Estimating the selective effects of heterozygous protein-truncating variants from human exome data

Christopher A Cassa1,2,9, Donate Weghorn1,9, Daniel J Balick1,9, Daniel M Jordan3,9, David Nusinow1, Kaitlin E Samocha4,5, Anne O’Donnell-Luria4,6, Daniel G MacArthur2,4, Mark J Daly2,4, David R Beier7,8 & Shamil R Sunyaev1,2

The evolutionary cost of gene loss is a central question in genetics and has been investigated in model organisms and human cell lines1–3. In humans, tolerance of the loss of one or both functional copies of a gene is related to the gene’s causal role in disease. However, estimates of the selection and dominance coefficients in humans have been elusive. Here we analyze exome sequence data from 60,706 individuals4 to make genome-wide estimates of selection against heterozygous loss of gene function. Using this distribution of selection coefficients for heterozygous protein-truncating variants (PTVs), we provide corresponding Bayesian estimates for individual genes. We find that genes under the strongest selection are enriched in embryonic lethal mouse knockouts, Mendelian disease-associated genes, and regulators of transcription. Screening by essentiality, we find a large set of genes under strong selection that are likely to have crucial functions but have not yet been thoroughly characterized.

The dispensability of individual genes for viability has interested generations of geneticists. Approaches used in nonhuman sexual organisms to estimate selection and dominance coefficients are not applicable to humans, as they generally require cross-breeding over several generations. This has precluded reliable estimates of these crucial parameters in humans to date. The analysis of patterns of natural genetic variation in humans provides an alternative approach to estimating selection intensity and dispensability of individual genes. Despite substantial methodological progress in the ascertainment and analysis of population sequence data5–9, estimation of parameters of natural selection in humans has been complicated by genetic drift, complexities of human demographic history5,6,8,10–13, and the role of nonadditive genetic variation14–16. Additionally, naturally occurring PTVs are infrequent in the population, so data sets of thousands of individuals are underpowered for the estimation of gene dispensability in humans.

The Exome Aggregation Consortium (ExAC) data set now provides a sample sufficiently powered to directly estimate the selection that constrains the number of gene-specific PTVs in the general population4. We restricted our analysis to PTVs predicted to be consequential17, and number of chromosomes sampled, we estimated the genome-wide distribution of selective effects for heterozygous PTVs ($s_{het}$). We parameterized the distribution of selective effects using an inverse Gaussian, which is fit using maximum likelihood (Fig. 1). We then estimated individual gene selection coefficients using the posterior probability for $s_{het}$ given gene-specific values of the observed number of PTVs, number of chromosomes sampled, and estimated mutation rate (Supplementary Table 1).

Although the distribution is broad, suggesting that the effect of losing one copy of a gene is variable, the mode of the distribution corresponds to a fitness loss of $\sim 0.5$% ($s_{het} = 0.005$). Despite the large sample size, resolution to distinguish between very high selective effects is limited. There are 2,984 genes with $s_{het} > 0.1$, a result concordant with previous estimates of loss-of-function intolerance.

1Division of Genetics, Brigham and Women’s Hospital, Harvard Medical School, Boston, Massachusetts, USA. 2Program in Medical and Population Genetics, Broad Institute, Cambridge, Massachusetts, USA. 3Department of Genetic and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York, USA. 4Analytic and Translational Genetics Unit, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA. 5Division of Genetics and Genomics, Boston Children’s Hospital, Boston, Massachusetts, USA. 6Division of Genetics and Genomics, Boston Children’s Hospital, Boston, Massachusetts, USA. 7Center for Developmental Biology and Regenerative Medicine, Seattle Children’s Research Institute, Seattle, Washington, USA. 8Department of Pediatrics, University of Washington School of Medicine, Seattle, Washington, USA. 9These authors contributed equally to this work. Correspondence should be addressed to S.R.S. (ssunyaev@rics.bwh.harvard.edu) or D.R.B. (david.beier@seattlechildrens.org).

Received 13 September 2016; accepted 7 March 2017; published online 3 April 2017; doi:10.1038/ng.3831
derived from population data. Even though some genes are heavily depleted of PTVs in ExAC as compared with mutational expectation, these values suggest that heterozygote PTVs in many genes are not necessarily responsible for observable, severe clinical consequences.

Unsurprisingly, however, Mendelian disease genes had higher values. Among them, genes annotated exclusively as autosomal dominant (AD, N = 867) had significantly higher values than those annotated as autosomal recessive (AR, N = 1,482) (P = 3.14 × 10⁻⁶⁴, Mann–Whitney) (Fig. 2a,b). This suggests that it may be possible to prioritize candidate disease-associated genes identified in clinical exome–sequencing analysis using the observed mode of inheritance and values.

In 504 clinical exome cases that resulted in Mendelian diagnosis, we found a similar enrichment of cases by mode of inheritance and selection value (Fig. 2c). We found that 90.4% of novel, dominant variants were associated with heterozygous fitness loss >0.04 (Fig. 2d). Among disease variants, a cutoff of values > 0.04 provided a 96% positive predictive value for discriminating between AD and AR.

To test the generalizable utility of prioritizing candidate genes in Mendelian sequencing studies using values, we compared the overall prevalence of genes with values > 0.04 to the corresponding fraction in an independently ascertained data set of new dominant Mendelian diagnoses (Fig. 2e). The results suggest that restricting to genes with values > 0.04 would provide a threefold reduction of candidate variants, given the overall distribution of values. Thus, initial effort in clinical cases can be focused on just a few genes for functional validation, familial segregation studies, and patient matching. We summarize the classification accuracy (area under the curve (AUC) = 0.931) and generate mode of inheritance probabilities for each gene using the full set of clinical sequencing cases (Supplementary Fig. 2 and Supplementary Table 2).

