Evaluating drug targets through human loss-of-function genetic variation

https://doi.org/10.1038/s41586-020-2267-z

Received: 25 January 2019

Accepted: 10 February 2020

Published online: 27 May 2020

Open access

Check for updates

Eric Vallabh Minikel[1,2,3,4,5,6,7,8]✉, Konrad J. Karczewski[1,4], Hilary C. Martin[9], Beryl B. Cummings[1,4,5], Nicola Whiffin[1,10], Daniel Rhodes[11], Jessica Alföldi[1,4], Richard C. Trembath[12], David A. van Heel[13], Mark J. Daly[1,4], Genome Aggregation Database Production Team*, Genome Aggregation Database Consortium*, Stuart L. Schreiber[3,14] & Daniel G. MacArthur[1,4,153,154]✉

Naturally occurring human genetic variants that are predicted to inactivate protein-coding genes provide an in vivo model of human gene inactivation that complements knockout studies in cells and model organisms. Here we report three key findings regarding the assessment of candidate drug targets using human loss-of-function variants. First, even essential genes, in which loss-of-function variants are not tolerated, can be highly successful as targets of inhibitory drugs. Second, in most genes, loss-of-function variants are sufficiently rare that genotype-based ascertainment of homozygous or compound heterozygous 'knockout' humans will await sample sizes that are approximately 1,000 times those presently available, unless recruitment focuses on consanguineous individuals. Third, automated variant annotation and filtering are powerful, but manual curation remains crucial for removing artefacts, and is a prerequisite for recall-by-genotype efforts. Our results provide a roadmap for human knockout studies and should guide the interpretation of loss-of-function variants in drug development.

Human genetics is an increasingly crucial source of evidence guiding the selection of new targets for drug discovery[1]. Most new clinical drug candidates eventually fail for lack of efficacy[2], and although in vitro, cell culture and animal model systems can provide preclinical evidence that the compound engages its target, too often the target itself is not causally related to human disease[1]. Candidates targeting genes with human genetic evidence for disease causality are more likely to reach approval[3,4], and identification of humans with loss-of-function (LoF) variants, particularly two-hit (homozygous or compound heterozygous) genotypes, has, for several genes, correctly predicted the safety and phenotypic effect of pharmacological inhibition[5]. Although these examples demonstrate the value of human genetics in drug development, important questions remain regarding strategies for identifying individuals with LoF variants in a gene of interest, interpretation of the frequency—or lack—of such individuals, and whether it is wise to pharmacologically target a gene in which LoF variants are associated with a deleterious phenotype.

Public databases of human genetic variation have catalogued predicted loss-of-function (pLoF) variants—nonsense, essential splice site, and frameshift variants expected to result in a non-functional allele. This presents an opportunity to study the effects of pLoF variation in genes of interest and to identify individuals with pLoF genotypes to understand gene function or disease biology, or to assess potential for therapeutic targeting. Although many variants initially annotated as pLoF do not, in fact, abolish gene function[6], rigorous automated filtering can remove common error modes[7]. True LoF variants are generally rare, and show important differences between outbred, bottlenecked[8] and consanguineous[9] populations[6,10]. Counting the number of distinct pLoF variants in each gene in a population sample allows the quantification of gene essentiality in humans through a metric named 'constraint'[10–13]. Specifically, the rate at which de novo pLoF mutations arise in each gene is predicted on the basis of rates of DNA mutation[10,12], and the ratio of the count of pLoF variants observed in a database to the number expected based on mutation rates—obs/exp, or constraint score—measures how strongly purifying natural selection has removed such variants from the population. The annotation of pLoF variants remains imperfect, and continued improvements are being made[14], but constraint usefully measures gene essentiality, as demonstrated by agreement with cell culture and mouse knockout experiments[7], by overlap with human disease genes[7,10] and genes depleted for structural variation[15], and by the power of constraint to enrich for deleterious variants in neurodevelopmental disorders[7,16].

[1]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [2]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [3]Chemical Biology and Therapeutics Science Program, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [4]Analytical and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA. [5]Program in Biological and Biomedical Sciences, Harvard Medical School, Boston, MA, USA. [6]Henry and Allison McCance Center for Brain Health, Massachusetts General Hospital, Boston, MA, USA. [7]Department of Neurology, Massachusetts General Hospital, Boston, MA, USA. [8]Prion Alliance, Cambridge, MA, USA. [9]Wellcome Sanger Institute, Hinxton, Cambridgeshire, UK. [10]National Heart and Lung Institute and MRC London Institute of Medical Sciences, Imperial College London, London, UK. [11]Centre for Translational Bioinformatics, William Harvey Research Institute, Barts and the London School of Medicine and Dentistry, Queen Mary University of London and Barts Health NHS Trust, London, UK. [12]School of Basic and Medical Biosciences, Faculty of Life Sciences and Medicine, King's College London, London, UK. [13]Blizard Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, UK. [14]Department of Chemistry & Chemical Biology, Harvard University, Cambridge, MA, USA. [153]Present address: Centre for Population Genomics, Garvan Institute of Medical Research and UNSW Sydney, Sydney, Australia. [154]Present address: Centre for Population Genomics, Murdoch Children's Research Institute, Melbourne, Australia. *Lists of authors and their affiliations appear at the end of the paper. ✉e-mail: eminikel@broadinstitute.org; d.macarthur@garvan.org.au

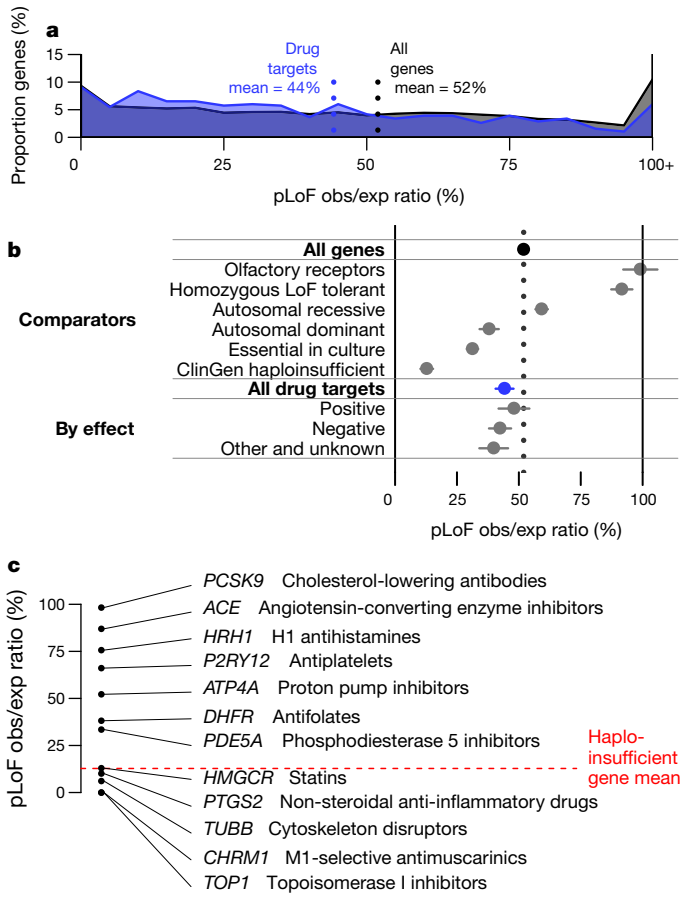


Fig. 1 | pLoF constraint in drug targets. a, Histogram of pLoF obs/exp values for all genes (black, $n = 17{,}604$) versus drug targets (blue, $n = 383$). **b**, Forest plot of means (dots) and 95% confidence intervals of the mean (line segments), for constraint in the indicated gene sets (data sources and n values in Extended Data Table 1). For drug effect, 'positive' indicates agonist, activator or inducer, whereas negative indicates antagonist, inhibitor or suppressor, for example. **c**, Examples of drug targets and corresponding drug classes from across the constraint spectrum. Details in Extended Data Table 2.

Building on these insights, here we leverage pLoF variation in the Genome Aggregation Database (gnomAD)[7] v2 dataset of 141,456 individuals to answer open questions in the interpretation of human pLoF variation in disease biology and drug development.

Constraint in human drug targets

We compared constraint in the targets of approved drugs extracted from DrugBank[17] ($n = 383$) versus all protein-coding genes ($n = 17{,}604$). Drug targets were, on average, just slightly more constrained than all genes (mean 44% versus 52%, nominal $P = 0.00028$, $D = 0.11$, two-sided Kolmogorov–Smirnov test), but the two gene sets had a qualitatively similar distribution of scores, ranging from intensely constrained (0% obs/exp) to not at all constrained (≥100% obs/exp) (Fig. 1a). Constraint scores showed clear divergence between categories of genes (Extended Data Table 1) expected to be more or less tolerant of inactivation (Fig. 1b), as previously reported[7,10], validating the usefulness of constraint as a measure of gene essentiality. Nonetheless, when drug targets were stratified by drug effect (Fig. 1b), modality, or indication (Extended Data Fig. 1), no statistically significant differences between subsets of drug targets were observed.

The slightly but significantly lower obs/exp value among drug targets may superficially appear to provide evidence that constrained genes make superior drug targets. Stratification of drug targets by protein family, human disease association, and tissue expression, however, argues against this interpretation. Drug targets are strongly enriched for a few canonically 'druggable' protein families, for genes known to be involved in human disease, and for genes with tissue-restricted expression; each of these properties is in turn correlated with either significantly stronger or weaker constraint (Extended Data Fig. 2). Although controlling for these correlations does not abolish the trend of stronger constraint among drug targets, the correlation of so many observed variables with the status of a gene as a drug target argues that many unobserved variables probably also confound interpretation of the lower mean obs/exp value among drug targets.

The overall constraint distribution of drug targets (Fig. 1a) also argues against the view that a gene in which LoF is associated with a deleterious phenotype cannot be successfully targeted. Indeed, 19% of drug targets ($n = 73$), including 52 targets of inhibitors, antagonists or other 'negative' drugs, have lower obs/exp values than the average (12.8%) for genes known to cause severe diseases of haploinsufficiency[18] (ClinGen level 3). To determine whether this finding could be explained by a particular class or subset of drugs, we examined constraint in several well-known example drug targets (Fig. 1c, Extended Data Table 2). Some heavily constrained genes are targets of cytotoxic chemotherapy agents such as topoisomerase inhibitors or cytoskeleton disruptors, a set of drugs intuitively expected to target essential genes. However, genes with near-complete selection against pLoF variants also include *HMGCR* and *PTGS2*, the targets of highly successful, chronically used inhibitors—statins and aspirin.

These human in vivo data further the evidence from other species and models that essential genes can be good drug targets. Homozygous knockout of *Hmgcr* and *Ptgs2* are lethal in mice[19–21]. Drug targets exhibit higher inter-species conservation than other genes[22]. Targets of negative drugs include 14 genes with lethal heterozygous knockout mouse phenotypes reported[23] and 6 reported as essential in human cell culture[24].

Prospects for finding human 'knockouts'

Athough constraint alone is not adequate to nominate or exclude drug targets, the study of individuals with single hit (heterozygous) or two-hit ('knockout') LoF genotypes in a gene of interest can be highly informative about the biological effect of engaging that target[5]. To assess prospects for ascertaining knockout individuals, we computed the cumulative allele frequency (CAF) of pLoF variants in each gene (Methods), and then used this to estimate the expected frequency of two-hit individuals under different population structures (Fig. 2) in the absence of natural selection.

Whereas gnomAD is now large enough to include at least one pLoF heterozygote for most (15,317 out of 19,194; 79.8%) genes, ascertainment of total knockout individuals in outbred populations will require 1,000-fold larger sample sizes for most genes: the median expected two-hit frequency of a gene is just six per billion (Fig. 2a). Even if every human on Earth were sequenced, there are 4,728 genes (24.6%) for which identification of even one two-hit individual would not be expected in outbred populations. Intuitively, because the sample size of gnomAD today is larger than the square root of the world population, variants so far seen in zero or only a few heterozygous individuals are not likely to ever be seen in a homozygous state in outbred populations, except where variants prove common in populations not yet well-sampled by gnomAD.

