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

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

1. Broad Institute of MIT and Harvard, Cambridge, MA, 02142, USA
2. Analytical and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, 02114, USA
3. Program in Biological and Biomedical Sciences, Harvard Medical School, Boston, MA, 02115, USA
4. Prion Alliance, Cambridge, MA, 02139, USA
5. Wellcome Sanger Institute, Hinxton, Cambridgeshire, CB10 1SA, UK
6. National Heart and Lung Institute and MRC London Institute of Medical Sciences, Imperial College London, London, SW7 2AZ, UK
7. 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, EC1M 6BQ, UK
8. Faculty of Life Sciences and Medicine, King's College London, London, WC2R 2LS, UK
9. Blizard Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, E1 2AT, UK
10. Department of Chemistry & Chemical Biology, Harvard University, Cambridge, MA, 02138, USA

†To whom correspondence should be addressed: eminikel@broadinstitute.org or danmac@broadinstitute.org
*A full list of authors appears at the end of this paper

SUMMARY

Naturally occurring human genetic variants predicted to cause loss of function of protein-coding genes provide an *in vivo* model of human gene inactivation that complements cell and model organism knockout studies. Here we investigate the application of human loss-of-function variants to assess genes as candidate drug targets, with three key findings. First, even essential genes, where 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 ~1,000 times those available at present. Third, automated variant annotation and filtering are powerful, but manual curation remains critical for removing artifacts and making biological inferences, and is a prerequisite for recall-by-genotype efforts. Our results provide a roadmap for human “knockout” studies and should guide interpretation of loss-of-function variants in drug development.

MAIN TEXT

Human genetics is an increasingly critical source of evidence guiding the selection of new targets for drug discovery[1]. Most drug candidates that enter clinical trials eventually fail for lack of efficacy[2], and while *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 become approved drugs[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 pharmacologically inhibiting the drug's target[5]. While 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 provided catalogs of predicted loss-of-function (pLoF) variants — nonsense, essential splice site, and frameshift variants expected to result in a non-functional allele — in humans. Such databases present a new opportunity to study the effects of pLoF variation in genes of interest and to identify individuals with pLoF genotypes in order to understand gene function or disease biology, or to assess potential for therapeutic targeting. While many variants initially annotated as pLoF do not, in fact, cause a loss of 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 quantification of gene essentiality in humans through a metric known as *constraint*[10–13]. Specifically, the rate at which *de novo* pLoF mutations arise in each gene is predicted based on DNA mutation rates[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 simply constraint score — measures how strongly purifying natural selection has removed such variants from the population. Annotation of pLoF variants remains imperfect, and continued improvements are being made[14], but the fact that constraint usefully measures gene essentiality is 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].

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% vs. 52%, *P*=0.00028), 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 modality,

indication, or effect, 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. 1). Although controlling for these correlations does not abolish the trend of stronger constraint among drug targets, the correlation of so many observed variables with a gene's status as drug target argues that many unobserved variables likely 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 obs/exp values lower 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 particular class or subset of drugs, we examined constraint in several well-known example drug targets (Table 1). 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 "knockout" individuals

While 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 (Online Methods), and then used this to estimate the expected frequency of two-hit individuals under different population structures (Fig. 2) absent natural selection.

Whereas gnomAD is now large enough to include at least one pLoF heterozygote for the majority (15,317/19,194; 79.8%) of genes, ascertainment of total "knockout" individuals in outbred populations will require 1,000-fold larger sample sizes for most genes: the median gene has an expected two-hit frequency of just 6 per billion (Fig. 2a). Even if every human on Earth were sequenced, there are 4,728 genes (25%) for which identification of even one two-hit individual would not be expected in outbred populations. Intuitively, because today's gnomAD sample size 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 the utility of bottlenecked populations for knockout discovery, 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 for a pre-specified gene of interest appears equally or more difficult in Finns than in outbred populations (Fig. 2b and Extended Data Fig. 2), because rare variants not present in a founder have been effectively removed from the population.

Finally, we considered consanguineous individuals, where 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 5 per million for the median gene (Fig. 2c).

These projections allow us to draft a roadmap for discovery of human knockouts (Fig. 2d-e). Of 19,194 genes, 3,367 (18%) already have a human disease association annotated in OMIM, although we note that the discovery of LoF individuals in population databases will still be valuable for assessing penetrance and for identifying LoF syndromes associated with known GoF genes. There are 3,421 genes (18%) without known human disease association that have two-hit pLoF genotypes reported in gnomAD[7], ELGH[26], PROMIS[27], DeCODE[28], or UK BioBank[29], suggesting this genotype may be tolerated. An additional 2,190 genes (11%) can be inferred likely intolerant of heterozygous inactivation (pLI > 0.9) in gnomAD, and would be expected to be enriched for genes with severe heterozygous and lethal homozygous LoF phenotypes. Another 2,781 genes (14%) have no pLoF variants yet 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 inferred, nor have two-hit genotypes been observed, nor is a human disease phenotype known. We projected the sample sizes required to identify “knockout” individuals for these genes (Fig. 2e). In outbred populations, current sample size would need to be increased by approximately 1,000-fold before ascertainment of a single two-hit LoF individual would be expected for the typical gene. In contrast, 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.

These calculations are based on variants annotated as predicted LoF 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, leading to underestimation of cumulative LoF allele frequency. Overall, however, our calculations likely represent an upper bound on the total frequency of two-hit individuals in the population. 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. More importantly, for some genes, complete knockout will not be 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 this 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 impact 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 disease), *MAPT* (tauopathies), *PRNP* (prion disease), *SOD1* (amyotrophic lateral sclerosis), and *LRRK2* and *SNCA* (Parkinson disease). The results (Table 2 and Fig. 3) illustrate four points about pLoF variant curation.

First, other things being equal, genes with longer coding sequences offer more opportunity for LoF variants to arise, and so tend to have a higher cumulative frequency 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 (Table 1).

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 impact 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 ~1 in 50,000 genetic prevalence[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 (~1 in 18,000). Similar arguments can be made for *HTT*, *LRRK2*, and *SOD1*. Of course, this does not rule out the possibility that heterozygous loss-of-function in these genes could be associated with less severe or less penetrant phenotypes.