Beyond mode of inheritance, we found that values helps predict phenotypic severity, age of onset, penetrance, and the fraction of de novo variants in a set of high-confidence haploinsufficient disease-associated genes (Fig. 3). In broader sets of known disease-associated genes, values estimated correlated significantly with the number of references in OMIM MorbidMap and the number of HGMD disease DM variants (Supplementary Fig. 3).

Gene-specific fitness loss values allow us to plot the distribution of selective effects for different disorders. This provides information about the breadth and severity of selection associated with various disorder groups using both well-established genes (Fig. 4a) and findings from Mendelian exome cases (Fig. 4b). Overall, genes involved in neurologic phenotypes and congenital heart disease appeared to be under more intense selection than other disorders or tolerated knockouts from a consanguineous cohort (Fig. 4c,d). Notably, genes recessive for these disorders appeared to have only partially recessive effects on fitness, so selection on heterozygotes is not negligible in these genes (Fig. 4).

In germline cancer predisposition, genes under stronger selection are enriched in individuals with cancer over those in ExAC (Supplementary Fig. 4). This suggests that genes with low values should not be prioritized in prospective genetic screening for cancer predisposition. Consistent with previous studies, we found that de novo mutations in patients with autism spectrum disorder were significantly enriched in genes under stronger selection than those identified in controls (Supplementary Fig. 5 and Supplementary Table 3).

Next, we analyzed values in the context of developmental and functional assays. In a large set of neutrally ascertained mouse knockouts (N = 2,179), mice that were null mutant for orthologous genes with higher values were enriched for embryonic lethality or subviability, whereas those with the lowest values were depleted for embryonic lethality (Mann–Whitney U-test, P = 2.95 × 10⁻²⁸) (Fig. 5a,b).

It is well known that mutations that are haploinsufficient in humans can often be well-tolerated when heterozygous in mice.
A classic example is SHH; heterozygous null mutations in this important developmental signaling gene result in holoprosencephaly.28,29 Haploinsufficiency for other genes in this signaling pathway also results in developmental defects, as with GLI3 (Pallister–Hall syndrome and Greig cephalopolysyndactyly syndrome)27–29 and GLI2 (holoprosencephaly 9)30. Interestingly, haploinsufficiency for these genes is tolerated in mouse models; mice heterozygous for null variation in the Shh signaling pathway are phenotypically normal, whereas homzygous mutant mice have defects that recapitulate features of the human syndrome.31–33 This extends to many other human developmental disorders, enabling the experimental characterization of the molecular consequences of these mutations. Thus, it is notable that homozygous null mice in orthologous genes with higher $s_{het}$ values are enriched for lethality.

High-throughput genetic analysis of cell essentiality provides an orthogonal data set for comparison with $s_{het}$. In genes putatively essential for human cell proliferation using CRISPR-based inactivation (Fig. 5c) and gene trap inactivation assays3 (Fig. 5d), essential genes were heavily enriched with high $s_{het}$ values ($P = 5.13 \times 10^{-16}$ and $4.90 \times 10^{-18}$, respectively).

Key developmental pathways were highly enriched in genes under strong selection (Fig. 6a). We also found a significant positive correlation between the number of protein–protein interactions for each gene and its $s_{het}$ value (Fig. 6b,c), identified from high-throughput mass spectrometry data Mann–Whitney U-test $P = 6.67 \times 10^{-9}$. In the context of molecular and cellular function, a set of genes with very high selective effects ($s_{het} > 0.15$; 2,072 genes) were enriched in biological process categories 'transcription regulation' (Benjamini $P = 2.99 \times 10^{-37}$), 'transcription' Benjamini association $P = 9.73 \times 10^{-38}$, and 'negative regulators of biosynthetic processes' (Supplementary Table 6). The nucleus was the most enriched cellular compartment for these genes (Benjamini-corrected association $P = 4.8 \times 10^{-96}$). The enrichment of transcription factors in these genes is consistent with literature that describes dosage dependence for enzymatic proteins and haploinsufficiency for transcriptional regulators.

Selection estimates from human PTVs provide a measure of gene dispensability that is unbiased with respect to existing knowledge. Thus, these estimates may highlight genes that have a key role in development or in maintaining core cellular functions. Many genes with high fitness costs have not previously been described in human genetics studies. Given the marked enrichment of genes with high $s_{het}$ values associated with Mendelian disorders, cell essentiality, embryonic lethality, and development, it is plausible that many genes with high $s_{het}$ values have not been previously associated with human disease because they are required for embryonic development.

We inspected genes with high $s_{het}$ values but for which there are few disease annotations or publications to determine whether they have functional and genetic features reminiscent of known genes with central roles in cell housekeeping and developmental biology. We measured the relative

---

**Figure 3** Enrichments of $s_{het}$ in known haploinsufficient disease-associated genes of high confidence (ClinGen Dosage Sensitivity Project). (a–d) High $s_{het}$ scores are associated with earlier age of onset (a, $P = 1.46 \times 10^{-2}$, Mann–Whitney U-test), a larger fraction of de novo variants (b, $P = 8 \times 10^{-5}$), high or unspecified (c, $P = 1.79 \times 10^{-2}$), and increased phenotypic severity (d, $P = 4.87 \times 10^{-3}$) in autosomal genes ($N = 127$) associated with each disease category and classification. Box plot data are as in Fig. 2b. (e) Classical annotations of genes with $s_{het}$ values in the highest decile.