Because population bottlenecks can result in very rare variants present in a founder rising to an unusually high frequency, we also considered knockout discovery in bottlenecked populations, using Finnish individuals in gnomAD as an example[8]. Although this population structure can enable well-powered association studies for the small fraction of genes in which pLoF variants drifted to high frequency due to the bottleneck, overall, identification of two-hit pLoF individuals

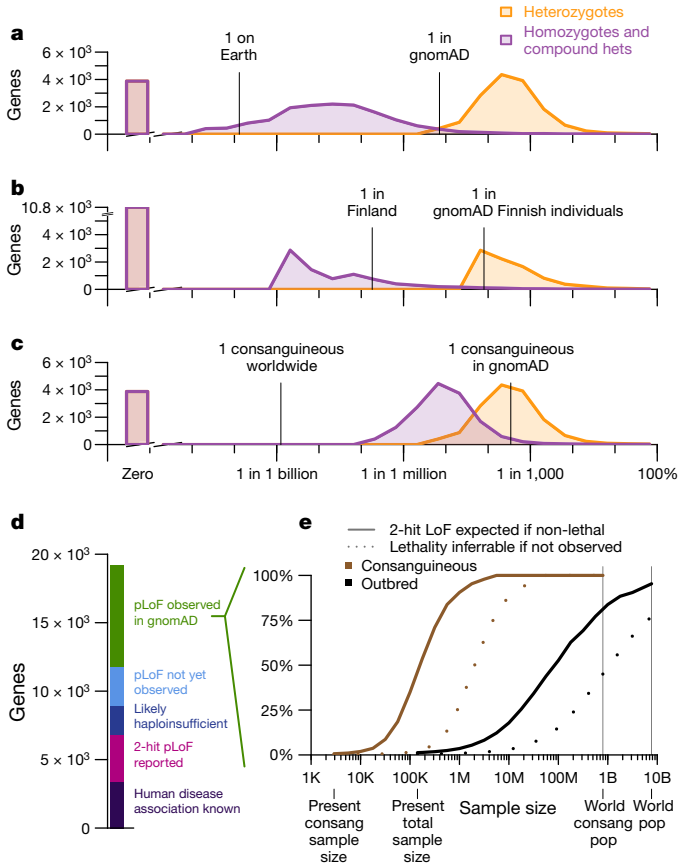


Fig. 2 | Prospects for discovery of human knockouts. a–**c**, Histograms (**a**–**c**): genes by expected heterozygote frequency (orange), and two-hit homozygote and compound heterozygote frequency (purple). **a**, Outbred populations. **b**, Finnish individuals; an example of a bottlenecked population. **c**, Consanguineous individuals. **d**, Current status of pLoF or disease association discovery for all protein-coding genes. **e**, Projected sample sizes required for discovery of two-hit individuals (solid lines) and for statistical inference that a two-hit genotype is lethal if no such individuals are observed (dashed lines), for 'pLoF observed in gnomAD' genes (**d**) for consanguineous and outbred individuals.

for a pre-specified gene of interest appears equally or more difficult in Finnish individuals than in outbred populations (Fig. 2b, Extended Data Fig. 3), because rare variants not present in a founder have been effectively removed from the population.

In consanguineous individuals, parental relatedness greatly increases the frequency of homozygous pLoF genotypes. The $n = 2{,}912$ individuals in the East London Genes & Health (ELGH) cohort[25] who report having parents who are second cousins or closer have on average 5.8% of their genomes autozygous. Here, the expected frequency of two-hit individuals is many times higher than in outbred populations, at five per million for the median gene (Fig. 2c).

These projections allow us to draft a roadmap for discovery of human knockouts across 19,194 genes (Fig. 2d, e). Online Mendelian Inheritance in Man (OMIM) already describes human disease association for 3,367 genes (18%), although the discovery of LoF individuals in population databases will still be valuable for assessing penetrance and identifying LoF syndromes of known gain-of-function genes. Another 3,421 genes (18%) without known human disease association have two-hit pLoF genotypes reported in gnomAD[7], ELGH[26], PROMIS[27], deCODE[28] or UK Biobank[29], which suggests that this genotype may be tolerated. An additional 2,190 genes (11%) appear intolerant of heterozygous inactivation (pLI score > 0.9) in gnomAD—a set expected to be enriched for genes with severe heterozygous and lethal homozygous LoF phenotypes. Another 2,781 genes (14%) have no pLoF variants observed in gnomAD, but our sample size is not yet large enough to robustly infer LoF intolerance. For these genes, observation of outbred two-hit individuals is not expected, and we cannot yet assess the feasibility of identifying consanguineous two-hit individuals because we lack an estimate of pLoF allele frequency.

This leaves 7,435 genes (39%) for which one or more pLoFs are observed in gnomAD, but strong LoF intolerance cannot be determined, two-hit genotypes have not been observed, and a human disease phenotype is not known. We projected the sample sizes required to identify knockout individuals for these genes (Fig. 2e). In outbred populations, current sample sizes would need to increase by approximately 1,000-fold before ascertainment of a single two-hit LoF individual would be expected for the typical gene. By contrast, around a 10- to 100-fold increase from current consanguineous sample size, meaning hundreds of thousands of individuals in absolute terms, would identify at least one two-hit LoF individual for the typical gene. Among other simplifying assumptions (Methods), these projections presume that complete knockout is tolerated. When only one or a few two-hit individuals are expected in a dataset, the absence of any such individuals can be due to either early lethality, a severe clinical phenotype incompatible with inclusion in gnomAD, or simply chance. Thus, the ability to infer lethality of the two-hit genotype based on statistical evidence will lag behind the identification of two-hit individuals where they do exist (Fig. 2e). For some genes, inference of lethality will always remain impossible in outbred populations, though it may be feasible in consanguineous individuals.

Curation of pLoF variants

Where pLoF variants can be identified, they are a valuable resource for assessing the effect of lifelong reduction in gene dosage. To highlight the challenges and opportunities of identifying such variants, we manually curated gnomAD data and the scientific literature for six genes associated with gain-of-function (GoF) neurodegenerative diseases, for which inhibitors or suppressors are under development[30–35]: *HTT* (Huntington's disease), *MAPT* (tauopathies), *PRNP* (prion disease), *SOD1* (amyotrophic lateral sclerosis), and *LRRK2* and *SNCA* (Parkinson's disease). The results (Fig. 3, Extended Data Table 3) illustrate four points about pLoF variant curation.

First, other things being equal, genes with longer coding sequences offer more opportunities for LoF variants to arise, and so tend to have a higher cumulative frequencies of LoF variants, unless they are heavily constrained. Ascertainment of LoF individuals is thus harder for shorter and/or more constrained genes, even though these may be good targets.

Second, many variants annotated as pLoF are false positives[6], and these are enriched for higher allele frequencies, so that both filtering and curation have an outsized effect on the cumulative allele frequency of LoF. Studies of human pLoF variants lacking stringent curation can therefore easily dilute results with false pLoF carriers.

Third, after careful curation, cumulative LoF allele frequency is sometimes sufficiently high to place certain bounds on what heterozygote phenotype might exist. For example, GoF mutations causing genetic prion disease have a genetic prevalence of approximately 1 in 50,000[36] and have been known for three decades, with thousands of cases identified, making it unlikely that a comparably severe and penetrant haploinsufficiency syndrome associated with *PRNP* would have gone unnoticed to the present day despite being more than twice as common (roughly 1 in 18,000). Similar arguments can be made for *HTT*, *LRRK2* and *SOD1* genes (Extended Data Tables 3, 4). Of course, this does not rule out a less severe or less penetrant heterozygous LoF phenotype.

Finally, careful inspection of the distributions of pLoF variants can reveal important error modes or disease biology. *HTT*, *MAPT*

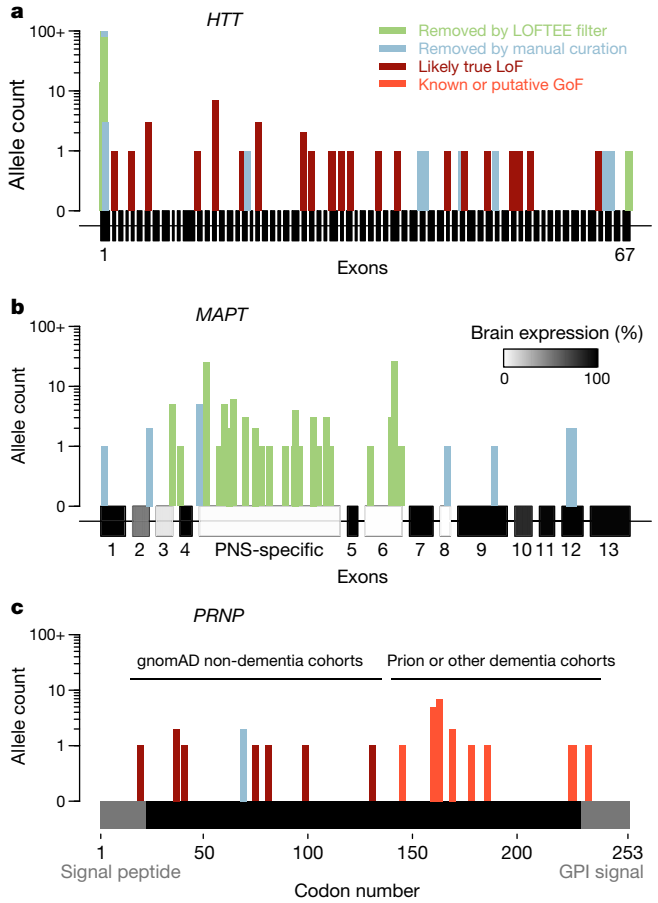


Fig. 3 | Insights from non-random positional distributions of pLoF variants. **a**–**c**, *HTT* (**a**), *MAPT*, with brain expression data from GTEx[40] (**b**) and *PRNP*, a single protein-coding exon with domains removed by post-translational modification in grey (**c**), showing previously reported variants[41] and those newly identified in gnomAD and in the literature (Extended Data Table 5). GPI, glycosylphosphatidylinositol. Detailed variant curation results are provided in Supplementary Table 1.

and *PRNP* genes each have different non-random positional distributions of pLoF variants (Fig. 3). High-frequency *HTT* pLoF variants cluster in the polyglutamine/polyproline repeat region of exon 1 and appear to be alignment artefacts (Fig. 3a). True *HTT* LoF variants are rare and the gene is highly constrained, which might suggest some fitness effect in a heterozygous state in addition to the known severe homozygous phenotype[37,38], although the frequency of LoF carriers still argues against a penetrant syndromic illness, consistent with the lack of phenotype reported in heterozygotes identified so far[38,39]. High-frequency *MAPT* pLoF variants cluster in exons not expressed in the brain in GTEx data[14,40], and all remaining pLoFs appear to be alignment or annotation errors (Fig. 3b). No true LoFs are observed in *MAPT*, although our sample size is insufficient to prove that *MAPT* LoF is not tolerated—among constitutive brain-expressed exons, we expect 12.6 LoFs and observe 0, giving a 95% confidence interval upper bound of 23.7% for obs/exp values. *PRNP*-truncating variants in gnomAD cluster in the N terminus; the sole C-terminal truncating variant in gnomAD is a dementia case (Extended Data Table 5), consistent with variants at codon ≥145 causing a pathogenic gain-of-function through change in localization (Fig. 3c). Within codons 1–144, *PRNP* is unconstrained (Extended Data Table 3), and no neurological phenotype has been identified in individuals with truncating variants so far, consistent with the hypothesis that N-terminal truncating variants are true LoF and are tolerated in a heterozygous state[41].

Discussion

Studying human gene inactivation can illuminate human biology and guide the selection of drug targets, complementing mouse knockout studies[42], but analysis of any one gene requires genome-wide context to set expectations and guide inferences. Here we have used gnomAD data to provide context to aid in the interpretation of human LoF variants.

