Heterozygous gene truncation delineates the human haploinsufficient genome

Istvan Bartha[1,2]†, Antonio Rausell[1,3]†, Paul J McLaren[1,2], Manuel Tardaguila[1], Pejman Mohammadi[1,4], Jacques Fellay[1,2], Amalio Telenti[5]*

[1]SIB Swiss Institute of Bioinformatics, Lausanne and Basel, Switzerland

[2]School of Life Sciences, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

[3]Vital-IT group, SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland

[4]Computational Biology Group, ETH Zurich, Switzerland

[5]J. Craig Venter Institute, La Jolla, CA, US

.

*Correspondence to: atelenti@jcvi.org

†Equal contribution

Sequencing projects have identified large numbers of rare stop gain and frameshift variants in the human genome. Because most of these are observed in the heterozygous state, they test the genome's tolerance to dominant loss of function. We analyzed the distribution of truncating variants in 16818 protein coding autosomal genes from the exomes of 11546 individuals. We observed 45044 truncating variants affecting 12735 genes, against an expectation of observing 13612 genes under a model of neutral evolution. Extrapolating this to a larger human population, we estimate that 8-9% of genes would not tolerate heterozygous truncation. The study of dominant loss of function variants delineates the essential genome, and underscores the functional importance of haploinsufficiency.

One Sentence Summary: Analysis of heterozygous gene truncating variants in >11,000 individuals identifies the properties of genes that require two functional copies.

The abundance of rare human genetic variation results from recent population expansion and limited purifying selection (*1-3*). Homozygosity of rare functional variants is a well-known cause of severe Mendelian disorders. However, since most variants have low allele frequency, they are observed almost exclusively in the heterozygous state and may not be prioritized for further analyses. We focused on stop-gain (nonsense) and frameshift variants as a tractable set of variants for the assessment of heterozygous gene loss across the genome. The study dataset included 11546 exomes in which we observed 45044 truncating variants. We considered transcripts from autosomal protein coding genes. The expectation of a neutral model of evolution, under which there is no selection against truncating variants, was that 13612 (IQR, 13575-13651) of 16818 genes reliably annotated by the Consensus CDS (CCDS) project (*4*) would be targeted by stop gains or frameshift variants. We generated this expectation by simulating 45044 variants according to the neutral per gene *de novo* probabilities (*5*). However, only 12735 such genes were observed harboring a total of 45044 truncating variants (**Figure 1**). We estimated the number of genes intolerant to truncating variants in a probabilistic framework assuming that most genes accumulate truncating variants according to their neutral de novo probabilities and that a small number of genes (*N*) do not accumulate truncation. The number of genes bearing truncation given *N* is simulated by sampling variants from the multinomial distribution of the neutral probabilities of *N* genes (**Figure S1**). By minimizing the difference between the observed number of genes without a truncating variant and the simulated scenarios we estimate the total size of the haploinsufficient genome (i.e. genes that can not tolerate a single functional copy) from 1500 to 1600, which is about 9% of the autosomal protein coding genes analyzed (**Figure 1**). The empirically testing of this prediction would require the observation of more than 150,000 truncating variants, in a study population of at least 80,000 individuals.

Truncation may not necessarily result in loss of function, and the severity of such variants can be explored through various pathogenicity scores. Prediction of functional consequences use gene-level features such as conservation, centrality, or tolerance to mutational burden (*6, 7*), or use sequence-level attributes. We used NutVar (http://nutvar.labtelenti.org), a sequence-based score that assesses loss of functional domains, differential impact on isoforms, and degradation by non-sense mediated decay (*8*), to estimate the functional consequences of a truncation variant. The predicted severity given by NutVar associates with levels of gene expression and thus with potential for haploinsufficieny (*8*). We partitioned the variants into mild and severe categories based on a cutoff of 0.4 in NutVar score that has the greatest discrimination power for consequences on gene expression (see Material and Methods). On this basis, we estimated that in addition to the 1500-1600 genes not tolerating any truncation, there are an estimated 2300 genes that may tolerate truncations with predicted limited functional consequences (truncations with a NutVar score lower than 0.4) but remain intolerant to severe truncations (**Figure 1**). Although we benchmarked on the two most definitive classes of truncating variants, stop gain and frameshifts, the literature frequently considers variants at splice site acceptors and donors. Functional consequences are diverse as splice site variation may lead to exon skipping or to truncation of the transcript. We thus extended the analysis to include 12378 splice site variants (bringing the total of observed truncating variants to 57422). We observed 13758 genes without any of these variants and estimated a lower bound of 1180-1260 genes intolerant to splice site, stop gain or frameshift variants. We did not considered large structural variants.