**Figure 4** Distribution of $s_{het}$ values for phenotypes in known disease-associated genes and clinical cases. (a–d) Distribution of $s_{het}$ values for known Mendelian disease-associated genes (Clinical Genomic Database) annotated as AR or AD (a) and clinical exome-sequencing cases (b), compared to all tolerated knockouts in a consanguineous cohort (PROMIS)23 (c) and the $s_{het}$ distribution in all scored genes (d).
knowledge about each gene in the primary literature from Entrez and
PubMed36 using the number of gene reports connected with each manu-
script, and summed the weighted contributions across all available manus-
cripts37 (PubMed score, Online Methods). Although the PubMed score was
positively correlated with \( s_{\text{het}} \) values, a substantial number of under-
studied genes fell in the highest \( s_{\text{het}} \) decile (Supplementary Fig. 6).

We selected the 250 most cited and least cited genes within the top
\( s_{\text{het}} \) decile and compared their frequency of protein–protein inter-
actions, viability of orthologous mouse knockouts, and cell essenti-
ality assays. Genes with the fewest publications (no more than one
citation) had nearly the same number of embryonic lethal mouse
knockouts as genes with the most publications. Other assay results
were only slightly depleted in genes with the fewest publications
(Supplementary Fig. 7). These findings suggest there may be essential
developmental pathways yet to be uncovered in genes under strong
selection that lack functional or disease annotations and provides a
promising gene set for further exploration. We have created a pri-
oritized list of genes developed from functional evidence to indi-
cate the most promising candidates for future functional screening
(Supplementary Table 4).

To place our inferences in the broader evolutionary context, we used
comparable estimates from model organisms, including flies and yeast,
based on knockout competition with wild type or explicit crosses. In
yeast, analysis of a library of PTV knockouts provided a mean esti-
mate of \( s_{\text{het}} \approx 0.013 \), which is close to our inferred results (\( s_{\text{het}} \approx 0.059 \))
in humans38, given that the functional experiments excluded genes
with very high \( s \) and we have excluded genes with high cumulative allelic frequency. Estimates in flies derived from homozygous lethal
mutations that reduce viability in heterozygotes (rather than only
allele frequency. Estimates in flies derived from homozygous lethal
mutations that reduce viability in heterozygotes (rather than only
allele frequency. Estimates in flies derived from homozygous lethal
mutations that reduce viability in heterozygotes (rather than only
allele frequency. Estimates in flies derived from homozygous lethal
mutations that reduce viability in heterozygotes (rather than only
allele frequency. Estimates in flies derived from homozygous lethal
mutations that reduce viability in heterozygotes (rather than only
allele frequency. Estimates in flies derived from homozygous lethal
mutations that reduce viability in heterozygotes (rather than only
allele frequency. Estimates in flies derived from homozygous lethal
mutations that reduce viability in heterozygotes (rather than only
allele frequency. Estimates in flies derived from homozygous lethal
mutations that reduce viability in heterozygotes (rather than only
allel frequency. Estimates in flies derived from homozygous lethal
mutations that reduce viability in heterozygotes (rather than only
allel frequency. Estimates in flies derived from homozygous lethal
mutations that reduce viability in heterozygotes (rather than only
allel frequency. Estimates in flies derived from homozygous lethal
mutations that reduce viability in heterozygotes (rather than only
allel frequency. Estimates in flies derived from homozygous lethal
mutations that reduce viability in heterozygotes (rather than only
allel frequency. Estimates in flies derived from homozygous lethal
mutations that reduce viability in heterozygotes (rather than only
allel frequency. Estimates in flies derived from homozygous lethal
mutations that reduce viability in heterozygotes (rather than only
allel frequency. Estimates in flies derived from homozygous lethal
mutations that reduce viability in heterozygotes (rather than only
allel frequency. Estimates in flies derived from homozygous lethal
mutations that reduce viability in heterozygotes (rather than only
allel frequency. Estimates in flies derived from homozygous lethal
mutations that reduce viability in heterozygotes (rather than only
allel frequency. Estimates in flies derived from homozygous lethal
mutations that reduce viability in heterozygotes (rather than only
allel frequency. Estimates in flies derived from homozygous lethal
mutations that reduce viability in heterozygotes (rather than only
allel frequency. Estimates in flies derived from homozygous lethal
mutations that reduce viability in heterozygotes (rather than only
allel frequency. Estimates in flies derived from homozygous lethal
mutations that reduce viability in heterozygotes (rather than only
allel frequency. Estimates in flies derived from homozygous lethal
mutations that reduce viability in heterozygotes (rather than only
allel frequency. Estimates in flies derived from homozygous lethal
mutations that reduce viability in heterozygotes (rather than only
allel frequency. Estimates in flies derived from homozygous lethal
mutations that reduce viability in heterozygotes (rather than only
allel frequency. Estimates in flies derived from homozygous lethal
mutations that reduce viability in heterozygotes (rather than only
allel frequency. Estimates in flies derived from homozygous lethal
mutations that reduce viability in heterozygotes (rather than only
allel frequency. Estimates in flies derived from homozygous lethal
mutations that reduce viability in heterozygotes (rather than only
allel frequency. Estimates in flies derived from homozygous lethal
mutations that reduce viability in heterozygotes (rather than only
allel frequency. Estimates in flies derived from homozygous lethal
mutations that reduce viability in heterozygotes (rather than only
allel frequency. Estimates in flies derived from homozygous lethal
mutations that reduce viability in heterozygotes (rather than only
allel frequency. Estimates in flies derived from homozygous lethal
mutations that reduce viability in heterozygotes (rather than only
allel frequency. Estimates in flies derived from homozygous lethal
mutations that reduce viability in heterozygotes (rather than only
allel frequency. Estimates in flies derived from homozygous lethal
mutations that reduce viability in heterozygotes (rather than only
allel frequency. Estimates in flies derived from homozygous lethal
mutations that reduce viability in heterozygotes (rather than only
allel frequency. Estimates in flies derived from homozygous lethal
mutations that reduce viability in heterozygotes (rather than only
allel frequency. Estimates in flies derived from homozygous lethal
mutations that reduce viability in heterozygotes (rather than only
allel frequency. Estimates in flies derived from homozygous lethal
mutations that reduce viability in heterozygotes (rather than only
allel frequency. Estimates in flies derived from homozygous lethal
mutations that reduce viability in heterozygotes (rather than only
allel frequency. Estimates in flies derived from homozygous lethal
mutations that reduce viability in heterozygotes (rather than only
allel frequency. Estimates in flies derived from homozygous lethal
mutations that reduce viability in heterozygotes (rather than only
allel frequency. Estimates in flies derived from homozygous lethal
mutations that reduce viability in heterozygotes (rather than only
allel frequency. Estimates in flies derived from homozygous lethal
mutations that reduce viability in heterozygotes (rather than only
allel frequency. Estimates in flies derived from homozygous lethal
mutations that reduce viability in heterozygotes (rather than only
allel frequency. Estimates in flies derived from homozygous lethal
mutations that reduce viability in heterozygotes (rather than only
allel frequency. Estimates in flies derived from homozygous lethal
mutations that reduce viability in heterozygotes (rather than only
allel frequency. Estimates in flies derived from homozygous lethal
mutations that reduce viability in heterozygotes (rather than only
allel frequency. Estimates in flies derived from homozygous lethal
mutations that reduce viability in heterozygotes (rather than only
allel frequency. Estimates in flies derived from homozygous lethal
mutations that reduce viability in heterozygotes (rather than only
allel frequency. Estimates in flies derived from homozygous lethal
mutations that reduce viability in heterozygotes (rather than only
allel frequency. Estimates in flies derived from homozygous lethal
mutations that reduce viability in heterozygotes (rather than only
allel frequency. Estimates in flies derived from homozygous lethal
mutations that reduce viability in heterozygotes (rather than only
allel frequency. Estimates in flies derived from homozygous lethal
strongest or weakest selection. These results may be useful in Mendelian disease gene-discovery efforts and provide clinical utility in the inference of severity and mode of inheritance underlying Mendelian disease.