Targets of approved drugs range from highly constrained to completely unconstrained. There may be several reasons why some genes apparently tolerate pharmacological inhibition but not genetic inactivation. LoF variants in constitutive exons should affect all tissues for life, whereas drugs differ in tissue distribution and timing and duration of use. Many drugs known or suspected to cause fetal harm are tolerated in adults[43], and might target developmentally important genes. Constraint is thought to primarily reflect selection against heterozygotes[13], the effective gene dosage of which may differ from that achieved by a drug. Constraint measures natural selection over centuries or millennia; the environment of our ancestors presented different selective pressures from what we face today. Actions of small-molecule drugs may not map one-to-one onto genes[44–47]. Regardless, these human in vivo data show that even a highly deleterious knockout phenotype is compatible with a gene being a viable drug target.

For most genes, the lack of total knockout individuals identified so far does not yet provide statistical evidence that this genotype is not tolerated. Indeed, for many genes, such evidence may never be attainable in outbred populations. Bottlenecked populations, individually, are unlikely to yield two-hit individuals for a pre-specified gene of interest, although the sequencing of many different, diverse bottlenecked populations will certainly expand the set of genes accessible by this approach. Identification of two-hit individuals will be most greatly aided by increased investment in consanguineous cohorts, in which the sample size required for any given gene is often orders of magnitude lower than in outbred populations. Our analysis is limited by sample size, insufficient diversity of sampled populations, and simplifying assumptions about population structure and distribution of LoF variants, so our calculations should be taken as rough, order-of-magnitude estimates. Nonetheless, this strategic roadmap for the identification of human knockouts should inform future research investments and rationalize the interpretation of existing data.

Recall-by-genotype efforts are only valuable if the variants in question are correctly annotated. Automated filtering[7] and transcript expression-aware annotation[14] are powerful tools, but we demonstrate the continued value of manual curation for excluding further false positives, assessing and interpreting the cumulative allele frequency of true LoF variants, and identifying error modes or biological phenomena that give rise to non-random distributions of pLoF variants across a gene. Such curation is essential before any recontact efforts, and establishing methods for high-throughput functional validation[48] of LoF variants is a priority. Our curation of pLoF variants in neurodegenerative disease genes is limited by a lack of functional validation and detailed phenotyping; a companion paper demonstrates a deeper investigation of the effects of LoF variants in the *LRRK2* gene[49].

Drug development projects may increasingly be accompanied by efforts to phenotype human carriers of LoF variants. With the cost of drug discovery driven overwhelmingly by failure[50], successful interpretation of LoF data to select the right targets and right clinical pathways will yield outsize benefits for research productivity and, ultimately, human health.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2267-z.

1. Plenge, R. M., Scolnick, E. M. & Altshuler, D. Validating therapeutic targets through human genetics. *Nat. Rev. Drug Discov.* **12**, 581–594 (2013).
2. Hay, M., Thomas, D. W., Craighead, J. L., Economides, C. & Rosenthal, J. Clinical development success rates for investigational drugs. *Nat. Biotechnol.* **32**, 40–51 (2014).
3. Nelson, M. R. et al. The support of human genetic evidence for approved drug indications. *Nat. Genet.* **47**, 856–860 (2015).
4. King, E. A., Davis, J. W. & Degner, J. F. Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. *PLoS Genet.* **15**, e1008489 (2019).
5. Musunuru, K. & Kathiresan, S. Genetics of common, complex coronary artery disease. *Cell* **177**, 132–145 (2019).
6. MacArthur, D. G. et al. A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823–828 (2012).
7. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* https://doi.org/10.1038/s41586-020-2308-7 (2020).
8. Lim, E. T. et al. Distribution and medical impact of loss-of-function variants in the Finnish founder population. *PLoS Genet.* **10**, e1004494 (2014).
9. Bittles, A. H. & Black, M. L. Evolution in health and medicine Sackler colloquium: consanguinity, human evolution, and complex diseases. *Proc. Natl Acad. Sci. USA* **107** (Suppl. 1), 1779–1786 (2010).
10. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
11. Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. & Goldstein, D. B. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet.* **9**, e1003709 (2013).
12. Samocha, K. E. et al. A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* **46**, 944–950 (2014).
13. Fuller, Z. L., Berg, J. J., Mostafavi, H., Sella, G. & Przeworski, M. Measuring intolerance to mutation in human genetics. *Nat. Genet.* **51**, 772–776 (2019).
14. Cummings, B. B. et al. Transcript expression-aware annotation improves rare variant discovery and interpretation. *Nature* https://doi.org/10.1038/s41586-020-2329-2 (2020).
15. Collins, R. L. et al. A structural variation reference for medical and population genetics. *Nature* https://doi.org/10.1038/s41586-020-2287-8 (2020).
16. Kosmicki, J. A. et al. Refining the role of de novo protein-truncating variants in neurodevelopmental disorders by using population reference samples. *Nat. Genet.* **49**, 504–510 (2017).
17. Wishart, D. S. et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* **46** (D1), D1074–D1082 (2018).
18. Rehm, H. L. et al. ClinGen—the clinical genome resource. *N. Engl. J. Med.* **372**, 2235–2242 (2015).
19. Morham, S. G. et al. Prostaglandin synthase 2 gene disruption causes severe renal pathology in the mouse. *Cell* **83**, 473–482 (1995).
20. Ohashi, K. et al. Early embryonic lethality caused by targeted disruption of the 3-hydroxy-3-methylglutaryl-CoA reductase gene. *J. Biol. Chem.* **278**, 42936–42941 (2003).
21. Nagashima, S. et al. Liver-specific deletion of 3-hydroxy-3-methylglutaryl coenzyme A reductase causes hepatic steatosis and death. *Arterioscler. Thromb. Vasc. Biol.* **32**, 1824–1831 (2012).
22. Lv, W. et al. The drug target genes show higher evolutionary conservation than non-target genes. *Oncotarget* **7**, 4961–4971 (2016).
23. Motenko, H., Neuhauser, S. B., O'Keefe, M. & Richardson, J. E. MouseMine: a new data warehouse for MGI. *Mamm. Genome* **26**, 325–330 (2015).
24. Hart, T. et al. Evaluation and design of genome-wide CRISPR/SpCas9 knockout screens. *G3 (Bethesda)* **7**, 2719–2727 (2017).
25. Finer, S. et al. Cohort profile: East London genes & health (ELGH), a community-based population genomics and health study of British Bangladeshi and British Pakistani people. *Int. J. Epidemiol.* **49**, 20–21 (2019).
26. Narasimhan, V. M. et al. Health and population effects of rare gene knockouts in adult humans with related parents. *Science* **352**, 474–477 (2016).
27. Saleheen, D. et al. Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. *Nature* **544**, 235–239 (2017).
28. Sulem, P. et al. Identification of a large set of rare complete human knockouts. *Nat. Genet.* **47**, 448–452 (2015).
29. DeBoever, C. et al. Medical relevance of protein-truncating variants across 337,205 individuals in the UK Biobank study. *Nat. Commun.* **9**, 1612 (2018).
30. Tabrizi, S. J. et al. Targeting Huntingtin expression in patients with Huntington's disease. *N. Engl. J. Med.* **380**, 2307–2316 (2019).
31. DeVos, S. L. et al. Tau reduction prevents neuronal loss and reverses pathological tau deposition and seeding in mice with tauopathy. *Sci. Transl. Med.* **9**, eaag0481 (2017).
32. Minikel, E. V. et al. Prion protein lowering is a disease-modifying therapy across prion strains, disease stages, and endpoints. Preprint at https://doi.org/10.1101/2020.03.27.011940 (2020).
33. McCampbell, A. et al. Antisense oligonucleotides extend survival and reverse decrement in muscle response in ALS models. *J. Clin. Invest.* **128**, 3558–3567 (2018).
34. Chen, J., Chen, Y. & Pu, J. Leucine-rich repeat kinase 2 in Parkinson's disease: updated from pathogenesis to potential therapeutic target. *Eur. Neurol.* **79**, 256–265 (2018).
35. Cole, T. A. et al. Alpha-synuclein antisense oligonucleotides as a disease-modifying therapy for Parkinson's disease. Preprint at https://doi.org/10.1101/830554 (2019).
36. Minikel, E. V. et al. Age at onset in genetic prion disease and the design of preventive clinical trials. *Neurology* **93**, e125–e134 (2019).
37. Duyao, M. P. et al. Inactivation of the mouse Huntington's disease gene homolog Hdh. *Science* **269**, 407–410 (1995).
38. Rodan, L. H. et al. A novel neurodevelopmental disorder associated with compound heterozygous variants in the huntingtin gene. *Eur. J. Hum. Genet.* **24**, 1826–1827 (2016).
39. Ambrose, C. M. et al. Structure and expression of the Huntington's disease gene: evidence against simple inactivation due to an expanded CAG repeat. *Somat. Cell Mol. Genet.* **20**, 27–38 (1994).
40. GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
41. Minikel, E. V. et al. Quantifying prion disease penetrance using large population control cohorts. *Sci. Transl. Med.* **8**, 322ra9 (2016).
42. Zambrowicz, B. P. & Sands, A. T. Knockouts model the 100 best-selling drugs—will they model the next 100? *Nat. Rev. Drug Discov.* **2**, 38–51 (2003).
43. Uhl, K., Kennedy, D. L. & Kweder, S. L. Risk management strategies in the Physicians' Desk Reference product labels for pregnancy category X drugs. *Drug Saf.* **25**, 885–892 (2002).
44. Haggarty, S. J., Koeller, K. M., Wong, J. C., Grozinger, C. M. & Schreiber, S. L. Domain-selective small-molecule inhibitor of histone deacetylase 6 (HDAC6)-mediated tubulin deacetylation. *Proc. Natl Acad. Sci. USA* **100**, 4389–4394 (2003).
45. Zhang, B. W. et al. T cell responses in calcineurin A alpha-deficient mice. *J. Exp. Med.* **183**, 413–420 (1996).
46. Jacinto, E. et al. Mammalian TOR complex 2 controls the actin cytoskeleton and is rapamycin insensitive. *Nat. Cell Biol.* **6**, 1122–1128 (2004).
47. Hoshi, N., Langeberg, L. K., Gould, C. M., Newton, A. C. & Scott, J. D. Interaction with AKAP79 modifies the cellular pharmacology of PKC. *Mol. Cell* **37**, 541–550 (2010).
48. Findlay, G. M. et al. Accurate classification of BRCA1 variants with saturation genome editing. *Nature* **562**, 217–222 (2018).
49. Whiffin, N. et al. Characterising the loss-of-function impact of 5′ untranslated region variants in whole genome sequence data from 15,708 individuals. *Nat. Commun.* https://doi.org/10.1038/s41467-019-10717-9 (2020).
50. DiMasi, J. A., Grabowski, H. G. & Hansen, R. W. Innovation in the pharmaceutical industry: new estimates of R&D costs. *J. Health Econ.* **47**, 20–33 (2016).