Homozygous truncations are observed at a much lower frequency than heterozygous variants and are expected to be more deleterious. In our dataset, 2808 genes carried homozygous

stop-gain or frameshift variants and can be considered to represent a subset of the dispensable genome (**Table 1, inset**). Given the relatively low frequency of homozygous loss-of-function genotypes, a full mapping of the dispensable genome will require a very large number of participants. In contrast, the collectively prevalent nature of rare heterozygous variants suggests that a map of "essentiality" on the basis of heterozygosity is within reach, and that it will be immediately relevant for human health. In mice, when homozygous knockout mutants are not viable, up to 71.7% of heterozygotes have a phenotypic hit (*9*). Systematic phenotyping of knockout mice revealed that haploinsufficiency might be more common than generally suspected (*10*). Indeed, our study indicates that there is considerable purifying selection acting on heterozygous truncating variants. Heterozygous truncation can lead to deleterious functional consequences through haploinsufficiency due to decreased gene dosage, or through a dominant-negative effect (*11, 12*). In 2008, Dang et al. established a list of 299 haploinsufficient genes linked to human diseases (*13*). The number has since increased: for example, a recent report highlights the consequence of haploinsufficiency in humans: frameshift and splice site variants in *CTLA-4*, encoding cytotoxic T-lymphocyte-associated protein 4, associated with undiagnosed or misdiagnosed autoimmune disorders in the second to fifth decades of life (*14*).

We next evaluated the characteristics of the genes not observed as carrying a truncation in our study datasets (n=4083). As previously reported (*13*), genes encoding transcription factors, and genes that function in development, the cell cycle, and nucleic acid metabolism are overrepresented among those genes not carrying a truncation (**Figure 2**). In line with expectations, genes without observed truncating mutations were highly conserved ($p=2x10^{-280}$). These genes had fewer paralogs ($p=3x10^{-5}$), a higher probability to affect cell survival in genome-wide CRISPR-Cas9 experiments (odds ratio=2.75, $p=1.4x10^{-13}$), and to be part of protein complexes (odds ratio=3.16, $p=1.7x10^{-28}$), and had higher posterior probabilities in computational prediction of haploinsufficiency (p=1E-119), **Table 1**.

The concept of the essential genome has been explored in analyses of minimal bacterial genomes (*15*), mice knockout studies (*16*), studies of transposon or chemical mutagenesis, and more recently in studies that used CRISPR-Cas9 genome-editing technology (*17, 18*). One limitation of these studies is that observation of phenotypes may require adequate exposure to specific environmental interactions – a practical obstacle in animal studies (*10*). In contrast, in humans, there is a possibility that life-long exposures will eventually reveal a phenotypic trait or disease associated with heterozygous gene truncations. In fact, the observation of a given heterozygous truncating variant does not imply the absence of biological consequences. Although this study has been performed using data from a presumably healthy adult human population, we cannot exclude a pathogenic role for any of the observed truncating variants. The potential impact of such genetic changes might be masked by incomplete penetrance (*19*) or by the simultaneous occurrence of compensatory variants (*20*). Because monoallelic lesions are more common than biallelic lesions, heterozygous variants associated with haploinsufficiency will play an important role in human disease. The rapidly increasing number of sequenced genomes will continue to improve our knowledge of the essential, the haploinsufficient and the dispensable parts of the human genome.

References and Notes:

1. J. A. Tennessen, A. W. Bigham, T. D. O'Connor, W. Fu, E. E. Kenny, S. Gravel, S. McGee, R. Do, X. Liu, G. Jun, H. M. Kang, D. Jordan, S. M. Leal, S. Gabriel, M. J. Rieder, G. Abecasis, D. Altshuler, D. A. Nickerson, E. Boerwinkle, S. Sunyaev, C. D. Bustamante, M. J. Bamshad, J. M. Akey, G. O. Broad, G. O. Seattle, N. E. S. Project, Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64-69 (2012); published online EpubJul 6 (10.1126/science.1219240).
2. M. R. Nelson, D. Wegmann, M. G. Ehm, D. Kessner, P. St Jean, C. Verzilli, J. Shen, Z. Tang, S. A. Bacanu, D. Fraser, L. Warren, J. Aponte, M. Zawistowski, X. Liu, H. Zhang, Y. Zhang, J. Li, Y. Li, L. Li, P. Woollard, S. Topp, M. D. Hall, K. Nangle, J. Wang, G. Abecasis, L. R. Cardon, S. Zollner, J. C. Whittaker, S. L. Chissoe, J. Novembre, V. Mooser, An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* **337**, 100-104 (2012); published online EpubJul 6 (10.1126/science.1217876).
3. D. G. MacArthur, T. A. Manolio, D. P. Dimmock, H. L. Rehm, J. Shendure, G. R. Abecasis, D. R. Adams, R. B. Altman, S. E. Antonarakis, E. A. Ashley, J. C. Barrett, L. G. Biesecker, D. F. Conrad, G. M. Cooper, N. J. Cox, M. J. Daly, M. B. Gerstein, D. B. Goldstein, J. N. Hirschhorn, S. M. Leal, L. A. Pennacchio, J. A. Stamatoyannopoulos, S. R. Sunyaev, D. Valle, B. F. Voight, W. Winckler, C. Gunter, Guidelines for investigating causality of sequence variants in human disease. *Nature* **508**, 469-476 (2014); published online EpubApr 24 (10.1038/nature13127).
4. K. D. Pruitt, J. Harrow, R. A. Harte, C. Wallin, M. Diekhans, D. R. Maglott, S. Searle, C. M. Farrell, J. E. Loveland, B. J. Ruef, E. Hart, M. M. Suner, M. J. Landrum, B. Aken, S. Ayling, R. Baertsch, J. Fernandez-Banet, J. L. Cherry, V. Curwen, M. Dicuccio, M. Kellis, J. Lee, M. F. Lin, M. Schuster, A. Shkeda, C. Amid, G. Brown, O. Dukhanina, A. Frankish, J. Hart, B. L. Maidak, J. Mudge, M. R. Murphy, T. Murphy, J. Rajan, B. Rajput, L. D. Riddick, C. Snow, C. Steward, D. Webb, J. A. Weber, L. Wilming, W. Wu, E. Birney, D. Haussler, T. Hubbard, J. Ostell, R. Durbin, D. Lipman, The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome research* **19**, 1316-1323 (2009); published online EpubJul (10.1101/gr.080531.108).
5. K. E. Samocha, E. B. Robinson, S. J. Sanders, C. Stevens, A. Sabo, L. M. McGrath, J. A. Kosmicki, K. Rehnstrom, S. Mallick, A. Kirby, D. P. Wall, D. G. MacArthur, S. B. Gabriel, M. DePristo, S. M. Purcell, A. Palotie, E. Boerwinkle, J. D. Buxbaum, E. H. Cook, Jr., R. A. Gibbs, G. D. Schellenberg, J. S. Sutcliffe, B. Devlin, K. Roeder, B. M. Neale, M. J. Daly, A framework for the interpretation of de novo mutation in human disease. *Nature genetics* **46**, 944-950 (2014); published online EpubSep (10.1038/ng.3050).
6. D. G. MacArthur, S. Balasubramanian, A. Frankish, N. Huang, J. Morris, K. Walter, L. Jostins, L. Habegger, J. K. Pickrell, S. B. Montgomery, C. A. Albers, Z. D. Zhang, D. F. Conrad, G. Lunter, H. Zheng, Q. Ayub, M. A. DePristo, E. Banks, M. Hu, R. E. Handsaker, J. A. Rosenfeld, M. Fromer, M. Jin, X. J. Mu, E. Khurana, K. Ye, M. Kay, G. I. Saunders, M. M. Suner, T. Hunt, I. H. Barnes, C. Amid, D. R. Carvalho-Silva, A. H. Bignell, C. Snow, B. Yngvadottir, S. Bumpstead, D. N. Cooper, Y. Xue, I. G. Romero, C. Genomes Project, J. Wang, Y. Li, R. A. Gibbs, S. A. McCarroll, E. T. Dermitzakis, J. K. Pritchard, J. C. Barrett, J. Harrow, M. E. Hurles, M. B. Gerstein, C. Tyler-Smith, A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823-828 (2012); published online EpubFeb 17 (10.1126/science.1215040).
7. S. Petrovski, Q. Wang, E. L. Heinzen, A. S. Allen, D. B. Goldstein, Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS genetics* **9**, e1003709 (2013)10.1371/journal.pgen.1003709).
8. A. Rausell, P. Mohammadi, P. J. McLaren, I. Bartha, I. Xenarios, J. Fellay, A. Telenti, Analysis of stop-gain and frameshift variants in human innate immunity genes. *PLoS computational biology* **10**, e1003757 (2014); published online EpubJul (10.1371/journal.pcbi.1003757).
9. A. Ayadi, M. C. Birling, J. Bottomley, J. Bussell, H. Fuchs, M. Fray, V. Gailus-Durner, S. Greenaway, R. Houghton, N. Karp, S. Leblanc, C. Lengger, H. Maier, A. M. Mallon, S. Marschall, D. Melvin, H. Morgan, G. Pavlovic, E. Ryder, W. C. Skarnes, M. Selloum, R. Ramirez-Solis, T. Sorg, L. Teboul, L. Vasseur, A. Walling, T. Weaver, S. Wells, J. K. White, A. Bradley, D. J. Adams, K. P. Steel, M. Hrabe de Angelis, S. D. Brown, Y. Herault, Mouse large-scale phenotyping initiatives: overview of the European Mouse Disease Clinic (EUMODIC) and of the Wellcome Trust Sanger Institute Mouse Genetics Project. *Mammalian*