**METHODS**

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

**ACKNOWLEDGMENTS**

We thank I. Adzhubei, K. Karzczewski, E. Minikel, and A. Kondrashov for helpful advice. This work was supported by US National Institutes of Health (NIH) grants HG007229 (C.A.C.), GM078598 (S.R.S., D.M.J., D.J.B.), and MH101244 (S.R.S., D.W.).

**AUTHOR CONTRIBUTIONS**

Overall concept and approach conceived and developed by C.A.C., D.R.B., and S.R.S. Implementation, data analysis, and interpretation conducted by D.W., C.A.C., D.I.B., D.M.J., and D.N. Data sets and advice were provided by D.G.M., M.J.D., K.E.S., and A.O’D.-L. The article was written by C.A.C. and S.R.S. with contributions from D.W. and D.I.B. All authors read and discussed the manuscript.

**COMPETING FINANCIAL INTERESTS**

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Mukai, T., Chigusa, S.I., Mettler, L.E. & Crow, J.F. Mutation rate and dominance of genes affecting viability in *Drosophila melanogaster*. Genetics 72, 335–355 (1972).

2. Deng, H.W. & Lynch, M. Estimation of deleterious-mutation parameters in natural populations. Genetics 144, 349–360 (1996).

3. Wang, T. et al. Identification and characterization of essential genes in the human genome. Science 350, 1096–1101 (2015).

4. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature 536, 285–291 (2016).

5. Williamson, S.H. et al. Simultaneous inference of selection and population growth from patterns of variation in the human genome. Proc. Natl. Acad. Sci. USA 102, 7882–7887 (2005).

6. Boyko, A.R. et al. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet.* 4, e1000083 (2008).

7. Kryukov, G.V., Pennacchio, L.A. & Sunyaev, S.R. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am. J. Hum. Genet.* 80, 727–739 (2007).

8. Kryukov, G.V., Shpunt, A., Stamatoyannopoulos, J.A. & Sunyaev, S.R. Power of deep, all-exon resequencing for discovery of human trait genes. *Proc. Natl. Acad. Sci. USA* 106, 3871–3876 (2009).

9. Eyre-Walker, A. & Keightley, P.D. The distribution of fitness effects of new mutations. *Nat. Rev. Genet.* 8, 610–618 (2007).

10. Do, R. et al. No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. Nat. Genet. 47, 126–131 (2015).

11. Fu, W., Gittelman, R.M., Bamshad, M.J. & Akey, J.M. Characteristics of neutral and deleterious protein-coding variation among individuals and populations. *Am. J. Hum. Genet.* 95, 421–436 (2014).

12. Lohmueller, K.E. The distribution of deleterious genetic variation in human populations. *Curr. Opin. Genet. Dev.* 29, 139–146 (2014).

13. Gravel, S. When is selection effective? *Genetics* 203, 451–462 (2016).

14. Williamson, S., Fledel-Alon, A. & Bustamante, C.D. Population genetics of polymorphism and divergence: empirical test of models with arbitrary dominance. *Genetics* 168, 463–475 (2004).

15. Balick, D.J., Do, R., Cassa, C.A., Reich, D. & Sunyaev, S.R. Dominance of deleterious alleles controls the response to a population bottleneck. *PLoS Genet.* 11, e1005436 (2015).

16. Simons, Y.B., Turchin, M.C., Pritchard, J.K. & Sella, G. The deleterious mutation load is insensitive to recent population history. Nat. Genet. 46, 220–224 (2014).

17. MacArthur, D.G. et al. A systematic survey of loss-of-function variants in human protein-coding genes. *Science* 335, 823–828 (2012).

18. Samocha, K.E. et al. A framework for the interpretation of de novo mutation in human disease. Nat. Genet. 46, 944–950 (2014).