Finally, careful inspection of the distributions of pLoF variants can reveal important error modes or disease biology. *HTT*, *MAPT*, and *PRNP* 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 artifacts (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 to date[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. *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 2), consistent with variants at codon ≥145 causing a pathogenic gain-of-function through change in localization (Fig. 3c). Within codon 1-144, *PRNP* is unconstrained, and no neurological phenotype has been identified in individuals with truncating variants to date, consistent with the hypothesis that N-terminal truncating variants are true LoF and are tolerated in a heterozygous state[41].

Discussion

The study of gene inactivation through human genetic databases can illuminate human biology and guide drug target selection, 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 span a spectrum from highly constrained to completely unconstrained. Why do some genes apparently tolerate pharmacological inhibition but not genetic inactivation? LoF variants, at least 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 believed to primarily reflect selection against heterozygotes[13], whose effective gene dosage may differ from that achieved by a drug. Constraint measures natural selection over centuries or millennia; our ancestors' environment presented different selective pressures than what we face today. Finally, the actions of small molecule drugs do not always 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 to date does not yet provide statistical evidence that this genotype is not tolerated, and, 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, though 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 the ascertainment and characterization of consanguineous cohorts, where 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 to characterize humans with genotypes of interest are only valuable if the variants in question are true LoF. 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 prior to any recontact efforts, and indeed, establishing methods for high-throughput functional validation[48] of LoF variants should be a high priority. Our curation of pLoF variants in neurodegenerative disease genes is limited by a lack of functional validation and detailed phenotyping; in a

companion paper we demonstrate a deeper investigation of the effects of LoF variants in *LRRK2*[49].

As the value of human genetics for drug discovery has been demonstrated repeatedly, we expect that drug development projects will increasingly be accompanied by efforts to study the phenotypes of human carriers of LoF variants. Because the cost of drug discovery is driven overwhelmingly by failure[50], successful interpretation of LoF data to select the right targets and the right clinical pathways will yield an outsize benefit for research productivity and, ultimately, human health.

DISPLAY ITEMS

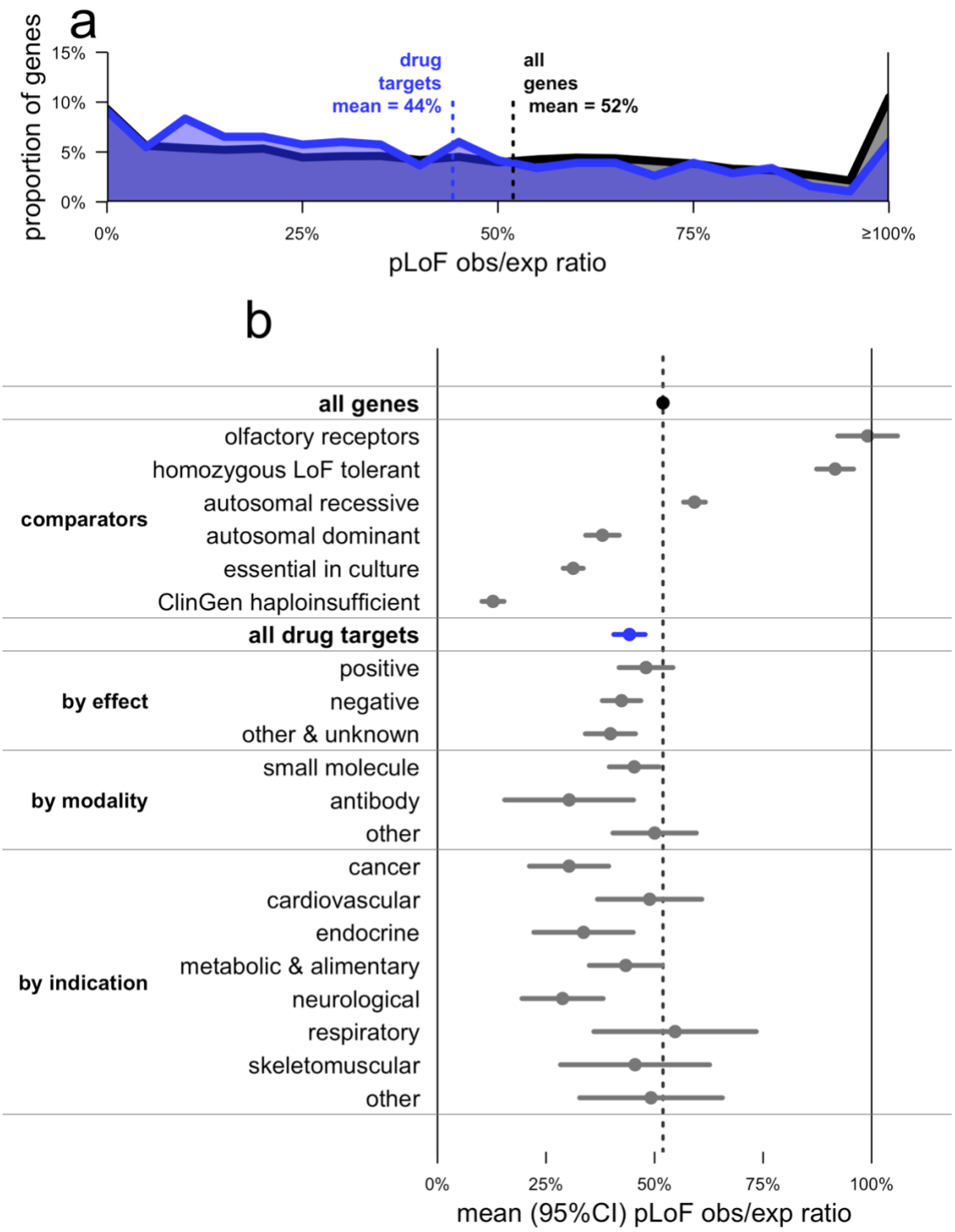


Figure 1 | pLoF constraint in drug targets. a) *Histogram of pLoF obs/exp for all genes (black) versus drug targets (blue).* ***b)*** *Forest plot of means (dots) and 95% confidence intervals indicating our certainty about the mean (line segments), for pLoF obs/exp ratio in the indicated*

gene sets. See Online Methods for data sources. For drug effect, 'positive' indicates agonist, activator, inducer, etc., while negative indicates antagonist, inhibitor, suppressor, etc.