genome : official journal of the International Mammalian Genome Society **23**, 600-610 (2012); published online EpubOct (10.1007/s00335-012-9418-y).

10. J. K. White, A. K. Gerdin, N. A. Karp, E. Ryder, M. Buljan, J. N. Bussell, J. Salisbury, S. Clare, N. J. Ingham, C. Podrini, R. Houghton, J. Estabel, J. R. Bottomley, D. G. Melvin, D. Sunter, N. C. Adams, P. Sanger Institute Mouse Genetics, D. Tannahill, D. W. Logan, D. G. Macarthur, J. Flint, V. B. Mahajan, S. H. Tsang, I. Smyth, F. M. Watt, W. C. Skarnes, G. Dougan, D. J. Adams, R. Ramirez-Solis, A. Bradley, K. P. Steel, Genome-wide generation and systematic phenotyping of knockout mice reveals new roles for many genes. *Cell* **154**, 452-464 (2013); published online EpubJul 18 (10.1016/j.cell.2013.06.022).
11. E. Fisher, P. Scambler, Human haploinsufficiency--one for sorrow, two for joy. *Nature genetics* **7**, 5-7 (1994); published online EpubMay (10.1038/ng0594-5).
12. N. Huang, I. Lee, E. M. Marcotte, M. E. Hurles, Characterising and predicting haploinsufficiency in the human genome. *PLoS genetics* **6**, e1001154 (2010); published online EpubOct (10.1371/journal.pgen.1001154).
13. V. T. Dang, K. S. Kassahn, A. E. Marcos, M. A. Ragan, Identification of human haploinsufficient genes and their genomic proximity to segmental duplications. *European journal of human genetics : EJHG* **16**, 1350-1357 (2008); published online EpubNov (10.1038/ejhg.2008.111).
14. H. S. Kuehn, W. Ouyang, B. Lo, E. K. Deenick, J. E. Niemela, D. T. Avery, J. N. Schickel, D. Q. Tran, J. Stoddard, Y. Zhang, D. M. Frucht, B. Dumitriu, P. Scheinberg, L. R. Folio, C. A. Frein, S. Price, C. Koh, T. Heller, C. M. Seroogy, A. Huttenlocher, V. K. Rao, H. C. Su, D. Kleiner, L. D. Notarangelo, Y. Rampertaap, K. N. Olivier, J. McElwee, J. Hughes, S. Pittaluga, J. B. Oliveira, E. Meffre, T. A. Fleisher, S. M. Holland, M. J. Lenardo, S. G. Tangye, G. Uzel, Immune dysregulation in human subjects with heterozygous germline mutations in CTLA4. *Science* **345**, 1623-1627 (2014); published online EpubSep 26 (10.1126/science.1255904).
15. C. A. Hutchison, S. N. Peterson, S. R. Gill, R. T. Cline, O. White, C. M. Fraser, H. O. Smith, J. C. Venter, Global transposon mutagenesis and a minimal Mycoplasma genome. *Science* **286**, 2165-2169 (1999); published online EpubDec 10 (
16. A. Bradley, K. Anastassiadis, A. Ayadi, J. F. Battey, C. Bell, M. C. Birling, J. Bottomley, S. D. Brown, A. Burger, C. J. Bult, W. Bushell, F. S. Collins, C. Desaintes, B. Doe, A. Economides, J. T. Eppig, R. H. Finnell, C. Fletcher, M. Fray, D. Frendewey, R. H. Friedel, F. G. Grosveld, J. Hansen, Y. Herault, G. Hicks, A. Horlein, R. Houghton, M. Hrabe de Angelis, D. Huylebroeck, V. Iyer, P. J. de Jong, J. A. Kadin, C. Kaloff, K. Kennedy, M. Koutsourakis, K. C. Lloyd, S. Marschall, J. Mason, C. McKerlie, M. P. McLeod, H. von Melchner, M. Moore, A. O. Mujica, A. Nagy, M. Nefedov, L. M. Nutter, G. Pavlovic, J. L. Peterson, J. Pollock, R. Ramirez-Solis, D. E. Rancourt, M. Raspa, J. E. Remacle, M. Ringwald, B. Rosen, N. Rosenthal, J. Rossant, P. Ruiz Noppinger, E. Ryder, J. Z. Schick, F. Schnutgen, P. Schofield, C. Seisenberger, M. Selloum, E. M. Simpson, W. C. Skarnes, D. Smedley, W. L. Stanford, A. F. Stewart, K. Stone, K. Swan, H. Tadepally, L. Teboul, G. P. Tocchini-Valentini, D. Valenzuela, A. P. West, K. Yamamura, Y. Yoshinaga, W. Wurst, The mammalian gene function resource: the International Knockout Mouse Consortium. *Mammalian genome : official journal of the International Mammalian Genome Society* **23**, 580-586 (2012); published online EpubOct (10.1007/s00335-012-9422-2).
17. O. Shalem, N. E. Sanjana, E. Hartenian, X. Shi, D. A. Scott, T. S. Mikkelsen, D. Heckl, B. L. Ebert, D. E. Root, J. G. Doench, F. Zhang, Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343**, 84-87 (2014); published online EpubJan 3 (10.1126/science.1247005).
18. T. Wang, J. J. Wei, D. M. Sabatini, E. S. Lander, Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343**, 80-84 (2014); published online EpubJan 3 (10.1126/science.1246981).
19. F. Rieux-Laucat, J. L. Casanova, Immunology. Autoimmunity by haploinsufficiency. *Science* **345**, 1560-1561 (2014); published online EpubSep 26 (10.1126/science.1260791).
20. B. Szamecz, G. Boross, D. Kalapis, K. Kovacs, G. Fekete, Z. Farkas, V. Lazar, M. Hrtyan, P. Kemmeren, M. J. Groot Koerkamp, E. Rutkai, F. C. Holstege, B. Papp, C. Pal, The genomic landscape of compensatory evolution. *PLoS biology* **12**, e1001935 (2014); published online EpubAug (10.1371/journal.pbio.1001935).
21. W. H. Chen, P. Minguez, M. J. Lercher, P. Bork, OGEE: an online gene essentiality database. *Nucleic acids research* **40**, D901-906 (2012); published online EpubJan (10.1093/nar/gkr986).