19. Francioli, L.C. et al. Genotype-wide patterns and properties of de novo mutations in humans. Nat. Genet. 47, 822–826 (2015).

20. Solomon, B.D., Nguyen, A.-D., Bear, K.A. & Wolfsberg, T.G. Clinical genomic database. *Proc. Natl. Acad. Sci. USA* 110, 9851–9855 (2013).

21. Yang, Y. et al. Molecular findings among patients referred for clinical whole-exome sequencing. *JAMA* 312, 1870–1879 (2014).

22. Lee, H. et al. Clinical exome sequencing for genetic identification of rare Mendelian disorders. *JAMA* 312, 1880–1887 (2014).

23. Saleheen, D. et al. Human knockouts in a cohort with a high rate of consanguinity. *Annu. Rev. Genomics Hum. Genet.* 15, 51–76 (2014).

24. Koscielny, G. et al. The International Mouse Phenotyping Consortium Web Portal, a unified point of access for knockout mice and related phenotyping data. *Nucleic Acids Res.* 42, D802–D809 (2014).

25. Georgi, B., Voight, B.F. & Buican, M. From mouse to human: evolutionary genomics analysis of human orthologs of essential genes. *PLoS Genet.* 9, e1003484 (2013).

26. Roessler, E. et al. Mutations in the human Sonic Hedgehog gene cause holoprosencephaly. *Nat. Genet.* 14, 357–360 (1996).

27. Kang, S., Graham, J.M., Olney, A.H. & Biesecker, L.G. *GLI3* frameshift mutations cause autosomal dominant Pallister–Hall syndrome. *Nat. Genet.* 15, 266–268 (1997).

28. Vorkamp, A., Gessler, M. & Grzeschik, K.H. *GLI3* zinc-finger gene interrupted by translocations in Greg syndrome families. *Nature* 352, 539–540 (1991).

29. Wild, A. et al. Point mutations in human *GLI3* cause Greg syndrome. *Hum. Mol. Genet.* 6, 1979–1984 (1997).

30. Roessler, E. et al. Loss-of-function mutations in the human *GLI2* gene are associated with pituitary anomalies and holoprosencephaly-like features. *Proc. Natl. Acad. Sci. USA* 100, 13424–13429 (2003).

31. Chiang, C. et al. Cyclopia and defective axial patterning in mice lacking Sonic hedgehog gene function. *Nature* 383, 407–413 (1996).

32. Hui, C.C. & Joyner, A.L. A mouse model of Greg cephalopolysyndactyly syndrome: the extra-toes*1* mutation contains an intragenic deletion of the *GLI3* gene. *Nat. Genet.* 3, 241–246 (1993).

33. Mo, R. et al. Specific and redundant functions of *GLI2* and *GLI3* zinc finger genes in skeletal patterning and development. *Development* 124, 113–123 (1997).

34. Huang, D.W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57 (2008).

35. Seidman, J.G. & Seidman, C. Transcription factor haploinsufficiency: when half a loaf is not enough. *J. Clin. Invest.* 109, 451–455 (2002).

36. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 41, D8–D20 (2013).

37. Raychaudhuri, S. et al. Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet.* 5, e1000534 (2009).

38. Argayelles, M. & Whitlock, M.C. Inferences about the distribution of dominance from yeast gene knockout data. *Genetics* 187, 553–566 (2011).

39. Simmons, M.J. & Crow, J.F. Mutations affecting fitness in *Drosophila* populations. *Annu. Rev. Genet.* 11, 49–78 (1977).

40. Wright, S. Evolution in Mendelian populations. *Bull. Math. Biol.* 52, 241–295 (1990).

41. Petrovski, S. et al. The intolerance of regulatory sequence to genetic variation predicts gene dosage sensitivity. *PLoS Genet.* 11, e1005492 (2015).

42. Kiezun, A. et al. Exome sequencing and the genetic basis of complex traits. *Nat. Genet.* 44, 623–630 (2012).
ONLINE METHODS

Model of deterministic mutation–selection balance. For most genes, protein-truncating alleles are both individually and collectively rare. For genes in which they are collectively rare, estimation of the selective effect against heterozygous PTVs ($s_{het}$) can be greatly simplified. We model each gene as a single biallelic locus with cumulative frequency

$$X = \sum_{j} x_j,$$

where the sum is over PTV sites $j$. This is motivated by the simplifying assumption of identical selection coefficients for all PTVs within a gene and the observation that the frequency of the vast majority of PTVs is extremely low such that the occurrence of multiple variable sites within a gene on a single haplotype is also extremely low ($2N_s x_j < 1$ for sample size $N$). Moreover, multiple PTVs in a gene in an individual would be functionally equivalent to a single PTV resulting in a loss-of-function state.

Then, for each gene, the cumulative allele frequency $X$ is influenced by incoming mutation, selection, and the random reassortment of alleles (genetic drift). When selection is strong, $s >> 2.5 \times 10^{-8}$ (i.e., when $4N_s s >> 1$, with effective population size $N_e = 10^5$), drift is much smaller than the contribution of selection. Furthermore, the strength of genetic drift is weakest for genes at low frequencies: for a variant with cumulative frequency of $X = 0.001$ the expected frequency change due to drift is only $(\Delta X)^2 - X/4N_e = 2.5 \times 10^{-8}$ per generation. Notably, at the locus level, assuming $X << 1$, the drift contribution is also much smaller than the mutational influx. Hence, under strong selection and for small allele frequencies, the expected cumulative contribution is also much smaller than the mutational influx. Hence, under strong selection and for small allele frequencies, the expected cumulative frequency of PTVs is determined by the equilibrium between the influx of de novo mutations (estimated to increase the cumulative frequency by an average $1.4 \times 10^{-8}$ per locus per generation by mutational model) and the outflow due to natural selection.