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 (NSAIDs)	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 (ACE) inhibitors	benazepril	*ACE*	87% (62/71.3)
PCSK9 antibodies	alirocumab	*PCSK9*	98% (26/26.5)

Table 1 | Spectrum of tolerance to genetic inactivation among human drug targets. *Example targets are arranged from most intolerant of inactivation (top) to most tolerant (bottom).*

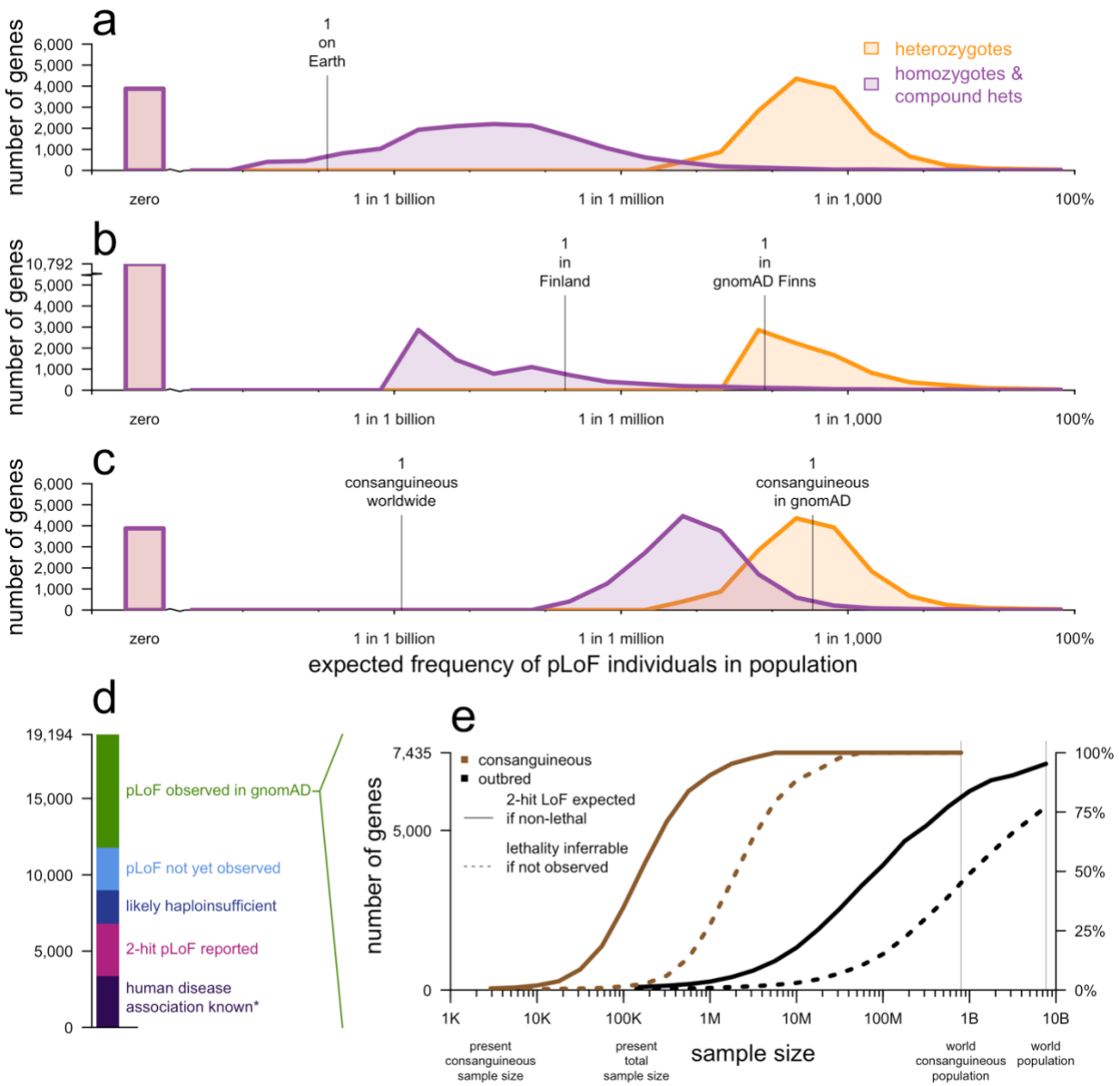


Figure 2 | Prospects for discovery of human "knockouts". *Each panel* ***a-c*** *shows a histogram where the y axis is number of genes and the x axis shows the theoretically expected population frequency of single hit heterozygotes (orange), versus two-hit homozygotes and compound heterozygotes (purple). Zero indicates the number of genes where no pLoF variants have been observed.* ***a)*** *Outbred populations, under random mating — heterozygotes have frequency 2p(1-p) and two-hit individuals have frequency* p^2*. The value of p is taken from all gnomAD exomes.* ***b)*** *Finnish individuals, an example of a bottlenecked population. Single and two-hit frequencies are again 2p(1-p) and* p^2 *but p is based on Finnish exomes only.* ***c)*** *Consanguineous individuals with a = 5.8% of their genome autozygous (both chromosomes inherited from the same recent ancestor); heterozygote frequency is 2p(1-p) and two-hit frequency is* $(1-a)p^2 + ap$*. See Online Methods for details.* ***d)*** *Current status of pLoF or disease association discovery for all protein-coding genes.* ***e)*** *For genes with pLoF observed in gnomAD (top category from d), 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 consanguineous and outbred individuals.

			cumulative pLoF allele frequency			
gene	**length (bp)**	**pLoF obs/exp**	**before filtering & curation**	**after filtering & curation**	**pLoF heterozygote frequency**	**GoF disease genetic prevalence**
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%*	14%	0%	—	1 in 5,000 – 31,000
PRNP	759	99%**	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