Genome Aggregation Database Production Team

Jessica Alföldi[1,4], Irina M. Armean[1,4,15], Eric Banks[16], Louis Bergelson[16], Kristian Cibulskis[16], Ryan L. Collins[1,17,18], Kristen M. Connolly[19], Miguel Covarrubias[16], Beryl B. Cummings[1,4,5], Mark J. Daly[1,4], Stacey Donnelly[1], Yossi Farjoun[16], Steven Ferriera[20], Laurent Francioli[1,4], Stacey Gabriel[20], Laura D. Gauthier[16], Jeff Gentry[16], Namrata Gupta[1,20], Thibault Jeandet[16], Diane Kaplan[16], Konrad J. Karczewski[1,4], Kristen M. Laricchia[1,4], Christopher Llanwarne[16], Eric V. Minikel[1,2,4], Ruchi Munshi[16], Benjamin M. Neale[1,4], Sam Novod[16], Anne H. O'Donnell-Luria[1,21,22], Nikelle Petrillo[16], Timothy Poterba[1,2,4], David Roazen[16], Valentin Ruano-Rubio[16], Andrea Saltzman[1], Kaitlin E. Samocha[9], Molly Schleicher[1], Cotton Seed[2,4], Matthew Solomonson[1,4], Jose Soto[16], Grace Tiao[1,4], Kathleen Tibbetts[16], Charlotte Tolonen[16], Christopher Vittal[2,4], Gordon Wade[16], Arcturus Wang[1,2,4], Qingbo Wang[1,4,18], James S. Ware[1,23,24], Nicholas A. Watts[1,4], Ben Weisburd[16] & Nicola Whiffin[1,23,24]

[15]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK. [16]Data Sciences Platform, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [17]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA. [18]Program in Bioinformatics and Integrative Genomics, Harvard Medical School, Boston, MA, USA. [19]Genomics Platform, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [20]Broad Genomics, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [21]Division of Genetics and Genomics, Boston Children's Hospital, Boston, MA, USA. [22]Department of Pediatrics, Harvard Medical School, Boston, MA, USA. [23]National Heart & Lung Institute and MRC London Institute of Medical Sciences, Imperial College London, London, UK. [24]Cardiovascular Research Centre, Royal Brompton & Harefield Hospitals NHS Trust, London, UK.

Genome Aggregation Database Consortium

Carlos A. Aguilar Salinas[25], Tariq Ahmad[26], Christine M. Albert[27,28], Diego Ardissino[29], Gil Atzmon[30,31,32], John Barnard[33], Laurent Beaugerie[34], Emelia J. Benjamin[35,36,37], Michael Boehnke[38], Lori L. Bonnycastle[39], Erwin P. Bottinger[40], Donald W. Bowden[41,42,43], Matthew J. Bown[44], John C. Chambers[45,46,47], Juliana C. Chan[48], Daniel Chasman[27], Judy Cho[40], Mina K. Chung[33], Bruce Cohen[49,50], Adolfo Correa[51], Dana Dabelea[52], Mark J. Daly[1,4], Dawood Darbar[53], Ravindranath Duggirala[54], Josée Dupuis[55,56], Patrick T. Ellinor[1,57], Roberto Elosua[58,59,60], Jeanette Erdmann[61,62,63], Tõnu Esko[1,64], Martti Färkkilä[65], Jose Florez[1], Andre Franke[66], Gad Getz[67], Benjamin Glaser[68], Stephen J. Glatt[69], David Goldstein[70,71], Clicerio Gonzalez[72], Leif Groop[73], Christopher Haiman[74], Craig Hanis[75], Matthew Harms[76,77], Mikko Hiltunen[78], Matti M. Holi[79], Christina M. Hultman[80], Mikko Kallela[81], Jaakko Kaprio[82,83], Sekar Kathiresan[1,17,84], Bong-Jo Kim[85], Young Jin Kim[85], George Kirov[86], Jaspal Kooner[10,46,47],

Seppo Koskinen[87], Harlan M. Krumholz[88], Subra Kugathasan[89], Soo Heon Kwak[90], Markku Laakso[91,92], Terho Lehtimäki[93], Ruth J. F. Loos[40,94], Steven A. Lubitz[1,95], Ronald C. W. Ma[96,97,98], Daniel G. MacArthur[1,4,153,154], Jaume Marrugat[59,99], Kari M. Mattila[93], Steven McCarroll[2,100], Mark I. McCarthy[101,102,103], Dermot McGovern[104], Ruth McPherson[105], James B. Meigs[1,84,106], Olle Melander[107], Andres Metspalu[64], Benjamin M. Neale[1,4], Peter M. Nilsson[108], Michael C. O'Donovan[86], Dost Ongur[49,84], Lorena Orozco[109], Michael J. Owen[86], Colin N. A. Palmer[110], Aarno Palotie[1,4,82], Kyong Soo Park[90,111], Carlos Pato[112], Ann E. Pulver[113], Nazneen Rahman[114], Anne M. Remes[115], John D. Rioux[116,117], Samuli Ripatti[1,82,118], Dan M. Roden[119,120], Danish Saleheen[121,122,123], Veikko Salomaa[124], Nilesh J. Samani[44,125], Jeremiah Scharf[1,2,17], Heribert Schunkert[126,127], Moore B. Shoemaker[128], Pamela Sklar[129,130,131,155], Hilkka Soininen[132], Harry Sokol[34], Tim Spector[133], Patrick F. Sullivan[80,134], Jaana Suvisaari[124], E. Shyong Tai[135,136,137], Yik Ying Teo[135,138,139], Tuomi Tiinamaija[82,140,141], Ming Tsuang[142,143], Teresa Dan Turner[144], Teresa Tusie-Luna[145,146], Erkki Vartiainen[147], James. S. Ware[1,23,24], Hugh Watkins[148], Rinse K. Weersma[149], Maija Wessman[82,140], James G. Wilson[150] & Ramnik J. Xavier[151,152]

[25]Unidad de Investigacion de Enfermedades Metabolicas, Instituto Nacional de Ciencias Medicas y Nutricion, Mexico City, Mexico. [26]Peninsula College of Medicine and Dentistry, Exeter, UK. [27]Division of Preventive Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. [28]Division of Cardiovascular Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. [29]Department of Cardiology, University Hospital, Parma, Italy. [30]Department of Biology, Faculty of Natural Sciences, University of Haifa, Haifa, Israel. [31]Department of Medicine, Albert Einstein College of Medicine, Bronx, NY, USA. [32]Department of Genetics, Albert Einstein College of Medicine, Bronx, NY, USA. [33]Department of Quantitative Health Sciences, Lerner Research Institute, Cleveland Clinic, Cleveland, OH, USA. [34]Sorbonne Université, APHP, Gastroenterology Department, Saint Antoine Hospital, Paris, France. [35]Framingham Heart Study, National Heart, Lung, & Blood Institute and Boston University, Framingham, MA, USA. [36]Department of Medicine, Boston University School of Medicine, Boston, MA, USA. [37]Department of Epidemiology, Boston University School of Public Health, Boston, MA, USA. [38]Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, MI, USA. [39]National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA. [40]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [41]Department of Biochemistry, Wake Forest School of Medicine, Winston-Salem, NC, USA. [42]Center for Genomics and Personalized Medicine Research, Wake Forest School of Medicine, Winston-Salem, NC, USA. [43]Center for Diabetes Research, Wake Forest School of Medicine, Winston-Salem, NC, USA. [44]Department of Cardiovascular Sciences and NIHR Leicester Biomedical Research Centre, University of Leicester, Leicester, UK. [45]Department of Epidemiology and Biostatistics, Imperial College London, London, UK. [46]Department of Cardiology, Ealing Hospital NHS Trust, Southall, UK. [47]Imperial College Healthcare NHS Trust, Imperial College London, London, UK. [48]Department of Medicine and Therapeutics, Chinese University of Hong Kong, Hong Kong, Hong Kong. [49]Program for Neuropsychiatric Research, McLean Hospital, Belmont, MA, USA. [50]Department of Psychiatry, Harvard Medical School, Boston, MA, USA. [51]Department of Medicine, University of Mississippi Medical Center, Jackson, MI, USA. [52]Department of Epidemiology, Colorado School of Public Health, Aurora, CO, USA. [53]Department of Medicine and Pharmacology, University of Illinois at Chicago, Chicago, IL, USA. [54]Department of Genetics, Texas Biomedical Research Institute, San Antonio, TX, USA. [55]Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA. [56]Framingham Heart Study, National Heart, Lung, & Blood Institute and Boston University, Framingham, MA, USA. [57]Cardiac Arrhythmia Service and Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA, USA. [58]Cardiovascular Epidemiology and Genetics, Hospital del Mar Medical Research Institute (IMIM), Barcelona, Catalonia, Spain. [59]Centro de Investigación Biomédica en Red Enfermedades Cardiovaculares (CIBERCV), Barcelona, Catalonia, Spain. [60]Departament of Medicine, Medical School, University of Vic-Central University of Catalonia, Vic, Catalonia, Spain. [61]Institute for Cardiogenetics, University of Lübeck, Lübeck, Germany. [62]DZHK (German Research Centre for Cardiovascular Research), partner site Hamburg/Lübeck/Kiel, Lübeck, Germany. [63]University Heart Center Lübeck, Lübeck, Germany. [64]Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia. [65]Clinic of Gastroenterology, Helsinki University and Helsinki University Hospital, Helsinki, Finland. [66]Institute of Clinical Molecular Biology (IKMB), Christian-Albrechts-University of Kiel, Kiel, Germany. [67]Cancer Genome Computational Analysis Group, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [68]Endocrinology and Metabolism Department, Hadassah-Hebrew University Medical Center, Jerusalem, Israel. [69]Department of Psychiatry and Behavioral Sciences, SUNY Upstate Medical University, Syracuse, NY, USA. [70]Institute for Genomic Medicine, Columbia University Medical Center, Hammer Health Sciences, New York, NY, USA. [71]Department of Genetics & Development, Columbia University Medical Center, Hammer Health Sciences, New York, NY, USA. [72]Centro de Investigacion en Salud Poblacional, Instituto Nacional de Salud Publica, Cuernavaca, Mexico. [73]Genomics, Diabetes and Endocrinology, Lund University, Malmo, Sweden. [74]Lund University Diabetes Centre, Malmo, Sweden. [75]Human Genetics Center, University of Texas Health Science Center at Houston, Houston, TX, USA. [76]Department of Neurology, Columbia University, New York, NY, USA. [77]Institute of Genomic Medicine, Columbia University, New York, NY, USA. [78]Institute of Biomedicine, University of Eastern Finland, Kuopio, Finland. [79]Department of Psychiatry, Helsinki University Central Hospital, Lapinlahdentie, Helsinki, Finland. [80]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [81]Department of Neurology, Helsinki University Central Hospital, Helsinki, Finland. [82]Institute for Molecular Medicine FIMM, University of Helsinki, Helsinki, Finland. [83]Department of Public Health, University of Helsinki, Helsinki, Finland. [84]Department of Medicine, Harvard Medical School, Boston, MA, USA. [85]Center for Genome Science, Korea National Institute of Health, Chungcheongbuk-do, South Korea. [86]MRC Centre for Neuropsychiatric Genetics & Genomics, Cardiff University School of Medicine, Cardiff, UK. [87]Department of Health, THL-National Institute for Health and Welfare, Helsinki, Finland. [88]Section of Cardiovascular Medicine, Department of Internal Medicine, Yale School of Medicine, New Haven, CT, USA. [89]Division of Pediatric Gastroenterology, Emory University School of Medicine, Atlanta, GA, USA. [90]Department of Internal Medicine, Seoul National University Hospital, Seoul, South Korea. [91]Institute of Clinical Medicine, The University of Eastern Finland, Kuopio, Finland. [92]Kuopio University Hospital, Kuopio, Finland. [93]Department of Clinical Chemistry, Fimlab Laboratories and Finnish Cardiovascular Research Center-Tampere, Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland. [94]The Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [95]Cardiac Arrhythmia Service, Massachusetts General Hospital, Boston, MA, USA. [96]Department of Medicine and Therapeutics, The Chinese University of Hong Kong, Hong Kong, China. [97]Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Hong Kong, China. [98]Hong Kong Institute of Diabetes and Obesity, The Chinese University of Hong Kong, Hong Kong, China. [99]Cardiovascular Research REGICOR Group, Hospital del Mar Medical Research Institute (IMIM), Barcelona, Catalonia, Spain. [100]Department of Genetics, Harvard Medical School, Boston, MA, USA. [101]Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Churchill Hospital, Headington, Oxford, UK. [102]Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK. [103]Oxford NIHR Biomedical Research Centre, Oxford University Hospitals NHS Foundation Trust, John Radcliffe Hospital, Oxford, UK. [104]F Widjaja Foundation Inflammatory Bowel and Immunobiology Research Institute, Cedars-Sinai Medical Center, Los Angeles, CA, USA. [105]Atherogenomics Laboratory, University of Ottawa Heart Institute, Ottawa, Canada. [106]Division of General Internal Medicine, Massachusetts General Hospital, Boston, MA, USA. [107]Department of Clinical Sciences, University Hospital Malmo Clinical Research Center, Lund University, Malmo, Sweden. [108]Department of Clinical Sciences, Lund University, Skane University Hospital, Malmo, Sweden. [109]Instituto Nacional de Medicina Genómica (INMEGEN), Mexico City, Mexico. [110]Medical Research Institute, Ninewells Hospital and Medical School, University of Dundee, Dundee, UK. [111]Department of Molecular Medicine and Biopharmaceutical Sciences, Graduate School of Convergence Science and Technology, Seoul National University, Seoul, South Korea. [112]Department of Psychiatry, Keck School of Medicine at the University of Southern California, Los Angeles, CA, USA. [113]Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, MD, USA. [114]Division of Genetics and Epidemiology, Institute of Cancer Research, London, UK. [115]Research Unit of Clinical Neuroscience, University of Oulu, Oulu, Finland. [116]Research Center, Montreal Heart Institute, Montreal, Quebec, Canada. [117]Department of Medicine, Faculty of Medicine, Université de Montréal, Quebec, Canada. [118]Department of Public Health, Faculty of Medicine, University of Helsinki, Helsinki, Finland. [119]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA. [120]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA. [121]Department of Biostatistics and Epidemiology, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA. [122]Department of Medicine, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA. [123]Center for Non-Communicable Diseases, Karachi, Pakistan. [124]National Institute for Health and Welfare, Helsinki, Finland. [125]NIHR Leicester Biomedical Research Centre, Glenfield Hospital, Leicester, UK. [126]Deutsches Herzzentrum München, Munich, Germany. [127]Technische Universität München, Munich, Germany. [128]Division of Cardiovascular Medicine, Nashville VA Medical Center and Vanderbilt University, School of Medicine, Nashville, TN, USA. [129]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [130]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [131]Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [132]Institute of Clinical Medicine Neurology, University of Eastern Finland, Kuopio, Finland. [133]Department of Twin Research and Genetic Epidemiology, King's College London, London, UK. [134]Department of Genetics and Psychiatry, University of North Carolina, Chapel Hill, NC, USA. [135]Saw Swee Hock School of Public Health, National University of Singapore, National University Health System, Singapore, Singapore. [136]Department of Medicine, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore. [137]Duke-NUS Graduate Medical School, Singapore, Singapore. [138]Life Sciences Institute, National University of Singapore, Singapore, Singapore. [139]Department of Statistics and Applied Probability, National University of Singapore, Singapore, Singapore. [140]Folkhälsan Institute of Genetics, Folkhälsan Research Center, Helsinki, Finland. [141]HUCH Abdominal Center, Helsinki University Hospital, Helsinki, Finland. [142]Center for Behavioral Genomics, Department of Psychiatry, University of California, San Diego, CA, USA. [143]Institute of Genomic Medicine, University of California, San Diego, CA, USA. [144]Juliet Keidan Institute of Pediatric Gastroenterology, Shaare Zedek Medical Center, The Hebrew University of Jerusalem, Jerusalem, Israel. [145]Instituto de Investigaciones Biomédicas UNAM, Mexico City, Mexico. [146]Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, Mexico City, Mexico. [147]Department of Public Health Solutions, National Institute for Health and Welfare, Helsinki, Finland. [148]Radcliffe Department of Medicine, University of Oxford, Oxford, UK. [149]Department of Gastroenterology and Hepatology, University of Groningen and University Medical Center Groningen, Groningen, The Netherlands. [150]Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, MS, USA. [151]Program in Infectious Disease and Microbiome, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [152]Center for Computational and Integrative Biology, Massachusetts General Hospital, Boston, MA, USA. [155]Deceased: Pamela Sklar.