In the presence of selection on both heterozygotes and homozygotes and ignoring back mutations, the dynamics of $X$ are captured by

$$\partial_t X = -s_{het} X (1 - X) - s_{hom} X^2 (1 - X) + U$$

(1)

Here $U$ represents the PTV mutation rate at the gene locus per individual per generation, and $s_{het} = hs > 0$ and $s_{hom} = s > 0$ represent the strength of negative selection against PTV heterozygotes and homozygotes, respectively. We note that compound heterozygotes (with a single PTV on each chromosome) are treated as homozygotes under the biallelic assumption. Provided $X << 1$, the case for PTVs under strong selection $(2N_s s >> 1)$, this equation simplifies dramatically:

$$\partial_t X = -s_{het} X - s_{hom} X^2 + U$$

(2)

Because $X << 1$, selection against heterozygotes (the linear term) generally also dominates over selection against homozygotes (the quadratic term), provided $s_{het}/s_{hom} >> X$. This is violated only in cases of extreme recessivity (where the dominance coefficient $h << 0.001$), but even in that case the expected cumulative frequency of PTVs in essential genes is unlikely to exceed 0.001 (the characteristic $X$ in the completely recessive case is $\sqrt{U/s} \sim 10^{-3}$ when $s << 1$; see simulations in Supplementary Fig. 1). The strong selection regime thus corresponds to mutation–selection balance in the heterozygote state of a PTV mutation. In our model, we do not assume that selection acts exclusively on heterozygotes but aim at estimating only fitness loss due to the lack of one functional copy of a gene. Even in the case of strong selection against homozygotes, the population frequency is primarily controlled by efficient selection against heterozygotes. Notably, there is no dependence on demography or population size in this regime, as the contribution from drift vanishes because selection drives alleles out of the population efficiently and on very short time scales. Classic papers by Li44,45 and Maruyama55,46 showed that relevant time scales are short, even in the case of exponential expansion, because individual deleterious alleles are predominantly recent, with an allelic age on the order of $1/s$. Current estimates of recent population histories for most populations included in the ExAC data set suggest that 4Ns safely exceeds 1. Even if individual alleles are subject to stochastic drift, this effect is mitigated by the aggregation of variants on the gene level. One possible concern is the inclusion of individuals with Finnish ancestry, as this population underwent an intense, relatively recent bottleneck. We address this population explicitly using forward simulations and by removing them from our analysis to show no significant deviation from our initial estimates in their absence (below).

From equation (2) follows that for a population sample of size $N$ chromosomes, sample allele counts

$$n = NX = N \sum_{j} x_j$$

are expected to be Poisson distributed around the expectation given by:

$$E(n) = \frac{NU}{s_{het}}$$

(3)

Generally, genes under the strongest and weakest selection are expected to have greater estimation uncertainty, as the resolution to estimate $s_{het}$ deteriorates when variants are so common that they may be controlled not only by heterozygote selection but also by drift or complex demography. However, the overwhelming majority of genes conform to our assumptions of cumulative PTV allele frequency not exceeding 0.001. Despite issues such as the admixture of populations, consanguineous samples in ExAC47, and the Wahlund effect, very few genes (1,201 of 17,199 covered genes) have higher estimated cumulative allele frequencies $X$, which we restrict from the estimation procedure. On the other end of the spectrum, genes under strong selection may lack PTVs by chance alone in ExAC, which limits the ability to distinguish between large selective effects.

Population genetics simulations of model assumptions. To validate the assumption that estimates of selection can be made under mutation–selection balance independently of demography or population size for variants under sufficiently strong selection, we used SLiM 2.0 to conduct forward population genetics simulations47. We ran 10,000 replicates of each with selection coefficients of $-5 \times 10^{-2}$, $-5 \times 10^{-3}$, $-5 \times 10^{-4}$, $-5 \times 10^{-5}$, and $-5 \times 10^{-6}$ through a realistic demography derived from previously published histories for African, non-Finnish European, and Finnish populations48,49 (Supplementary Fig. 1). We compare the theoretical expectation of cumulative allele frequency ($U/s_{het}$ (equation (3))) with the simulated cumulative allele frequency. We do this in three populations (African, non-Finnish European, and Finnish), plus a ‘combined’ population that includes pooled site frequency spectra from all three populations in proportions represented in the ExAC data set. The simulations support our assumption of mutation–selection balance in the strong selection regime ($s_{het} \geq 10^{-4}$), which appears to be appropriate for PTVs. This is true for all three populations examined and for the combined population, demonstrating that this assumption is robust to differences in the strength of drift due the distinct demographic histories of included human populations.

All simulations had a length of 100 bp, mutation rate of $2 \times 10^{-8}$ per generation per base pair, and recombination rate of $1 \times 10^{-5}$ per generation per base pair. The high recombination rate was chosen to simulate largely unlinked sites, as we are simulating PTVs that are infrequent enough that they are expected not to be in linkage with other PTVs in the same gene.