Acknowledgments: Funded by the Swiss National Science Foundation (CRSII3-147665). The authors would like to thank the NHLBI GO Exome Sequencing Project and its ongoing studies; The 1000 Genomes Project, the TwinsUK Cohort; The Avon Longitudinal Study of Parents and Children; The

Genome of the Netherlands Project; The CoLaus cohor; The Swiss HIV Cohort Study and The National Institute of Environmental Health Science Environmental Genome Project.

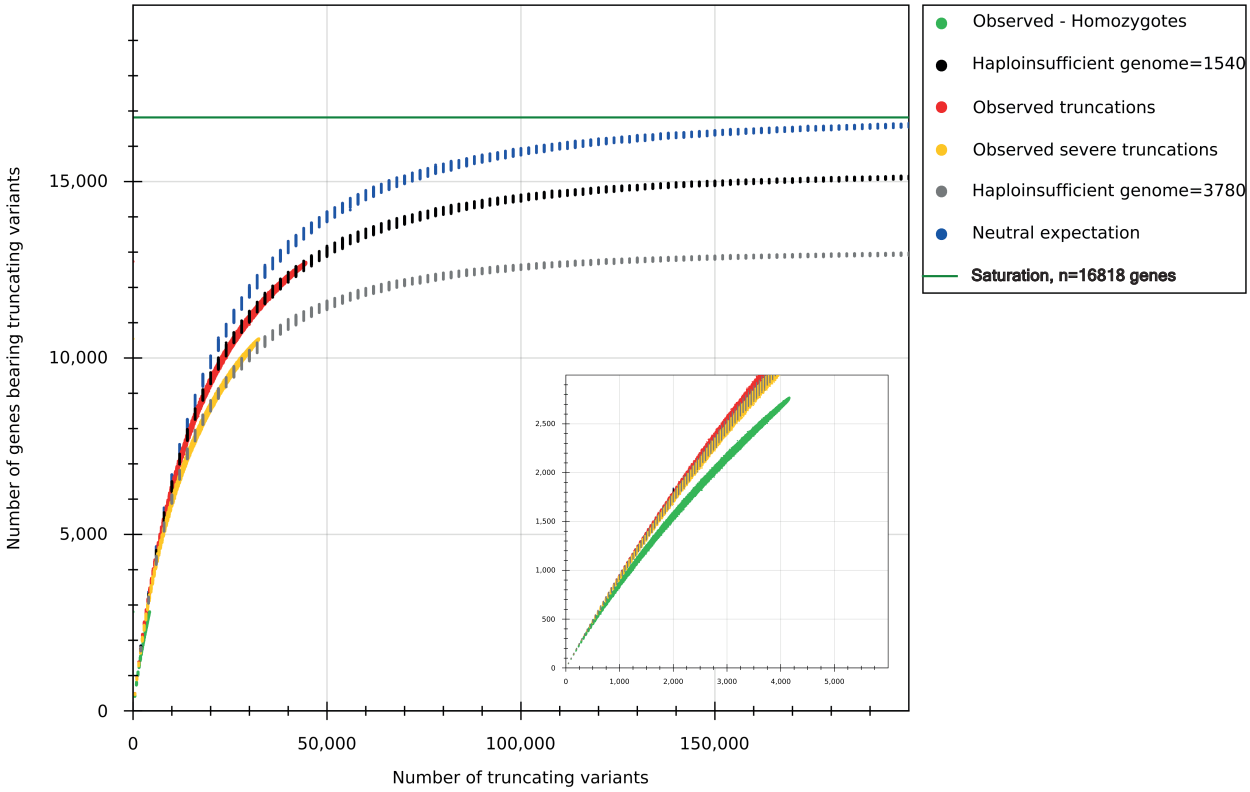

Fig. 1. Observed and expected truncating variants in the study population. Number of genes with at least one stop gain or frameshift variant as a function of the number of sampled truncation variants. The dotted blue curve follows the neutral expectation calculated from the *de novo* probabilities of Samocha et al. (*5*). The red curve shows gene counts when all observed variants are considered. The orange curve shows gene counts when only highly