Table 2 | Curation of pLoF variation in six neurodegenerative disease genes. *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 gain-of-function (GoF) disease associated with the gene. Genetic prevalence calculations are described in Extended Data Table 2, and variant curation details are provided in Supplementary Table 1, except for LRRK2 which is described in detail in Whiffin et al*[49]*. *Constitutive brain-expressed exons only. **PRNP codons 1-144, see Fig. 3c for rationale.*

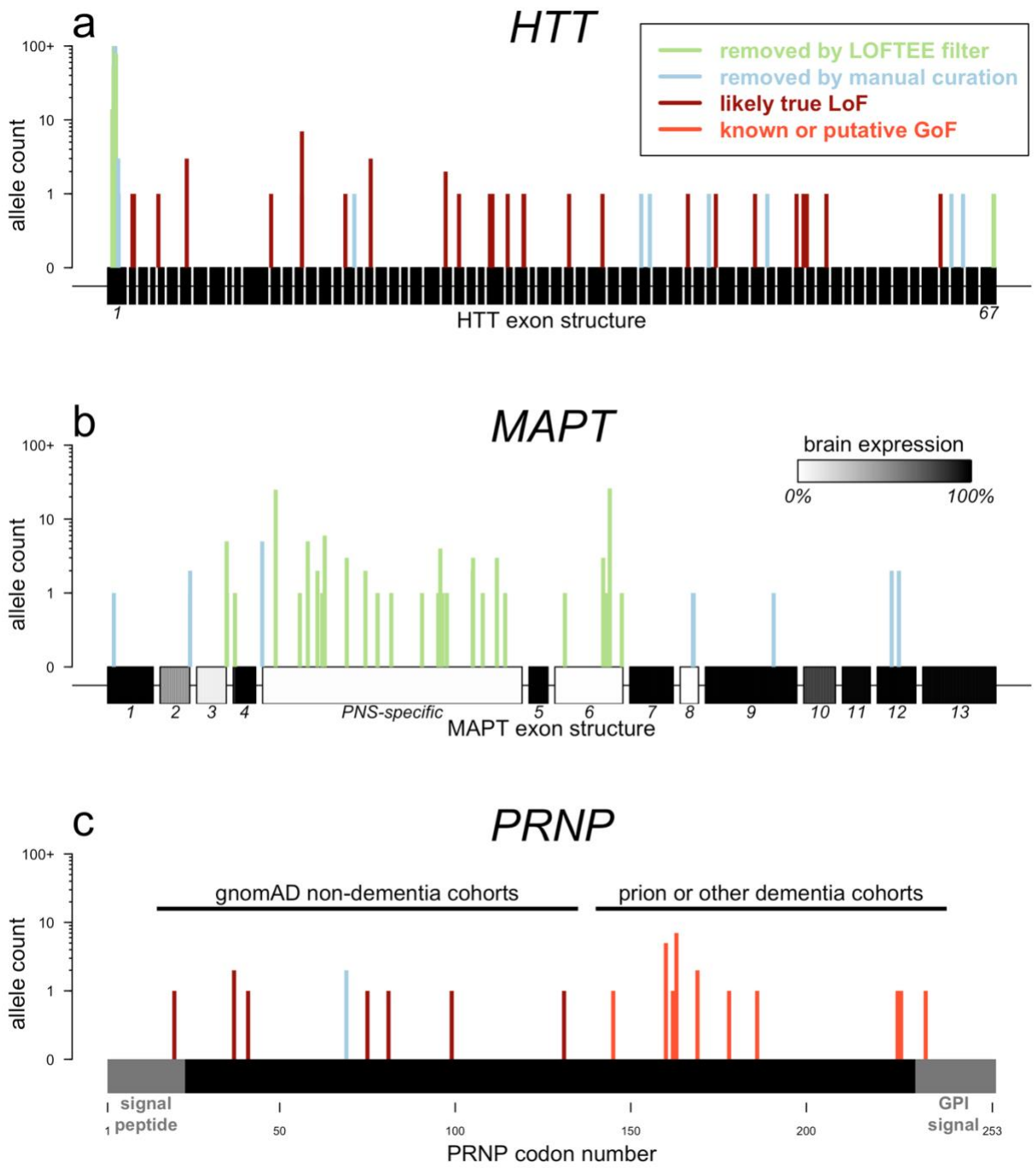


Figure 3 | Insights from non-random positional distributions of pLoF variants. a) *HTT,* ***b)*** *MAPT, with exon numbering and annotation from Andreadis*[51] *and brain expression data from GTEx*[40]*, and* ***c)*** *PRNP, a single protein-coding exon with domains removed by post-translational modification in gray, showing previously reported variants*[41] *as well as those newly identified in gnomAD and in the literature*[52,53]*. See text for interpretation, Extended Data Table 3 for details of PRNP variants, and Supplementary Table 1 for detailed variant curation results.*

REFERENCES (including for Online Methods and Extended Data)