Data set for $s_{het}$ estimation. In this analysis, we used ExAC data set version 0.3, a set of jointly called exomes from 60,706 individuals ascertained with no known severe, early-onset Mendelian disorders. The mean coverage depth was calculated for each gene (canonical transcript from Ensembl v75, GENCODE v19) in the ExAC data set (mean 57.75; s.d. 20.96). Genes with average coverage depth of at least $30 \times$ were used in further analysis ($N = 17,199$). Single nucleotide substitution variants annotated as PASS quality with predicted functional effects in the canonical transcript of ‘stop_gained’, ‘splice_donor’, or ‘splice_acceptor’ (as annotated by Variant Effect Predictor) were included in the analysis. Variants such as indels, in-frame mutations, and frameshift variants were excluded from this analysis, as many of these variants may have annotation issues or may not be functionally impactful. Along the same lines, we are mindful that not all PTVs will result in complete loss of gene function, due to alternative transcripts or nonsense mediated decay. To address this, variants were filtered using LOFTEE and restricted to those predicted with high confidence to have consequences in the canonical transcript.
For each of the 17,199 genes, we have observable values for \((n, U, N)\), where \(n\) denotes the total number of observed PTV alleles in the population sample of \(N\) chromosomes covered in the gene, and \(U\) the PTV mutation rate across the canonical gene transcript from a mutational model. Values of \(U\) for each gene from Samocha et al.\(^{18}\) were used along with the number of well-covered chromosomes \(N\) in each gene to generate the null mutational expectation of neutral evolution, \(NU\). Incorrectly specified values from this mutational model could alter estimates of selection for individual genes, as higher estimates of selection are made in genes with greater depletions from the null expectation model. Our inference of selection coefficients relies on the assumption that the cumulative population frequency of PTV mutations, \(X\), is small owing to strong negative selection, so genes with \(X = n/N > 0.001\) are omitted from the analysis, leaving 15,998 genes.

**Estimation of \(P(\nu)\).** A genome-wide ensemble of observed \((n)\) and expected \((NU = v)\) genic PTV counts enables the inference of the distribution of heterozygous loss-of-function fitness effects, \(P(\nu)\), which underlies the evolutionary dynamics of this class of mutations. We estimate the parameters \((\alpha, \beta)\) of this distribution by fitting the observed distribution of PTV counts across genes:

\[
P(n|\alpha, \beta; v) = \int P(n|\nu; v)P(\nu|\alpha, \beta)\,d\nu
\]

For a given gene under negative selection, PTV mutations are rare events, such that we expect a Poisson distribution for the likelihood of the observed number of PTVs, \(P(n|\nu; v) = \text{Pois}(n; \lambda)\), where \(\lambda = v/\nu\) (equation (3)). We parameterize by using the functional form of an inverse Gaussian distribution, i.e., \(P(\nu|\alpha, \beta) = \text{IG}(\nu|\alpha, \beta)\), with mean \(\alpha\) and shape parameter \(\beta\), so equation (4) becomes:

\[
P(n|\alpha, \beta; v) = \left[\text{Pois}(n; \lambda = v/\nu)\right]\text{IG}(\nu|\alpha, \beta)\,d\nu
\]

\[
\frac{1}{n!}\alpha^{\frac{n}{2}}\beta^{\frac{1-n}{2}}\left(\frac{\nu}{\alpha}\right)^n\left(\frac{\beta}{\beta + 2v}\right)^{\frac{1}{2}n}\frac{\Gamma_K\left(\frac{\beta + 2v}{\beta}\right)}{\Gamma_K\left(\frac{\beta}{\beta}\right)}
\]

where \(K_\nu(z)\) is the modified Bessel function of the second kind. To estimate parameters of the distribution of selection coefficients, \(P(\nu|\alpha, \beta)\), we fit equation (5) to the observed distribution of PTV counts, \(Q(n)\), by maximizing the log-likelihood

\[
\log \left[\mathcal{L}(\alpha, \beta | \{n_i\})\right] = \log \sum_{j=1}^{G} P(n_j | \alpha, \beta; v_j)
\]

on the regime \(\alpha \in [10^{-5}, 2]\) and \(\beta \in [10^{-4}, 2]\), where \(G\) is the number of genes. In order to account for a slightly positive correlation between the mutation rate and selection strength (Supplementary Fig. 8), we separately perform the fit on U terciles of the data set and combine the results in a mixture distribution with equal weights. The mean mutation rates in the three terciles are \(\bar{\alpha}_i = 4.6 \times 10^{-7}\), \(\bar{\alpha}_2 = 1.1 \times 10^{-6}\), and \(\bar{\alpha}_3 = 2.6 \times 10^{-6}\). We estimate \(\hat{\alpha}_i, \hat{\beta}_i\) = \((0.057 \pm 0.000, 0.052 \pm 0.0000), \((0.046 \pm 0.005, 0.087 \pm 0.0004)\), and \((0.074 \pm 0.005, 0.0160 \pm 0.0005)\), with error margins denoting 2 s.d. from 100 bootstrapping replicates of the set of -5,333 genes in each tercile. This error estimate is intended to quantify the effect of the sampling noise in the data set on the parameter inference while local mutation rate estimates are assumed fixed. The resulting fitted distributions of counts are shown in Supplementary Figure 9 together with the corresponding \(Q(n)\), while Figure 1 shows the inferred projection of \(P(\nu|\alpha, \beta)\) motivated by the shape of the empirical distribution of the naive estimator \(v/n\) (given by a simple inversion of equation (3)). We also compared the log-likelihood of the fit to \(Q(n)\) obtained with this model to that obtained from two other two-parameter distributions, \(\text{Gamma}\) and \(\text{InvGamma}\), and chose the model with the highest likelihood, which is \(\text{Gamma}\) in each gene to generate the null mutational expectation of neutral evolution, \(NU\). We found high concordance between the overall distribution generated using all ExAC samples and NFE-specific estimates. We also separately removed Finnish individuals from the estimation of the distribution of selection coefficients and found very high concordance between estimates made using all ExAC samples and ExAC without Finnish individuals (Supplementary Fig. 11). These analyses demonstrate that the model is robust to concerns about recent demographic history in Finnish individuals, supporting the validity of the deterministic approximation. We cannot completely rule out the possibility that other included populations may have issues related to complexities of their recent demographic history.