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment.

Data sources

pLoF analyses used the gnomAD dataset of 141,456 individuals[7]. For data consistency, all genome-wide constraint and CAF analyses used only the 125,748 gnomAD exomes. Curated analyses of individual genes used all 141,456 individuals including 15,708 whole genomes. Gene lists used in this study were extracted from public data sources between September 2018 and June 2019. Data sources and criteria for gene list extraction are shown in Extended Data Table 1. This study was performed under ethical approval from the Partners Healthcare Institutional Research Board (2013P001339/MGH) and the Broad Institute Office of Research Subjects Protection (ORSP-3862). All research participants provided informed consent.

Calculation of pLoF constraint

The calculation of constraint values for genes has been described in general elsewhere[10,12] and for this dataset specifically by Karczewski et al.[7]. Constraint calculations used LOFTEE-filtered ('high confidence') single-nucleotide variants (which for pLoF means nonsense and essential splice site mutations) found in gnomAD exomes with minor allele frequency <0.1%. Only unique canonical transcripts for protein-coding genes were considered, yielding 17,604 genes with available constraint values. For curated genes (Extended Data Table 2), the number of observed variants passing curation was divided by the expected number of variants to yield a curated constraint value. For *PRNP*, the expected number of variants was adjusted by multiplying by the ratio of the sum of mutation frequencies for all possible pLoF variants in codons 1–144 to the sum of mutation frequencies for all possible pLoF variants in the entire transcript, yielding 6 observed out of 6.06 expected. For *MAPT*, the expected number of variants was taken from Ensembl transcript ENST00000334239, which includes only the exons identified as constitutively brain-expressed in Fig. 3b (exon numbering previously described[51]).

Calculation of pLoF heterozygote and homozygote/compound heterozygote frequencies

LOFTEE-filtered high-confidence pLoF variants with minor allele frequency <5% in 125,748 gnomAD exomes were used to compute the proportion of individuals without a loss-of-function variant (q); the CAF was computed as $p = 1 - \mathrm{sqrt}(q)$. This approach conservatively assumes that, if an individual has two different pLoF variants, they are in *cis* to each other and count as only one pLoF allele.

For outbred populations (Fig. 2a), we used the value of p from all 125,748 gnomAD exomes, as this allows the largest possible sample size. This includes some individuals from bottlenecked populations, for which the distribution of p does differ from outbred populations, but these individuals are a small proportion of gnomAD exomes (12.6%). This also includes some consanguineous individuals, but these are an even smaller proportion of gnomAD exomes (2.3%), and any difference in the value of p between consanguineous and outbred populations is expected to be very small. Heterozygote frequency was calculated as $2p(1-p)$ and homozygote and compound heterozygote frequency was calculated as p^2. Lines indicate the size of gnomAD (141,456 individuals) and the world population (6.69 billion).

For bottlenecked populations (Fig. 2b), we used the value of p from the 10,824 Finnish exomes only. Lines indicate the number of Finnish individuals in gnomAD (12,526) and the population of Finland (5.5 million).

For consanguineous individuals (Fig. 2c), we again used the value of p from all gnomAD exomes, because p is not expected to differ greatly in consanguineous versus outbred populations. We used the mean proportion of the genome in runs of autozygosity (a) from individuals self-reporting second cousin or closer parents in East London Genes & Health, $a = 0.05766$ (rounded to 5.8%). Heterozygote frequency was calculated as $2p(1-p)$ and homozygote and compound heterozygote frequency was calculated as $(1 - a)p^2 + ap$. Lines indicate the number of consanguineous South Asian individuals in gnomAD ($n = 2{,}912$, by coincidence the same number as report second cousin or closer parents in ELGH) based on $F > 0.05$ (a conservative estimate, because second cousin parents are expected to yield $F = 0.015625$), and the estimated number of individuals in the world with second cousin or closer parents (10.4% of the world population)[9].

Several caveats apply to our CAF analysis. First, our approach naively treats genes with no pLoFs observed as having $P = 0$, even though pLoFs might be discovered at a larger sample size. Second, we naively group all populations together, even though the distribution of populations sampled in gnomAD does not reflect the world population[7]; we believe that this is reasonable because CAF for many genes is driven by singletons and other ultra-rare variants for which frequency is not expected to differ appreciably by continental population[10]. (It is important to note that the histograms shown in Fig. 2 reflect the expected frequency of heterozygotes and homozygotes/compound heterozygotes, based on gnomAD allele frequency, rather than the actual observed frequency of individuals with these genotypes in gnomAD.) Third, we use only protein-truncating variants annotated as pLoF in gnomAD. Structural and non-coding variation resulting in a loss of function may be missed in exomes, and missense variants resulting in a loss of function cannot be rigorously annotated. Fourth, we naively treat genes with one pLoF allele observed as having $P = 1/(2 \times 125{,}748)$, even though on average singleton variants have a true allele frequency lower than their nominal allele frequency[10]. Fifth, the variants included in this analysis are filtered but have not been manually curated or functionally validated, so some will ultimately prove not to be true LoF. These false positives tend to be more common and will have disproportionately contributed to the cumulative LoF allele frequency. Sixth, as described in the main text, our calculations assume that complete knockout is tolerated, which will not be true for some genes. We therefore also include a projection of the sample size needed to infer lethality from the absence of two-hit knockout individuals (Fig. 2e). Points one to three will tend to lead to underestimation of the true complete knockout frequency, whereas points four to six will tend to lead to overestimation. On balance, our calculations may reflect an upper bound of complete knockout frequency for most genes owing to the strong influence of factors five and six. Finally, as a matter of comparison between population structures, the sample size for all gnomAD exomes (Fig. 2a, c) is larger than for only Finnish exomes (Fig. 2b). For a version of Fig. 2 with the global gnomAD population downsampled to the same sample size as the gnomAD Finnish population, see Extended Data Fig. 2.

Knockout roadmap

For the knockout 'roadmap' (Fig. 2d, e), we classified genes according to the current status of human disease association and LoF ascertainment. Genes were classified as having a Mendelian disease association if they were present in OMIM with the filters described in Extended Data Table 1.

Remaining genes were classified as '2-hit LoF reported' based on presence in one or more of the following gene lists: homozygous LoF genotypes in gnomAD curated as previously described[7]; filtered homozygous LoF genotypes in runs of autozygosity with minor allele frequency <1% in canonical transcripts in the Bradford, Birmingham and ELGH[25] cohorts (total $n = 8{,}925$); observed number of imputed homozygotes >1 or number of compound heterozygous carriers where minor allele frequency <2% (for both variants) in deCODE[28]; homozygous LoF reported in PROMIS[27]; homozygous LoF with minor allele frequency <1% in UK Biobank[29].

The remainder of genes were sequentially classified as 'likely haploinsufficient' if pLI >0.9 in gnomAD, 'pLoF not yet observed' if CAF = 0 in gnomAD, and, finally, 'pLoF observed in gnomAD' if CAF >0 in gnomAD.

Genetic prevalence estimation

Here, we define 'genetic prevalence' for a given gene as the proportion of individuals in the general population at birth who have a pathogenic variant in that gene that will cause them to later develop disease. Genetic prevalence has not been well-studied or estimated for most disease genes.

In principle, it should be possible to estimate genetic prevalence simply by examining the allele frequency of reported pathogenic variants in gnomAD. In practice, three considerations usually preclude this approach. First, the present gnomAD sample size of 141,456 exomes and genomes is still too small to permit accurate estimates for very rare diseases. Second, the mean age of gnomAD individuals is approximately 55, which is above the age of onset for many rare genetic diseases, and individuals with known Mendelian disease are deliberately excluded, so pathogenic variants will be depleted in this sample relative to the whole birth population. Third and most importantly, a large fraction of reported pathogenic variants lack strong evidence for pathogenicity and are either benign or low penetrance[10,41], so without careful curation of pathogenicity assertions, summing the frequency of reported pathogenic variants in gnomAD will in most cases vastly overestimate the true genetic prevalence of a disease.