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. *bioRxiv* 513945 (2019). doi:10.1101/513945
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.* Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv* 531210 (2019). doi:10.1101/531210
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. U. S. A.* **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. *bioRxiv* 554444 (2019). doi:10.1101/554444
15. Collins, R. L. *et al.* An open resource of structural variation for medical and population genetics. *bioRxiv* 578674 (2019). doi:10.1101/578674
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**, 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 Off. J. Int. Mamm. Genome Soc.* **26**, 325–330 (2015).
24. Hart, T. *et al.* Evaluation and Design of Genome-Wide CRISPR/SpCas9 Knockout Screens. *G3 Bethesda Md* **7**, 2719–2727 (2017).
25. Finer, S. *et al.* Cohort Profile: East London Genes & Health (ELGH), a community based population genomics and health study in people of British-Bangladeshi and -Pakistani heritage. *bioRxiv* 426163 (2018). doi:10.1101/426163
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.* (2019). doi:10.1056/NEJMoa1900907
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**, (2017).
32. Vallabh, S. M. Antisense oligonucleotides for the prevention of genetic prion disease. PhD dissertation. (Harvard University, 2019).
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. Bennett, C. F., Freier, S. M. & Mallajosyula, J. Modulation of alpha synuclein expression. (2014).
36. Minikel, E. V. *et al.* Age at onset in genetic prion disease and the design of preventive clinical trials. *Neurology* (2019). doi:10.1212/WNL.0000000000007745
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. EJHG* (2016). doi:10.1038/ejhg.2016.74
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 *et al.* 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. U. S. A.* **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.* Human loss-of-function variants suggest that partial LRRK2 inhibition is a safe therapeutic strategy for Parkinsons disease. *bioRxiv* 561472 (2019). doi:10.1101/561472
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).
51. Andreadis, A. Tau splicing and the intricacies of dementia. *J. Cell. Physiol.* **227**, 1220–1225 (2012).
52. Bommarito, G. *et al.* A novel prion protein gene-truncating mutation causing autonomic neuropathy and diarrhea. *Eur. J. Neurol.* **25**, e91–e92 (2018).
53. Capellari, S. *et al.* Two novel PRNP truncating mutations broaden the spectrum of prion amyloidosis. *Ann. Clin. Transl. Neurol.* **5**, 777–783 (2018).
54. Yates, B. *et al.* Genenames.org: the HGNC and VGNC resources in 2017. *Nucleic Acids Res.* **45**, D619–D625 (2017).
55. 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).
56. Blekhman, R. *et al.* Natural selection on genes that underlie human disease susceptibility. *Curr. Biol. CB* **18**, 883–889 (2008).
57. 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**, D1091–D1106 (2018).
58. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2017).
59. Pringsheim, T. *et al.* The incidence and prevalence of Huntington's disease: a systematic review and meta-analysis. *Mov. Disord. Off. J. Mov. Disord. Soc.* **27**, 1083–1091 (2012).
60. 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).
61. Kay, C. *et al.* Huntington disease reduced penetrance alleles occur at high frequency in the general population. *Neurology* **87**, 282–288 (2016).
62. Fisher, E. R. & Hayden, M. R. Multisource ascertainment of Huntington disease in Canada: prevalence and population at risk. *Mov. Disord. Off. J. Mov. Disord. Soc.* **29**, 105–114 (2014).
63. Pringsheim, T., Jette, N., Frolkis, A. & Steeves, T. D. L. The prevalence of Parkinson's disease: a systematic review and meta-analysis. *Mov. Disord. Off. J. Mov. Disord. Soc.* **29**, 1583–1590 (2014).
64. 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).
65. Onyike, C. U. & Diehl-Schmid, J. The epidemiology of frontotemporal dementia. *Int. Rev. Psychiatry Abingdon Engl.* **25**, 130–137 (2013).
66. Bang, J., Spina, S. & Miller, B. L. Frontotemporal dementia. *Lancet Lond. Engl.* **386**, 1672–1682 (2015).
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. Mead, S. *et al.* A novel prion disease associated with diarrhea and autonomic neuropathy. *N. Engl. J. Med.* **369**, 1904–1914 (2013).
86. Matsuzono, K. *et al.* A novel familial prion disease causing pan-autonomic-sensory neuropathy and cognitive impairment. *Eur. J. Neurol. Off. J. Eur. Fed. Neurol. Soc.* **20**, e67-69 (2013).
87. Jansen, C. *et al.* Prion protein amyloidosis with divergent phenotype associated with two novel nonsense mutations in PRNP. *Acta Neuropathol. (Berl.)* **119**, 189–197 (2010).

Online Methods

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 3.

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 (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.

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-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 populaton (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 Finns 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, since 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. Our approach naively treats genes with no pLoFs observed as having p=0, even though pLoFs might be discovered at a larger sample size. It also naively treats genes with one pLoF allele observed as having p=1/(2*125748), even though on average singleton variants have a true allele frequency lower than their nominal allele frequency[10]. We naively group all populations together, even though the distribution of populations sampled in gnomAD does not reflect the world population[7]; we believe 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. Finally, the sample size for all gnomAD exomes (Fig. 2a and 2c) 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 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 harbor 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 ~55, 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, 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 2.

Data and source code 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

Acknowledgments

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). We thank all of the research participants for contributing their data. EVM acknowledges support from the National Institutes of Health (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 National Institutes of Health. ELGH 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). NW 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 Jaakko Kaprio and Mitja 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 Jeffrey B. Carroll, Karl Heilbron, J. Fah Sathirapongsasuti, and Laurent 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/).

Extended Data

Extended Data Table 1 | Data sources for gene lists used in this study. *For analysis all lists were subsetted to protein-coding genes with unambiguous mapping to current approved gene symbols; numbers in the table reflect this. Note that the gene counts here reflect totals from the full universe of 19,194 genes; some numbers quoted in the main text reflect only the subset of genes with non-missing constraint values.*

List	*N*	Description
All	19,194	HGNC protein-coding genes[54].
Olfactory receptors	371	As reported by Mainland et al[55].
Homozygous LoF tolerant	330	Genes with at least two different high-confidence pLoF variants each found in a homozygous state in at least one individual in gnomAD exomes.
Autosomal recessive	527	OMIM disease genes deemed to follow autosomal recessive inheritance according to extensive manual curation by the Przeworski group[56].
Autosomal dominant	307	OMIM disease genes deemed to follow autosomal dominant inheritance according to extensive manual curation by the Przeworski group[56].
Essential in culture	683	Genes deemed essential in cultured cell lines based on CRISPR screens[24].
ClinGen haploinsufficient	294	Genes with sufficient evidence for dosage pathogenicity (level 3) as determined by the ClinGen Dosage Sensitivity Map[18]
Approved drug targets	386	Genes listed as the top-ranked mechanistic target of approved drugs in the DrugBank 5.0 XML release[17] (accessed Sep 12, 2018). Includes products approved by a variety of agencies including FDA, EMA, and Health Canada. Genes were extracted from the XML file using a custom python script with the criteria target.attrib['position'] == '1', known-action=='yes', and group=='approved'.
Positive targets	143	Action listed in DrugBank as: activator, agonist, chaperone, cofactor, gene replacement, inducer, partial agonist, positive allosteric modulator, positive modulator, potentiator, or stimulator
Negative targets	243	Action listed in DrugBank as: antagonist, blocker, degradation, inhibitor, inverse agonist, negative modulator, neutralizer, or suppressor
Other & unknown (effect)	94	Action not listed in DrugBank, or any action other than those listed above for positive and negative targets.
Small molecule	176	DrugBank type == 'small'
Antibody	18	DrugBank type == 'biotech' and DrugBank categories include 'Antibodies'
Other (modality)	35	DrugBank type == 'biotech' and DrugBank categories do not include 'Antibodies'
Oncology	45	ATC level 1 code L
Cardiovascular	38	ATC level 1 code C
Endocrine	24	ATC level 1 code G or H