damaging variants (NutVar score >= 0.4) are considered. The black and grey curves denote the simulation results after fitting to 1540 or 3780 the number of genes intolerant to truncating variants. The green curve shows gene counts when homozygous variants are considered.

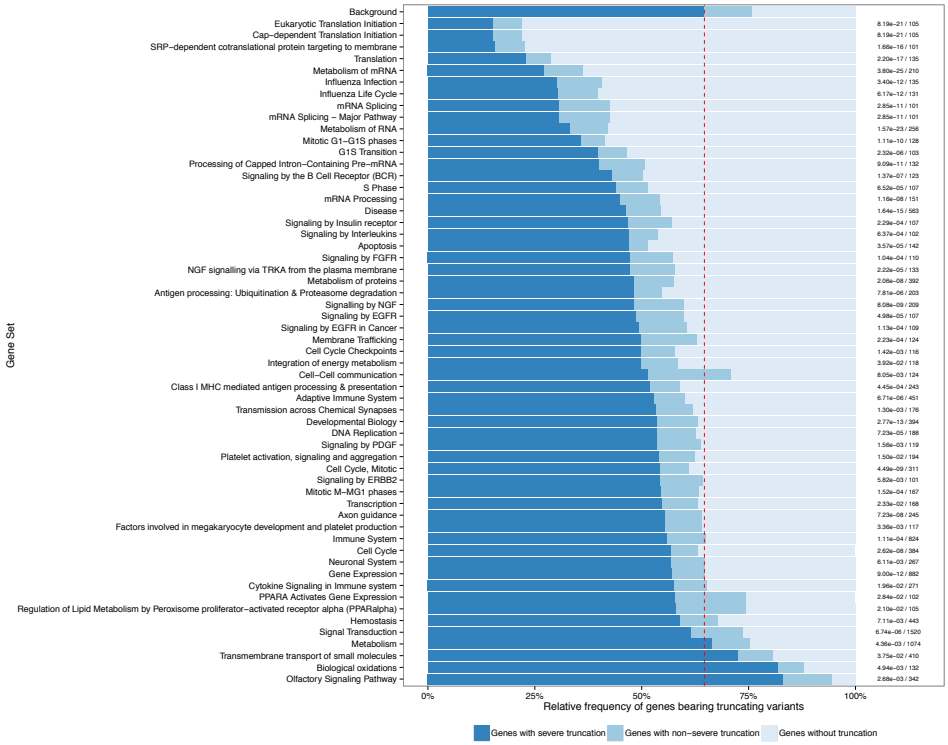

Fig. 2. Gene set overrepresentation results in genes with no observed truncating variants. The dark blue bar indicates the size of that part of the gene set which does not bear truncating variants. The middle bar indicates genes that bear slightly deleterious variants (NutVar score <= 0.4), while the light blue bar denotes the percentage of genes with highly deleterious variants. Each line is a gene set defined by Gene Ontology, the sets are not necessarily independent. P-values are adjusted by the Benjamini-Hochberg method and only significant gene sets are shown. Bars are colored according to clustering of gene sets.