Instead, we searched the literature and very roughly estimated genetic prevalence based on available data. In most cases, we took disease incidence (new cases per year per population), multiplied by proportion of cases due to variants in a gene of interest, and multiplied by average age at death in cases. In some cases, estimates of at-risk population or direct measures of genetic prevalence were available. Details of the calculations undertaken for each gene are provided in Extended Data Table 4.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The gnomAD v2 data are available via the gnomAD browser (https://gnomad.broadinstitute.org).

Code availability

Additional data and the R 3.5.1 and Python 2.7.10 source code for this study are available via GitHub (https://github.com/ericminikel/drug_target_lof).

51. Andreadis, A. Tau splicing and the intricacies of dementia. *J. Cell. Physiol.* **227**, 1220–1225 (2012).
52. Yates, B. et al. Genenames.org: the HGNC and VGNC resources in 2017. *Nucleic Acids Res.* **45** (D1), D619–D625 (2017).
53. Mainland, J. D., Li, Y. R., Zhou, T., Liu, W. L. L. & Matsunami, H. Human olfactory receptor responses to odorants. *Sci. Data* **2**, 150002 (2015).
54. Blekhman, R. et al. Natural selection on genes that underlie human disease susceptibility. *Curr. Biol.* **18**, 883–889 (2008).
55. Harding, S. D. et al. The IUPHAR/BPS Guide to PHARMACOLOGY in 2018: updates and expansion to encompass the new guide to IMMUNOPHARMACOLOGY. *Nucleic Acids Res.* **46** (D1), D1091–D1106 (2018).
56. MacArthur, J. et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45** (D1), D896–D901 (2017).
57. Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. & McKusick, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33**, D514–D517 (2005).
58. Pringsheim, T. et al. The incidence and prevalence of Huntington's disease: a systematic review and meta-analysis. *Mov. Disord.* **27**, 1083–1091 (2012).
59. Keum, J. W. et al. The HTT CAG-expansion mutation determines age at death but not disease duration in huntington disease. *Am. J. Hum. Genet.* **98**, 287–298 (2016).
60. Kay, C. et al. Huntington disease reduced penetrance alleles occur at high frequency in the general population. *Neurology* **87**, 282–288 (2016).
61. Fisher, E. R. & Hayden, M. R. Multisource ascertainment of Huntington disease in Canada: prevalence and population at risk. *Mov. Disord.* **29**, 105–114 (2014).
62. Pringsheim, T., Jette, N., Frolkis, A. & Steeves, T. D. L. The prevalence of Parkinson's disease: a systematic review and meta-analysis. *Mov. Disord.* **29**, 1583–1590 (2014).
63. Healy, D. G. et al. Phenotype, genotype, and worldwide genetic penetrance of LRRK2-associated Parkinson's disease: a case-control study. *Lancet Neurol.* **7**, 583–590 (2008).
64. Onyike, C. U. & Diehl-Schmid, J. The epidemiology of frontotemporal dementia. *Int. Rev. Psychiatry* **25**, 130–137 (2013).
65. Bang, J., Spina, S. & Miller, B. L. Frontotemporal dementia. *Lancet* **386**, 1672–1682 (2015).
66. Maddox, R. A. et al. Prion disease incidence in the United States, 2003–2015. *Neurology* **94**, e153–e157 (2019).
67. Trinh, J., Guella, I. & Farrer, M. J. Disease penetrance of late-onset parkinsonism: a meta-analysis. *JAMA Neurol.* **71**, 1535–1539 (2014).
68. Chiò, A. et al. Prevalence of SOD1 mutations in the Italian ALS population. *Neurology* **70**, 533–537 (2008).
69. Cudkowicz, M. E. et al. Epidemiology of mutations in superoxide dismutase in amyotrophic lateral sclerosis. *Ann. Neurol.* **41**, 210–221 (1997).
70. Renton, A. E., Chiò, A. & Traynor, B. J. State of play in amyotrophic lateral sclerosis genetics. *Nat. Neurosci.* **17**, 17–23 (2014).
71. Byrne, S. et al. Rate of familial amyotrophic lateral sclerosis: a systematic review and meta-analysis. *J. Neurol. Neurosurg. Psychiatry* **82**, 623–627 (2011).
72. Rowland, L. P. & Shneider, N. A. Amyotrophic lateral sclerosis. *N. Engl. J. Med.* **344**, 1688–1700 (2001).
73. Hirtz, D. et al. How common are the "common" neurologic disorders? *Neurology* **68**, 326–337 (2007).
74. Logroscino, G. et al. Incidence of amyotrophic lateral sclerosis in Europe. *J. Neurol. Neurosurg. Psychiatry* **81**, 385–390 (2010).
75. Hernandez, D. G., Reed, X. & Singleton, A. B. Genetics in Parkinson disease: Mendelian versus non-Mendelian inheritance. *J. Neurochem.* **139** (Suppl. 1), 59–74 (2016).
76. Funayama, M. et al. A new locus for Parkinson's disease (PARK8) maps to chromosome 12p11.2-q13.1. *Ann. Neurol.* **51**, 296–301 (2002).
77. Zimprich, A. et al. Mutations in LRRK2 cause autosomal-dominant parkinsonism with pleomorphic pathology. *Neuron* **44**, 601–607 (2004).
78. Goldwurm, S. et al. Evaluation of LRRK2 G2019S penetrance: relevance for genetic counseling in Parkinson disease. *Neurology* **68**, 1141–1143 (2007).
79. Do, C. B. et al. Web-based genome-wide association study identifies two novel loci and a substantial genetic component for Parkinson's disease. *PLoS Genet.* **7**, e1002141 (2011).
80. Kinoshita, T. & Fujita, M. Biosynthesis of GPI-anchored proteins: special emphasis on GPI lipid remodeling. *J. Lipid Res.* **57**, 6–24 (2016).
81. Kitamoto, T., Iizuka, R. & Tateishi, J. An amber mutation of prion protein in Gerstmann–Sträussler syndrome with mutant PrP plaques. *Biochem. Biophys. Res. Commun.* **192**, 525–531 (1993).
82. Finckh, U. et al. High prevalence of pathogenic mutations in patients with early-onset dementia detected by sequence analyses of four different genes. *Am. J. Hum. Genet.* **66**, 110–117 (2000).
83. Jayadev, S. et al. Familial prion disease with Alzheimer disease-like tau pathology and clinical phenotype. *Ann. Neurol.* **69**, 712–720 (2011).
84. Fong, J. C. et al. Genetic prion disease caused by PRNP Q160X mutation presenting with an orbitofrontal syndrome, cyclic diarrhea, and peripheral neuropathy. *J. Alzheimers Dis. JAD* **55**, 249–258 (2017).
85. Bommarito, G. et al. A novel prion protein gene-truncating mutation causing autonomic neuropathy and diarrhea. *Eur. J. Neurol.* **25**, e91–e92 (2018).
86. Mead, S. et al. A novel prion disease associated with diarrhea and autonomic neuropathy. *N. Engl. J. Med.* **369**, 1904–1914 (2013).
87. Capellari, S. et al. Two novel *PRNP* truncating mutations broaden the spectrum of prion amyloidosis. *Ann. Clin. Transl. Neurol.* **5**, 777–783 (2018).
88. Matsuzono, K. et al. A novel familial prion disease causing pan-autonomic-sensory neuropathy and cognitive impairment. *Eur. J. Neurol.* **20**, e67–e69 (2013).
89. Jansen, C. et al. Prion protein amyloidosis with divergent phenotype associated with two novel nonsense mutations in PRNP. *Acta Neuropathol.* **119**, 189–197 (2010).

Acknowledgements This study was performed under ethical approval from the Partners Healthcare Institutional Research Board (2013P001339/MGH) and the Broad Institute Office of Research Subjects Protection (ORSP-3862) in compliance with all relevant ethical regulations; written informed consent was obtained from all research participants. We thank all of the research participants for contributing their data. E.V.M. acknowledges support from the National Institutes of Health (NIH) (F31 AI122592) and an anonymous organization. gnomAD data aggregation was supported primarily by the Broad Institute, gnomAD analysis was supported in part by NIDDK U54 DK105566, and development of LOFTEE by NIGMS R01 GM104371. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. E.L.G.H. is funded by the Wellcome Trust (102627, 210561), the Medical Research Council (M009017), Higher Education Funding Council for England Catalyst, Barts Charity (845/1796), Health Data Research UK (for London substantive site), and research delivery support from the NHS National Institute for Health Research Clinical Research Network (North Thames). N.W. is supported by a Rosetrees and Stoneygate Imperial College Research Fellowship. The results published here are in part based upon data: (1) generated by The Cancer Genome Atlas managed by the NCI and NHGRI (accession: phs000178.v10.p8). Information about TCGA can be found at http://cancergenome.nih.gov; (2) generated by the Genotype-Tissue Expression Project (GTEx) managed by the NIH Common Fund and NHGRI (accession: phs000424.v7.p2); (3) generated by the Exome Sequencing Project, managed by NHLBI; (4) generated by the Alzheimer's Disease Sequencing Project (ADSP), managed by the NIA and NHGRI (accession: phs000572.v7.p4). We thank J. Kaprio and M. Kurki (Finnish Twins AD cohort) and Academy of Finland grant 312073, and Ruth McPherson

(Ottawa Genomics Heart Study) for providing information on individuals with *PRNP*-truncating variants. We thank J. B. Carroll, K. Heilbron, J. Fah Sathirapongsasuti, and L. C. Francioli for comments and suggestions. A subset of the analyses reported here originally appeared as a blog post on CureFFI.org (http://www.cureffi.org/2018/09/12/lof-and-drug-safety/).

Author contributions Conceived and designed the study: E.V.M., S.L.S., D.G.M. Performed analysis: E.V.M., K.J.K., H.C.M., B.B.C., N.W., D.R. Supervised the research: J.A., R.C.T., D.A.v.H., M.J.D., S.L.S., D.G.M. Provided data: gnomAD consortium (Genome Aggregation Database Production Team and Genome Aggregation Database Consortium), H.C.M., R.C.T., D.A.v.H. Wrote the paper: E.V.M. Edited and approved the final manuscript: all authors

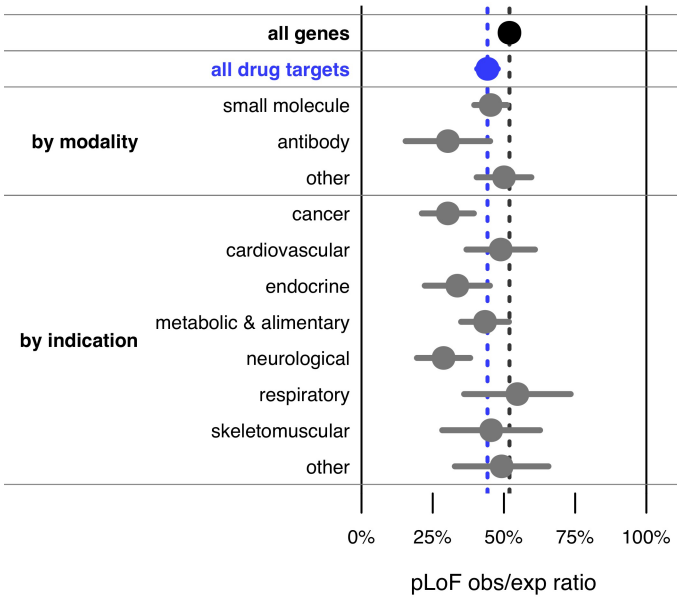
Competing interests E.V.M. has received research support in the form of charitable contributions from Charles River Laboratories and Ionis Pharmaceuticals, and has consulted for Deerfield Management. K.J.K. is a shareholder of Personalis. H.C.M., B.B.C., M.W., D.R. and J.A. have no competing interests to declare. R.C.T. serves on the Scientific Advisory Board of Ipsen Ltd and has current funding from the Wellcome Trust and the National Institute for Health Research UK. D.A.v.H. is a shareholder of Nexpep Pty Ltd; has current or recent research funding from Wellcome Trust, Medical Research Council UK, National Institute for Health Research UK, Alnylam Pharmaceuticals; and serves on the Population & Systems Medicine Board of the Medical Research Council UK. MJD is a founder of Maze Therapeutics. S.L.S. serves on the Board of Directors of the Genomics Institute of the Novartis Research Foundation ('GNF'); is a shareholder and serves on the Board of Directors of Jnana Therapeutics; is a shareholder of Forma Therapeutics; is a shareholder and advises Decibel Therapeutics and Eikonizo Therapeutics; serves on the Scientific Advisory Boards of Eisai Co., Ltd., Ono Pharma Foundation, Exo Therapeutics, and F-Prime Capital Partners; and is a Novartis Faculty Scholar. D.G.M. is a founder with equity in Goldfinch Bio, and has received research support from AbbVie, Astellas, Biogen, BioMarin, Eisai, Merck, Pfizer, and Sanofi-Genzyme.