Metabolic & alimentary	38	ATC level 1 code A
Neurology	35	ATC level 1 code N
Respiratory	12	ATC level 1 code R
Skeletomuscular	14	ATC level 1 code M
Other (indication)	29	ATC level 1 code B, D, J, P, S, or V
Rhodopsin-like GPCRs	689	HGNC gene set 140: "G protein-coupled receptors, Class A rhodopsin-like"[54].
Ion channels	326	HGNC gene set 177: "Ion channels"[54].
Nuclear receptors	48	IUPHAR/BPS Guide to Pharmacology "Nuclear receptors" list[57] .
Enzymes	1,178	IUPHAR/BPS Guide to Pharmacology "Enzymes" list[57].
GWAS hits	6,336	Closest gene to GWAS hits with P < 5-e8 in the EBI GWAS catalog (MAPPED_GENE column)[58].
OMIM genes	3,367	Associated to a phenotype with a 6-digit MIM number assigned, no qualifying '?', '{' or '[' in the phenotype description, and the words "response", "susceptibility", and "somatic" absent from the phenotype description, in the OMIM Synopsis of the Human Gene Map (morbidmap.txt) as of June 11, 2019.
All (tissue expression)	7,931	Expressed at >1 TPM in all 53 tissues in GTEx v7
Some (tissue expression)	9,698	Expressed at >1 TPM in >0 and <53 tissues in GTEx v7
None (tissue expression)	1,076	Expressed at >1 TPM in 0 tissues in GTEx v7
Mouse heterozygous lethal knockout	401	Human orthologs mapping to at least one of the 404 genes with a lethal heterozygous phenotype reported in at least one knockout mouse line per MouseMine[23].

Extended Data Table 2 | Estimation of genetic prevalence for gain-of-function genetic neurodegenerative diseases. *Data sources were identified through 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.*

Gene	Basis	Estimate
HTT	A reported HD incidence of 0.38 cases per 100,000 per year based on meta-analysis[59] multiplied by an average age at death of ~60 for the most common CAG lengths[60]. Finally, a genetic screen of a general population sample[61] 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[62]. 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[61] 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[63], 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[64], 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†.	1 in 3,300
MAPT	Estimation of the genetic prevalence of *MAPT* gain-of-function mutations is difficult because 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, we considered that frontotemporal dementia has a reported incidence of 2.7-4.1 per 100,000 per year[65] 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[66]. Multiplying all these figures results in range of 0.0032% to 0.020%	1 in 5,000 – 31,000
PRNP	We have recently considered the lifetime risk of genetic prion disease in detail[36]. All forms of prion disease (sporadic, genetic, and acquired) appear to be the cause of death of ~1 in 5,000 people based on either death certificate analysis or division of disease incidence by the overall death rate. ~10% of cases are attributable to *PRNP* variants with evidence for Mendelian segregation (although additional cases harbor lower-penetrance variants). 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	1 in 27,000-83,000

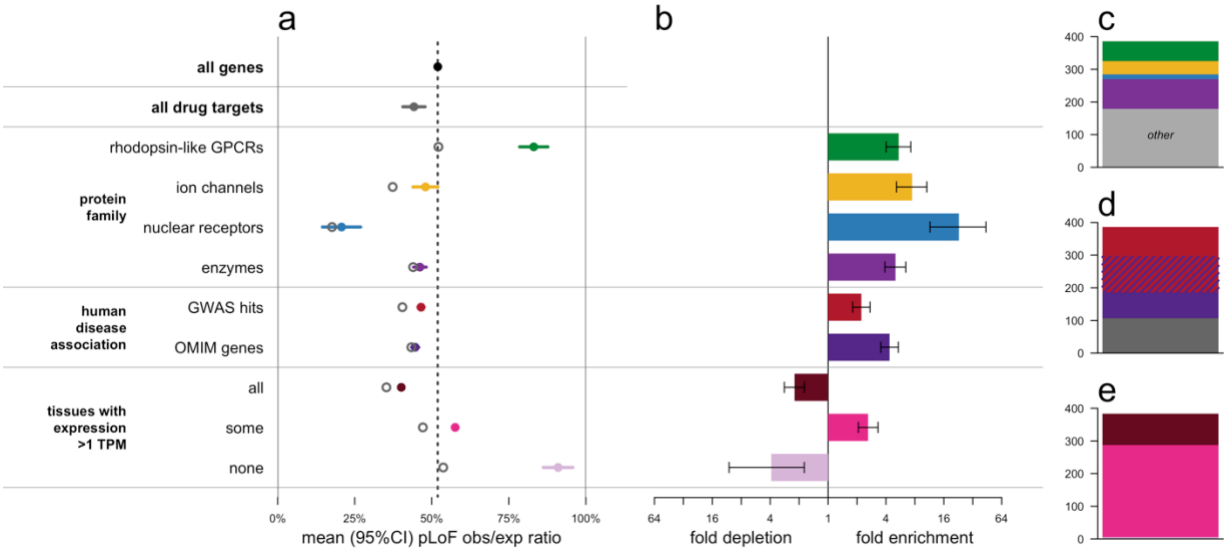
	prevalence of pathogenic *SOD1* mutations.	

†It is important to consider for a moment 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 ~32% penetrance[78]; and other common variants are risk factors with odds ratios of only ~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[64], but G2019S does account for the majority of 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%).