Table 1. **Orthogonal characteristics of the genes observed with no stop gain or frameshift variants**. Binary traits were evaluated with logistic regression; continuous traits were evaluated by Mann-Whitney U test.

Annotation	Effect in non-truncated genes	P-value	Test	Data Source
Paralog count	Lower	3E-05	rank-sum test	Ensembl Biomart
In protein complexes	Enrichment	1E-28	logistic regression	Gene Ontology term GO_0043234
Probability of haploinsufficiency	Higher	1E-119	rank-sum test	Huang et al. (*12*)
Loss of cell viability	Enrichment	1E-13	logistic regression	Shalem et al. (*17*)
Pathogenic genes	-	NS	logistic regression	ClinVar
dN/dS	Lower (conservation)	2E-280	rank-sum test	Ensembl primate genomes
Essentiality	Higher	6E-11	logistic regression	OGEE (http://ogeedb.embl.de/)
Centrality	Higher	5E-52	rank-sum test	OGEE (http://ogeedb.embl.de/)

Supplementary Materials:

Materials and Methods

Exomes. We collected exome data from public and non-public sources (**Table S1**). Individuals include the general population without severe disorders before reproductive age. Variants were filtered based on Hardy-Weinberg equilibrium (discarded if p<1E-8). For public data sets, variants were called at the data source with their respective pipelines. For in-house data sets, sequence reads were aligned using BWA, and called with Haplotypecaller from GATK 3.1. Variants were annotated with SnpEff 3.1 and filtered as described in Rausell et al. (*8*). Only genes and transcripts from autosomal protein coding genes reliably annotated by the Consensus CDS (CCDS) project (*4*) were considered. As a reference background throughout all analyses, a total number of 16818 autosomal protein coding genes was obtained by considering genes with synonymous variants detected in the exome data, passing all previous filters and for which neutral probabilities of generating de novo truncating variants are available; see below. In downstream analyses, we represented each gene with its predicted most damaging stop-gain or frameshift variant according to the NutVar score (http://nutvar.labtelenti.org). In the analysis including splice site variants we considered all stop gain, frameshift, splice site acceptor and splice site donor variants.

Neutral models and simulations. Neutral gene probabilities of generating a *de novo* truncating variant (stop-gain, frameshift, splice-disrupting) were obtained from Samocha et al. (*5*). We estimated the number of genes intolerant to truncating variants in a probabilistic framework assuming that most genes accumulate truncating variants according to their neutral *de novo* probabilities and that a small number of genes (*N*) do not accumulate truncation. We estimated *N* by fitting the expected number of genes bearing a truncation with 25k, 30k, 35k, 40k and 45k variants given *N*, to the observed number of genes bearing a truncation at these variant counts. In this model the expected number of genes bearing truncation by observing *V* variants given *N* is estimated by Monte Carlo by sampling of *V* variants from a multinomial distribution specified by the neutral probabilities of the *N* genes. For a given *N* we repeated the selection of genes 100 times and for each of these cases we repeated the multinomial sampling 50 times. We then chose the value which had the minimum median residuals of the former repeats. The observed number of genes bearing a truncation at 25k, 30k, 35k, 40k and 45k variants were calculated from

random subsets of total observed variant pool. This sampling process was repeated 100 times and we report in the Results the minimum and maximum estimate of *N* out of these replications.

Characteristics of haploinsufficient genes. Functional gene sets tested included Gene Ontology (GO) terms (downloaded from Ensembl Biomart, using the attribute 'GO Term Name'). dN/dS values were assessed as described in Rausell et al. (*8*). Degree of connectivity in the protein-protein interaction network was taken from the OGEE database (*21*). Paralogs were counted using Ensembl Biomart's 'Human Paralog Ensembl Gene ID' attribute). Genes in protein complexes were taken from the GO term GO:0043234. Predicted haploinsufficiency probabilities were taken from (*12*). Genes affecting cell viability in CRISPR experiments were collected from (*17*). For the assessment of depletion or enrichment of functional gene sets in genes without a truncating variant we used logistic regression with set membership as the dependent variable. We reported the p-values of the likelihood ratio test between a null model consisting of the neutral probability of occurrence of a truncating variant (*5*), and an alternative model including the null model plus either an indicator variable or the NutVar score as predictor. We only tested gene sets with at least 10 members. We adjusted the p-values by the Benjamini-Hochberg method to correct for multiple testing.