Additional information

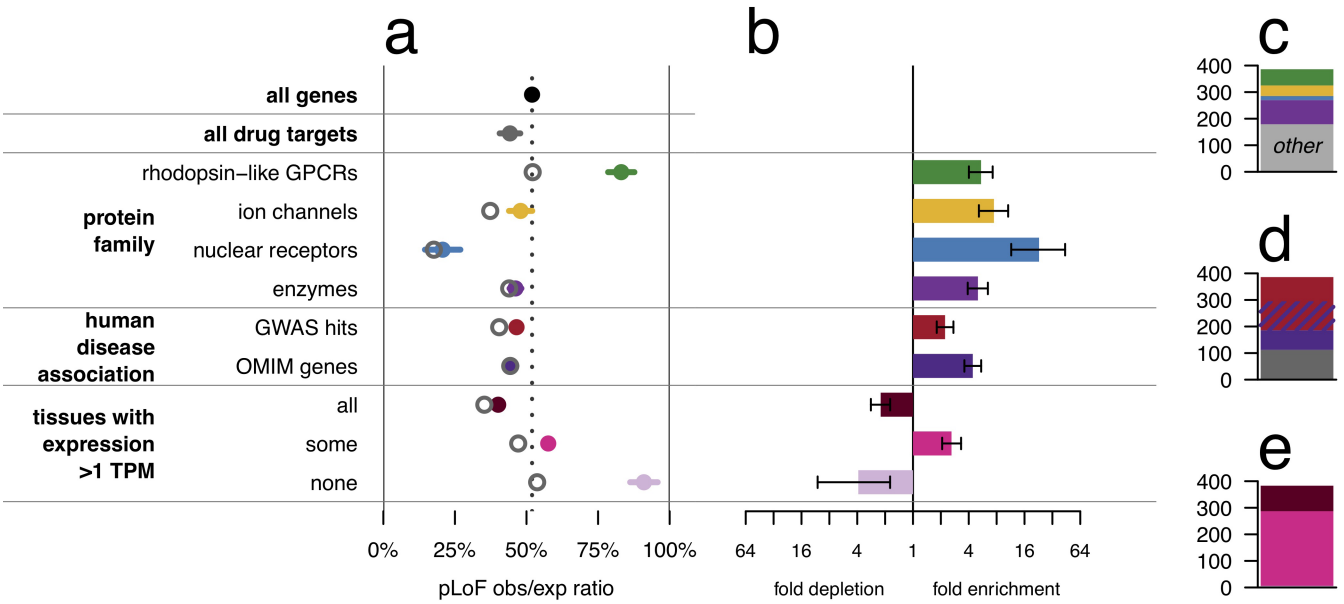
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-020-2267-z.

Correspondence and requests for materials should be addressed to E.V.M. or D.G.M.

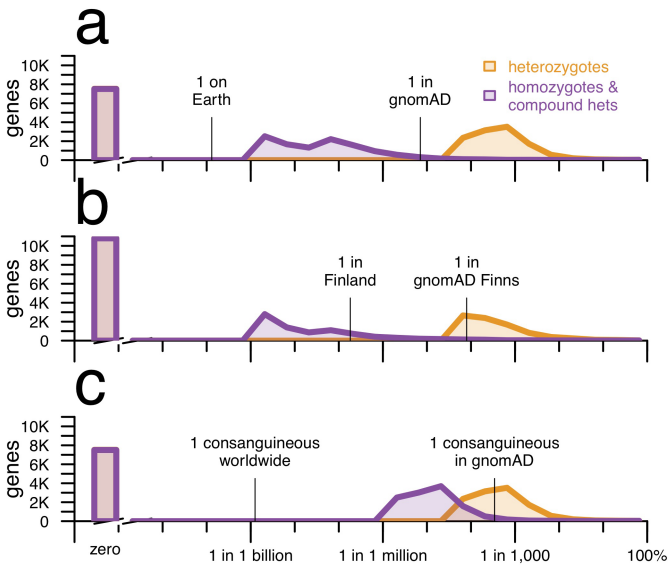
Peer review information *Nature* thanks Paul de Bakker and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.



Extended Data Fig. 1 | Drug target constraint by modality and indication. Mean (dots) and 95% confidence interval (line segments) for constraint in subsets of drug-targets sets (data sources and number of genes for each list are provided in Extended Data Table 1). Modality information was extracted from DrugBank and indication information from ATC codes; see Extended Data Table 1 for details.



Extended Data Fig. 2 | Drug-target gene set confounding. a, Forest plot of means (dots) and 95% confidence intervals of the mean (line segments) for gene sets evaluated for confounding with drug-target status. Data sources and number of genes for each list are provided in Extended Data Table 1. LoF obs/exp ratios differ significantly from the set of all genes for four canonically druggable protein families (top), human disease-associated genes (middle), and genes by broadness of tissue expression (bottom). Within each class, the genes that are drug targets have a lower mean obs/exp ratio (hollow grey circles) than the class overall. **b**, The druggable protein families, disease-associated genes, and genes expressed in some tissues but not others are enriched several-fold among the set of drug targets. Bars indicate fold enrichment and error bars indicate 95% confidence intervals. **c**–**e**, Composition of drug targets when broken down by protein family (**c**), disease association (**d**), or broadness of tissue expression (**e**). The enriched classes account for most drug targets. In a linear model, after controlling for protein family, disease association status, and number of tissues with expression >1 transcript per million (TPM), drug targets are still more constrained than other genes (−8.0% obs/exp, nominal $P = 0.00011$, $t = -3.9$, df = 17,325 for the contribution of drug_target in the linear regression obs/exp ~ drug_target + family + dz_assoc + n_tissues), but the probable existence of additional unobserved confounders cautions against over-interpretation of this observation (see main text).



Extended Data Fig. 3 | Expected frequency of individuals with one or two null alleles for every protein-coding gene across different population models, with sample size held constant. This is identical to Fig. 2 except as follows. As noted in the Methods, one caveat about Fig. 2 is that the sample size is larger for the plots using all gnomAD exomes (Fig. 2a, c) than for Finnish exomes (Fig. 2b). This figure shows the same analysis, but with the global gnomAD population downsampled to 10,824 randomly chosen exomes so that the same size is identical to that of Finnish exomes. Computation of $P = 1 - \mathrm{sqrt}(q)$ as described in the Methods is computationally expensive for downsampled datasets because it requires individual-level genotypes. Instead, this analysis uses 'classic' CAF, which is simply the sum of allele frequencies of all high-confidence pLoF variants each at allele frequency <5%, capped at a total of 100%, for both global and Finnish exomes. The results show that even when the sample size is held constant, the number of genes with zero pLoF variants observed is higher in a bottlenecked population than in a mostly outbred population. A constant *y* axis with no axis breaks is used in this figure to make this difference more clearly visible.

Extended Data Table 1 | Data sources for gene lists used in this study

list	*N*	*N**	reference / criteria
All	19,194	17,604	HGNC[52].
Olfactory receptors	371	325	Mainland et al[53].
Homozygous LoF tolerant	330	325	≥2 different high-confidence pLoF variants each homozygous in ≥1 individual in gnomAD exomes.
Autosomal recessive	527	519	Blekhman et al[54].
Autosomal dominant	307	305	Blekhman et al[54].
Essential in culture	683	659	Hart et al[24].
ClinGen haploinsufficient	294	288	ClinGen Dosage Sensitivity Map[18] level 3
Approved drug targets	386	383	DrugBank 5.0 XML release[17] (acc. Sep 12, 2018); Top-ranked mechanistic target of approved drugs. group==’approved’, target.attrib[‘position’] == '1', known-action==’yes’
Positive targets	143	142	DrugBank action: activator, agonist, chaperone, cofactor, gene replacement, inducer, partial agonist, positive allosteric modulator, positive modulator, potentiator, stimulator
Negative targets	243	241	DrugBank action: antagonist, blocker, degradation, inhibitor, inverse agonist, negative modulator, neutralizer, suppressor
Other & unknown (effect)	94	94	DrugBank action other or unlisted.
Small molecule	176	175	DrugBank type == ‘small’
Antibody	18	18	DrugBank type == ‘biotech’ and ‘Antibodies’ in categories
Other (modality)	35	35	DrugBank type == ‘biotech’ and ‘Antibodies’ not in categories
Oncology	45	45	ATC level 1 code L
Cardiovascular	38	38	ATC level 1 code C
Endocrine	24	24	ATC level 1 code G or H
Metabolic & alimentary	38	38	ATC level 1 code A
Neurology	35	35	ATC level 1 code N
Respiratory	12	11	ATC level 1 code R
Skeletomuscular	14	14	ATC level 1 code M
Other (indication)	29	28	ATC level 1 code B, D, J, P, S, or V
Rhodopsin-like GPCRs	689	604	HGNC[52] gene set 140: “G protein-coupled receptors, Class A rhodopsin-like”.
Ion channels	326	323	HGNC gene set 177: “Ion channels”[52].
Nuclear receptors	48	47	IUPHAR/BPS Guide to Pharmacology “Nuclear receptors”[55] .
Enzymes	1,178	1,144	IUPHAR/BPS Guide to Pharmacology “Enzymes”[55].
GWAS hits	6,336	6,080	GWAS Catalog[56] MAPPED_GENE column (P < 5-e8)
OMIM genes	3,367	3,294	OMIM[57] (acc. June 11, 2019) phenotypes with MIM number, lacking ‘?’, ‘{‘, ‘[‘, “response”, “susceptibility”, or “somatic”.
All (tissues)	7,931	7,550	>1 TPM in all 53 tissues in GTEx[40] v7
Some (tissues)	9,698	9,009	>1 TPM in >0 and <53 tissues in GTEx[40] v7
None (tissues)	1,076	776	>1 TPM in 0 tissues in GTEx[40] v7
Mouse heterozygous lethal knockout	401	395	MouseMine[23]

For analysis, only protein-coding genes with unambiguous mapping to current approved gene symbols were used; numbers in the table reflect this. Values in the *N* column indicate totals from the full universe of 19,194 genes; values in the *N** column indicate the subset of genes with non-missing constraint values, used for Fig. 1 and Extended Data Figs. 1, 2. The following references are cited in the table: refs. [17,18,23,24,40,52–57].

Extended Data Table 2 | Spectrum of tolerance to genetic inactivation among human drug targets

drug class	example	gene	obs/exp pLoF
topoisomerase I inhibitors	irinotecan	*TOP1*	0% (0/50.5)
M1-selective antimuscarinics	pirenzepine	*CHRM1*	0% (0/14.1)
cytoskeleton disruptors	paclitaxel	*TUBB*	6% (1/16.4)
non-steroidal anti-inflammatory drugs	aspirin	*PTGS2*	10% (3/29.7)
statins	atorvastatin	*HMGCR*	13% (6/46.3)
phosphodiesterase 5 inhibitors	sildenafil	*PDE5A*	33% (16/47.8)
antifolates	methotrexate	*DHFR*	38% (4/10.5)
proton pump inhibitors	omeprazole	*ATP4A*	52% (25/47.9)
antiplatelets	clopidogrel	*P2RY12*	66% (5/7.6)
H1 antihistamines	cetirizine	*HRH1*	76% (11/14.5)
angiotensin converting enzyme inhibitors	benazepril	*ACE*	87% (62/71.3)
cholesterol-lowering antibodies	alirocumab	*PCSK9*	98% (26/26.5)

Example targets are arranged from the most intolerant (top) to the most tolerant (bottom) of inactivation.