**Inference of \(\nu\) on individual genes.** From the inferred distributions \(P(\nu|\alpha, \beta)\) in each tercile \(i\) of the mutation rate \(U\), we construct a per-gene estimator of \(\nu\) for genes in the tercile using the posterior probability given \(n\), which mitigates the stochasticity of the observed PTV count:

\[
P(\nu|\alpha, \beta) = \frac{P(n_i|\nu, v_i)}{P(n_i|\alpha, \beta)}P(\nu|\alpha, \beta)
\]

where the denominator is given by equation (5). Supplementary Table 1 provides the mean values derived from these posterior probabilities for each gene.

**Predicted mode of inheritance in clinical exome cases.** We trained a naive Bayes classifier to predict the mode of inheritance in a set of solved clinical exome-sequencing cases from Baylor College of Medicine \((N = 283 cases)\)\(^{21}\) and UCLA\(^{22}\) \((N = 176 cases)\). Using data from UCLA as the training data set, we were able to cross-predict the mode of inheritance in separately ascertained Baylor cases with classification accuracy of 88.0%, sensitivity of 86.1%, specificity of 90.2%, and an AUC of 0.931. Genes that were related to diagnosis in both clinics (overlapping genes) were removed from the larger Baylor set (Supplementary Fig. 2).

Using a logistic regression based on the full set of cases from Baylor and UCLA, we generated predictions for all 15,998 genes where there is a \(\nu\) value (Supplementary Table 4).

**Mouse knockout comparative analysis.** We reviewed mouse knockout enrichments from the full set of mouse knockouts from a neutrally ascertained mouse knockout screen \((N = 2,179 genes)\) generated by the International Mouse Phenotyping Consortium\(^{23}\). Genes were classified as viable, sub-viable, or lethal on the basis of the results for the assay.

**PubMed gene score and enrichment analysis.** We developed a score to estimate the relative importance of each gene in the published medical and scientific literature. First, we connected literature from Entrez, which included both PubMed citations and references to Entrez genes. We assigned a weight to each article referring to four genes would receive a weighted article score. These scores ranged from 4,672 (for p53) to 0.0001. Finally, we assigned each gene a score which was the sum of the weight of 1/4. Next, we focused on genes that are estimated to be under very strong selection but that lack functional or clinical annotations. In the top decile of \(\nu\) values, we separated the top 250 and bottom 250 genes by PubMed score. We then annotated each of these with unbiased genome-wide assays, including the number of protein-protein interactions (as determined by a genome-wide mass spectrometry assay)\(^{26}\), whether each gene is determined to be cell-essential in genome-wide CRISPR and gene trap assays\(^3\), and whether there is a mouse knockout in the neutrally ascertained orthologous nonviable mouse knockout\(^{24}\). To limit the number of genes with incorrect \(\nu\) estimates in this set of 500 genes, we pre-filtered any genes with only a single exon, as they may be enriched for recent pseudogenes, and also removed any olfactory, mucin, and zinc-finger proteins. Genes are also ordered by a heuristic score based on each of the forms of functional evidence in the analysis.
**Functional enrichment analysis.** We inspected the functional annotations related to approximately the top 10% of selectively disadvantageous genes (with $s_{het} > 0.15$, $N = 2,072$ genes) that were successfully mapped using Database for Annotation, Visualization, and Integrated Discovery (DAVID) version 6.7 (ref. 34). Separately, two other cutoffs ($s_{het} > 0.25$, $N = 897$ genes and $s_{het} > 0.5$, $N = 32$ genes) were also tested and similar results were identified.

Using DAVID, we identified functional annotation terms and keywords that were enriched and clustered. Functional annotation terms were generated using the Functional Annotation tool, which includes protein information resource keywords, Gene Ontology (GO) terms, biological processes and pathways, and protein domains. Using the default settings (Count 2 and EASE 0.1), 247 statistically significant (Bonferroni-corrected) terms were identified and are included in **Supplementary Table 5**.

Using the DAVID functional annotation clustering feature, we identified clusters using the same set of 2,072 genes with the default settings. The first annotation cluster includes core, essential cellular components including the nuclear lumen, nucleoplasm, organelle lumen (enrichment score 32.63), and the second includes transcription regulation and transcription factor activity (enrichment score 27.94), detailed in **Supplementary Table 6**.

**Data availability.** The authors declare that the data supporting the findings of this study are available within the paper and its supplementary information files. All original population frequency data are available through the ExAC Aggregation Consortium [http://exac.broadinstitute.org/](http://exac.broadinstitute.org/). Updated selection estimates are available at [http://genetics.bwh.harvard.edu/genescores/](http://genetics.bwh.harvard.edu/genescores/).

43. Li, W.H. & Nei, M. Total number of individuals affected by a single deleterious mutation in a finite population. *Am. J. Hum. Genet.* **24**, 667–679 (1972).
44. Li, W.H. The first arrival time and mean age of a deleterious mutant gene in a finite population. *Am. J. Hum. Genet.* **27**, 274–286 (1975).
45. Maruyama, T. The age of a rare mutant gene in a large population. *Am. J. Hum. Genet.* **26**, 669–673 (1974).
46. Maruyama, T. The age of an allele in a finite population. *Genet. Res.* **23**, 137–143 (1974).
47. Messer, P.W. SLiM: simulating evolution with selection and linkage. *Genetics* **194**, 1037–1039 (2013).
48. Tennessen, J.A. et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64–69 (2012).
49. Wang, S.R. et al. Simulation of Finnish population history, guided by empirical genetic data, to assess power of rare-variant tests in Finland. *Am. J. Hum. Genet.* **94**, 710–720 (2014).
50. Huttiin, E.L. et al. The BioPlex Network: a systematic exploration of the human interactome. *Cell* **162**, 425–440 (2015).
51. Ayadi, A. et al. Mouse large-scale phenotyping initiatives: overview of the European Mouse Disease Clinic (EUMODIC) and of the Wellcome Trust Sanger Institute Mouse Genetics Project. *Mamm. Genome* **23**, 600–610 (2012).