Partitioning variants into mild and severe categories. We used GEUVADIS z-score to assess the consequence of a truncating variant on gene expression. The largest difference in the mean occurred when we used a NutVar threshold of 0.4 (mean difference in z-score between variants with NutVar score < 0.4 and $>= 0.4$ is -0.47, p-value=4.9×10^{-5}).

Table S1. Data sources.

Project name	sample n	URL
NHLBI Exome Sequencing Project	6502	http://evs.gs.washington.edu/EVS/
UK10K	2432	http://www.bristol.ac.uk/alspac/
1000 Genomes Project	1092	http://www.1000genomes.org/
Genome of the Netherlands	498	http://www.nlgenome.nl
Cohort Lausanne	426	http://www.colaus.ch/
Swiss HIV Cohort Study	500	http://www.shcs.ch/
NIEHS Environmental Genome Project	95	http://evs.gs.washington.edu/niehsExome/
Genome of J. Craig Venter	1	http://huref.jcvi.org/
Total	11546	

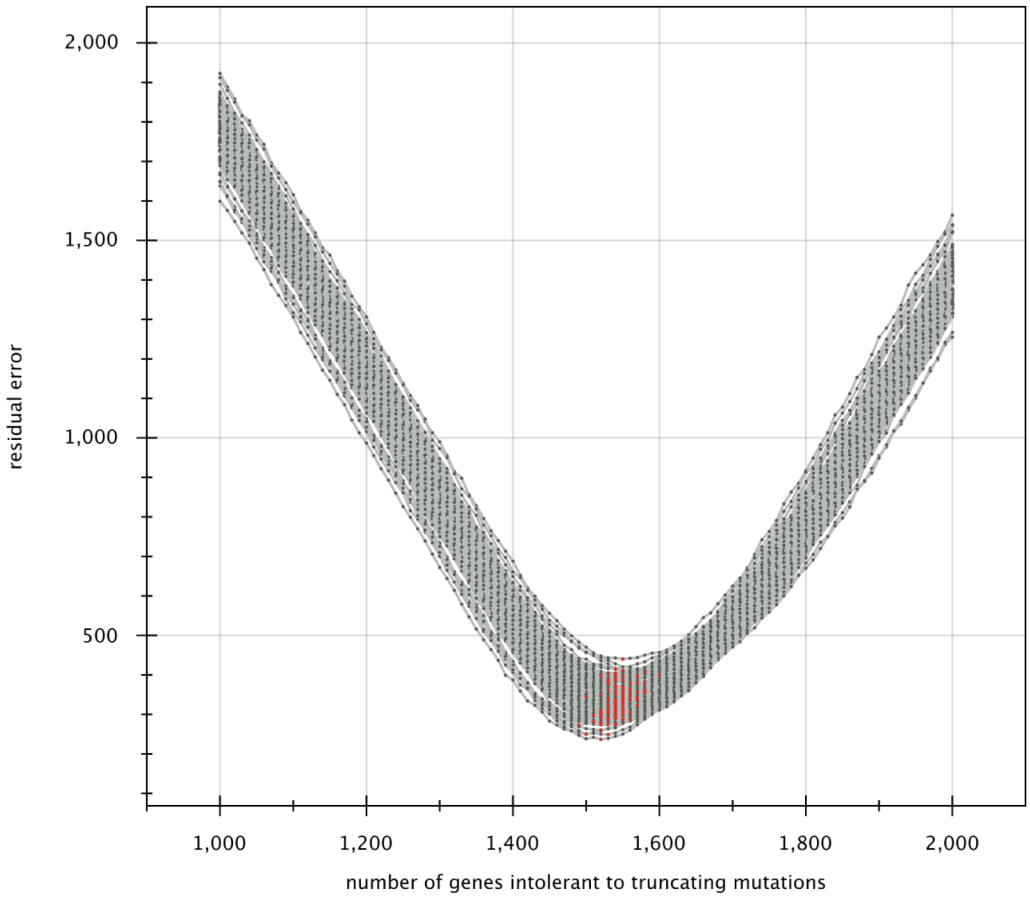

Figure S1. Residuals of the fitting procedure. Horizontal axis shows the assumed number of genes intolerant to truncating mutations. Vertical axis shows the difference in the number of genes without truncating mutations between observed and simulated data.