Extended Data Table 3 | Curation of pLoF variation in six neurodegenerative disease genes

gene	length (bp)	pLoF obs/exp	cumulative pLoF allele frequency		pLoF heterozygote frequency	GoF disease genetic prevalence
			before filtering & curation	after filtering & curation		
HTT	9,426	8.2%	6.2%	0.013%	1 in 3,800	1 in 2,400-4,400
LRRK2	7,581	41%	0.23%	0.09%	1 in 500	1 in 3,300
MAPT	2,328	0%[a]	14%	0%	—	1 in 5,000 – 31,000
PRNP	759	99%[b]	0.0035%	0.0021%	1 in 18,000	1 in 50,000
SNCA	420	0%	0.0012%	0%	—	1 in 360,000
SOD1	462	18%	0.0060%	0.0038%	1 in 26,000	1 in 27,000-83,000

Shown are the coding sequence length (base pairs, bp), constraint value (pLoF obs/exp) after filtering and curation, cumulative allele frequency before and after filtering and manual curation, estimated frequency of true pLoF heterozygotes in the population, and genetic prevalence (population frequency including pre-symptomatic individuals) of the GoF disease associated with the gene. Genetic prevalence calculations are described in Extended Data Table 4, and variant curation details are provided in Supplementary Table 1, except for *LRRK2*, which is described in detail elsewhere[49].

[a]Constitutive brain-expressed exons only.

[b]*PRNP* codons 1–144; see Fig. 3c for rationale.

Extended Data Table 4 | Estimation of genetic prevalence for GoF genetic neurodegenerative diseases

gene	basis for genetic prevalence estimation	estimate
HTT	A reported HD incidence of 0.38 cases per 100,000 per year based on meta-analysis[58] multiplied by an average age at death of ~60 for the most common CAG lengths[59]. Finally, a genetic screen of a general population sample[60] found ≥40 CAG repeat alleles, which are presumed to be fully penetrant, in 3 individuals out of 7,315, for a genetic prevalence of 1 in 2,438.	1 in 4,386
HTT	Prevalence of 13.7 per 100,000 symptomatic plus 81.6 per 100,000 at 25-50% risk in an exhaustive ascertainment study[61]. Assuming there are twice as many individuals at 25% risk as at 50% risk, then on average 33.3% of the 81.6, or 27.1 per 100,000 have the mutation. Thus, 13.7 + 27.1 = 40.8 per 100,000 individuals have an *HTT* CAG expansion.	1 in 2,451
HTT	A genetic screen of a general population sample[60] found ≥40 CAG repeat alleles, which are presumed to be fully penetrant, in 3 individuals out of 7,315.	1 in 2,438
LRRK2	Based on meta-analysis[62], Parkinson's disease (PD) has an estimated prevalence of 1,903 per 100,000 at age ≥80, meaning the general population's lifetime risk of PD is ~1.9%. It is generally stated that about 10% of PD cases are "familial" and the remainder sporadic; in a diverse worldwide case series, *LRRK2* mutations were found in 179/14,253 (1.3%) sporadic cases and 201/5,123 (3.9%) familial cases[63], implying that *LRRK2* mutations are present in ~1.6% of all PD cases. Thus, *LRRK2* mutations account for a 1.6% * 1.9% = ~0.030% lifetime risk of PD in the general population[a].	1 in 3,300
MAPT	Pathogenic *MAPT* mutations can present with a variety of clinical phenotypes, and common *MAPT* haplotypes are associated with risk for a variety of different neurodegenerative disorders; we were unable to identify any studies of genetic prevalence nor any large case series for any *MAPT*-associated phenotype. As a crude estimate, frontotemporal dementia has a reported incidence of 2.7-4.1 per 100,000 per year[64] with typical age at death of perhaps 60, and *MAPT* mutations accounting for 5-20% of familial cases, and familial cases accounting for 40% of all cases[65]. Multiplying all these figures results in range of 0.0032% to 0.020%.	1 in 5,000 – 31,000
PRNP	We recently considered the lifetime risk of genetic prion disease in detail[36]. Prion disease (including sporadic, genetic, and acquired) causes ~1 in 5,000 people based on either death certificate analysis or division of disease incidence by the overall death rate[41,66]. ~10% of cases are attributable to *PRNP* variants with evidence for Mendelian segregation (although additional cases harbor lower-penetrance variants)[36]. Thus, we expect a genetic prevalence of 1 in 50,000. On the order of ~1 in 100,000 people in gnomAD and 23andMe harbor high-penetrance *PRNP* variants[36,41], although as noted above, we expect these datasets to be depleted compared to the population at birth, because prion disease is rapidly fatal and many individuals in these databases are above the typical age of onset.	1 in 50,000
SNCA	As explained above for *LRRK2*, we assumed a 1.9% lifetime risk of Parkinson's disease (PD) in the general population, with 10% of cases being familial. *SNCA* point mutations, duplications, and triplications all appear to be highly penetrant, and in a familial PD case series these accounted for 103/709 = 15% of individuals[67]. Thus, we estimate that *SNCA* mutations account for a 1.9% * 10% * 15% = 0.00028% risk of PD in the general population.	1 in 360,000
SOD1	*SOD1* mutations are believed to account for ~12% to 24% of familial ALS[68,69] and 1% of sporadic ALS[68,70]. One a meta-analysis found that ~4.6% of ALS is familial[71], although a figure of 10% is also often used[72]. These figures imply that ~1.5 – 3.3% of all ALS is attributable to *SOD1*. The overall incidence of ALS is reported at ~1.6 – 2.2 per 100,000 per year[73,74], so the incidence of *SOD1* ALS might be estimated at ~0.024 – 0.073 per 100,000 per year. Age at death of ~50 is around average for many *SOD1* mutations[69], implying a 1.2 – 3.7 per 100,000 population prevalence of pathogenic *SOD1* mutations.	1 in 27,000-83,000

Data sources were identified by PubMed and Google Scholar searches. Genetic prevalence was defined as the proportion of the population at birth carrying a mutation and destined to later develop disease, and estimated as described for each gene. The following references are cited in the table: refs. [36,41,58–74].

[a]It is important to consider how this figure relates to the penetrance of *LRRK2* mutations, as *LRRK2* variants appear to occupy a spectrum of penetrance[75]. Some variants exhibit Mendelian segregation with disease[76,77], implying high risk; the G2019S variant is estimated to have approximately 32% penetrance[78]; and other common variants are risk factors with odds ratios of only around 1.2 estimated through genome-wide association studies (GWAS)[79]. The GWAS-implicated common variants were not included in the case series on which our estimate is based[63], but G2019S does account for most cases in that series. Because the 0.03% estimate here is based on counting symptomatic cases rather than asymptomatic individuals, it will appropriately underestimate the number of G2019S carriers. In essence, in this calculation each G2019S carrier in the population only counts as 1/3 of a person, because they have only a 1/3 probability of developing a disease. It is therefore appropriate that our estimate of genetic prevalence (0.03%) is actually lower than double the allele frequency of G2019S in gnomAD (0.1%).

Extended Data Table 5 | Details of *PRNP*-truncating variants

variant	allele count	neurological phenotype	comments	reference
G20Gfs84X	1	healthy	As previously reported.	[41]
R37X	2	healthy, unknown	One previously reported, one new.	[41], this work
Q41X	1	unknown		this work
H69 frameshifts	2	N/A	False variant calls in gnomAD, apparent alignment artifact due to octapeptide repeat region.	this work
Q75X	1	healthy	As previously reported	[41]
W81X	1	unknown		this work
W99X	1	unknown		this work
G131X	1	healthy	The presence of this variant in the ExAC database was previously reported, but without phenotype information. We now report that this individual is a 77-year-old male, cognitively well with no family history of dementia. Ascertained as a case in a study of coronary artery disease, this individual has hypertension and well-controlled dyslipidemia and has undergone one bypass surgery. He has two adult children.	[41], this work
Y145X	1	dementia		[81]
Q160X	5	dementia		[82–84]
Y162X	1	dementia		[85]
Y163X	7	dementia		[86,87]
Y169X	2	dementia		[87]
D178Efs25X	1	dementia		[88]
Q186X	1	dementia		[41]
Y226X	1	dementia		[89]
Q227X	1	dementia		[89]
L234Pfs7X	1	dementia	Ascertained as a female case in the Finnish twins Alzheimer disease cohort. Died at age >90 of proximal cause pneumonia, ultimate cause diagnosed as Alzheimer disease based on clinical examination only. Had a dizygotic twin not included in gnomAD.	this work

Allele count for variants from the literature in Fig. 3c is the total number of definite or probable cases with sequencing performed in the studies cited in this table. The L234Pfs7X variant changes the C-terminal GPI signal of prion protein from SMVLFSSPPVILLISFLIFLIVGX to SMVPSPLHLX. This new sequence does not adhere to the known rules of GPI anchor attachment[80]: GPI signals must contain a 5–10-polar-residue spacer followed by 15–20 hydrophobic residues. Thus, this frameshifted prion protein would be predicted to be secreted and thus may be pathogenic, explaining the Alzheimer's disease diagnosis in this individual. However, it is also possible that the new C-terminal sequence found here interferes with prion formation, and/or that this variant is incompletely penetrant, and that the diagnosis of Alzheimer's disease in this individual is merely a coincidence. The following references are cited in the table: refs. [41,81–89].

nature research

Corresponding author(s): Eric Vallabh Minikel and Daniel G. MacArthur

Last updated by author(s): 11/13/2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Analyses utilized Python 2.7.10 and R 3.5.1. Data and code sufficient to produce the plots and analyses in this paper are available at https://github.com/ericminikel/drug_target_lof
Data analysis	Analyses utilized Python 2.7.10 and R 3.5.1. Data and code sufficient to produce the plots and analyses in this paper are available at https://github.com/ericminikel/drug_target_lof

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Analyses utilized Python 2.7.10 and R 3.5.1. Data and code sufficient to produce the plots and analyses in this paper are available at https://github.com/ericminikel/drug_target_lof

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	This study was opportunistic, and involved secondary use of all available genome and exome data. No sample size was predetermined. Our flagship analysis of gnomAD loss-of-function variants (Karczewski et al, https://doi.org/10.1101/531210) indicates that the dataset is well-powered to examine constraint against such variants — for instance, 72% of genes have at least 10 pLoF variants expected in this sample size based on mutation rates.
Data exclusions	Sample QC and variant QC for the gnomAD database are described extensively by Karczewski et al, https://doi.org/10.1101/531210. Notably, individuals with severe pediatric disease, and known first disease relatives of those with severe pediatric disease were excluded.
Replication	We did not attempt to reproduce any findings in a separate dataset, as no other exome or genome sequencing dataset of comparable size exists.
Randomization	As this was a population-based study, and not a case-control study, no randomization was performed.
Blinding	As this was a population-based study, and not a case-control study, blinding was not relevant.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	As an opportunistic collection of data, the participants in gnomAD were not selected based on age, gender, or genotypic information. As described above, individuals with severe pediatric disease, and known first disease relatives of those with severe pediatric disease were excluded. The population and dataset inclusion criteria are described in more detail by Karczewski et al, https://doi.org/10.1101/531210
Recruitment	The generation of the gnomAD database was an opportunistic secondary use study, we did not recruit any participants. The study is described in more detail by Karczewski et al, https://doi.org/10.1101/531210
Ethics oversight	This study was performed under ethical approval from the Partners Healthcare Institutional Research Board (2013P001339/MGH) and the Broad Institute Office of Research Subjects Protection (ORSP-3862) in compliance with all relevant ethical regulations; informed consent was obtained from all research participants.

Note that full information on the approval of the study protocol must also be provided in the manuscript.