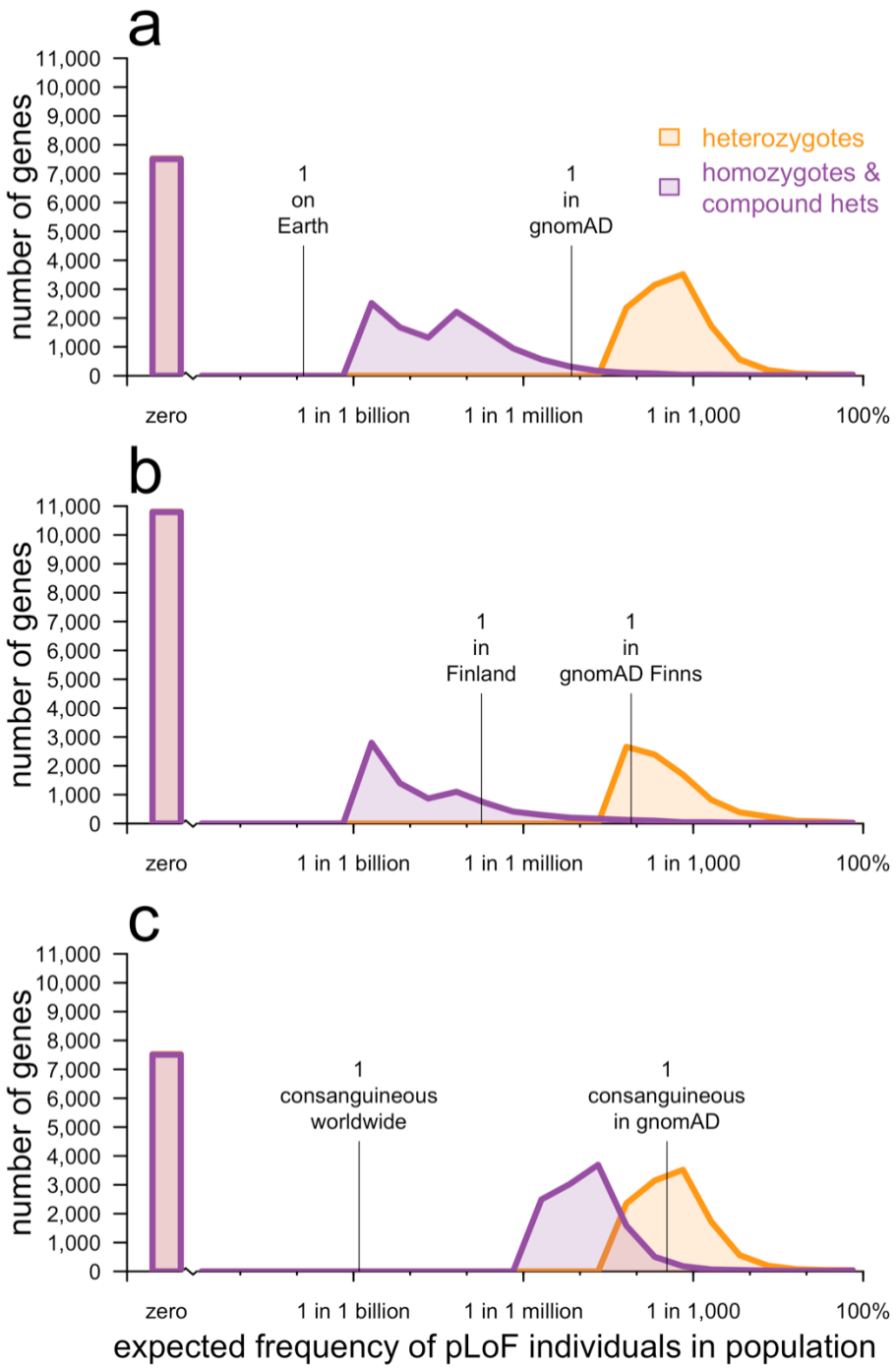
Extended Data Table 3 | Details of PRNP truncating variants. *Allele count for variants from the literature in Fig. 3c is the total number of definite or probable cases with sequencing performed in the studies cited in this table. The L234Pfs7X variant changes PrP's C-terminal GPI signal from SMVLFSSPPVILLISFLIFLIVGX to SMVPSPLHLX. This novel 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 PrP would be predicted to be secreted and thus may be pathogenic, explaining the Alzheimer disease diagnosis in this individual. However, it is also possible that the novel 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.*

variant	allele count	neurological phenotype	comments	reference
G20Gfs84X	1	healthy	As previously reported.	[41]
R37X	2	healthy, unknown	One previously reported, one new.	[41]
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		52
Y163X	7	dementia		53,85
Y169X	2	dementia		53
D178Efs25X	1	dementia		86
Q186X	1	dementia		41
Y226X	1	dementia		87
Q227X	1	dementia		87
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



Extended Data Figure 1 | Drug target gene set confounding. *a) 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 gray 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. c-e) Composition of drug targets when broken down by c) protein family, d) disease association, or e) broadness of tissue expression. 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 TPM, drug targets are still more constrained than other genes (-8.0% obs/exp,* P*=0.00012), but the probable existence of additional unobserved confounders cautions against over-interpretation of this observation (see main text).*



Extended Data Figure 2 | 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 Online Methods, one caveat about Fig. 2 is that the sample size is larger for the plots using all gnomAD exomes (Fig. 2a and 2c) 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-sqrt(q) as described in 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 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.

