/668/suppl/DC1

Materials and Methods
Figs. S1 to S42
Tables S1 to S7
Data File S1
References (28–56)

29 September 2014; accepted 2 April 2015
10.1126/science.1261877
Effect of predicted protein-truncating genetic variants on the human transcriptome


Sci. 348 (6235), 666-669.
DOI: 10.1126/science.1261877

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

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

Use of this article is subject to the Terms of Service

© The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. The title Science is a registered trademark of AAAS.