Group authors

Genome Aggregation Database Production Team: Jessica Alföldi[1,2], Irina M. Armean[3,1,2], Eric Banks[4], Louis Bergelson[4], Kristian Cibulskis[4], Ryan L Collins[1,5,6], Kristen M. Connolly[7], Miguel Covarrubias[4], Beryl Cummings[1,2,8], Mark J. Daly[1,2,9], Stacey Donnelly[1], Yossi Farjoun[4], Steven Ferriera[10], Laurent Francioli[1,2], Stacey Gabriel[10], Laura D. Gauthier[4], Jeff Gentry[4], Namrata Gupta[10,1], Thibault Jeandet[4], Diane Kaplan[4], Konrad J. Karczewski[1,2], Kristen M. Laricchia[1,2], Christopher Llanwarne[4], Eric V. Minikel[1], Ruchi Munshi[4], Benjamin M Neale[1,2], Sam Novod[4], Anne H. O'Donnell-Luria[1,11,12], Nikelle Petrillo[4], Timothy Poterba[9,2,1], David Roazen[4], Valentin Ruano-Rubio[4], Andrea Saltzman[1], Kaitlin E. Samocha[13], Molly Schleicher[1], Cotton Seed[9,2], Matthew Solomonson[1,2], Jose Soto[4], Grace Tiao[1,2], Kathleen Tibbetts[4], Charlotte Tolonen[4], Christopher Vittal[9,2], Gordon Wade[4], Arcturus Wang[9,2,1], Qingbo Wang[1,2,6], James S Ware[14,15,1], Nicholas A Watts[1,2], Ben Weisburd[4], Nicola Whiffin[14,15,1]

1. Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA
2. Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts 02114, USA
3. European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom
4. Data Sciences Platform, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA
5. Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA 02114, USA
6. Program in Bioinformatics and Integrative Genomics, Harvard Medical School, Boston, MA 02115, USA
7. Genomics Platform, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA
8. Program in Biological and Biomedical Sciences, Harvard Medical School, Boston, MA, 02115, USA
9. Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA
10. Broad Genomics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA
11. Division of Genetics and Genomics, Boston Children's Hospital, Boston, Massachusetts 02115, USA
12. Department of Pediatrics, Harvard Medical School, Boston, Massachusetts 02115, USA
13. Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, UK
14. National Heart & Lung Institute and MRC London Institute of Medical Sciences, Imperial College London, London UK
15. Cardiovascular Research Centre, Royal Brompton & Harefield Hospitals NHS Trust, London UK

Genome Aggregation Database Consortium: Carlos A Aguilar Salinas[1], Tariq Ahmad[2], Christine M. Albert[3,4], Diego Ardissino[5], Gil Atzmon[6,7], John Barnard[8], Laurent Beaugerie[9], Emelia J. Benjamin[10,11,12], Michael Boehnke[13], Lori L. Bonnycastle[14], Erwin P. Bottinger[15], Donald W Bowden[16,17,18], Matthew J Bown[19,20], John C Chambers[21,22,23], Juliana C. Chan[24], Daniel Chasman[3,25], Judy Cho[15], Mina K. Chung[26], Bruce Cohen[27,25], Adolfo Correa[28], Dana Dabelea[29], Mark J. Daly[30,31,32], Dawood Darbar[33], Ravindranath Duggirala[34], Josée Dupuis[35,36], Patrick T. Ellinor[30,37], Roberto Elosua[38,39,40], Jeanette Erdmann[41,42,43], Tõnu Esko[30,44], Martti Färkkilä[45], Jose Florez[46], Andre Franke[47], Gad Getz[48,49,25], Benjamin Glaser[50], Stephen J. Glatt[51], David Goldstein[52,53], Clicerio Gonzalez[54], Leif Groop[55,56], Christopher Haiman[57], Craig Hanis[58], Matthew Harms[59,60], Mikko Hiltunen[61], Matti M. Holi[62], Christina M. Hultman[63,64], Mikko Kallela[65], Jaakko Kaprio[56,66], Sekar Kathiresan[67,68,25], Bong-Jo Kim[69], Young Jin Kim[69], George Kirov[70], Jaspal Kooner[23,22,71], Seppo Koskinen[72], Harlan M. Krumholz[73], Subra Kugathasan[74], Soo Heon Kwak[75], Markku Laakso[76,77], Terho Lehtimäki[78], Ruth J.F. Loos[15,79], Steven A. Lubitz[30,37], Ronald C.W. Ma[24,80,81], Daniel G. MacArthur[31,30], Jaume Marrugat[82,39], Kari M. Mattila[78], Steven McCarroll[32,83], Mark I McCarthy[84,85,86], Dermot McGovern[87], Ruth McPherson[88], James B. Meigs[89,25,90], Olle Melander[91], Andres Metspalu[44], Benjamin M Neale[30,31], Peter M Nilsson[92], Michael C O'Donovan[70], Dost Ongur[27,25], Lorena Orozco[93], Michael J Owen[70], Colin N.A. Palmer[94], Aarno Palotie[56,32,31], Kyong Soo Park[75,95], Carlos Pato[96], Ann E. Pulver[97], Nazneen Rahman[98], Anne M. Remes[99], John D. Rioux[100,101], Samuli Ripatti[56,66,102], Dan M. Roden[103,104],

Danish Saleheen[105,106,107], Veikko Salomaa[108], Nilesh J. Samani[19,20], Jeremiah Scharf[30,32,67], Heribert Schunkert[109,110], Moore B. Shoemaker[111], Pamela Sklar*[112,113,114], Hilkka Soininen[115], Harry Sokol[9], Tim Spector[116], Patrick F. Sullivan[63,117], Jaana Suvisaari[108], E Shyong Tai[118,119,120], Yik Ying Teo[118,121,122], Tuomi Tiinamaija[56,123,124], Ming Tsuang[125,126], Dan Turner[127], Teresa Tusie-Luna[128,129], Erkki Vartiainen[66], James S Ware[130,131,30], Hugh Watkins[132], Rinse K Weersma[133], Maija Wessman[123,56], James G. Wilson[134], Ramnik J. Xavier[135,136]

1. Unidad de Investigacion de Enfermedades Metabolicas. Instituto Nacional de Ciencias Medicas y Nutricion. Mexico City
2. Peninsula College of Medicine and Dentistry, Exeter, UK
3. Division of Preventive Medicine, Brigham and Women's Hospital, Boston, Massachusetts, USA.
4. Division of Cardiovascular Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA.
5. Department of Cardiology, University Hospital, 43100 Parma, Italy
6. Department of Biology, Faculty of Natural Sciences, University of Haifa, Haifa, Israel
7. Departments of Medicine and Genetics, Albert Einstein College of Medicine, Bronx, NY, USA, 10461
8. Department of Quantitative Health Sciences, Lerner Research Institute, Cleveland Clinic, Cleveland, OH 44122, USA
9. Sorbonne Université, APHP, Gastroenterology Department, Saint Antoine Hospital, Paris, France
10. NHLBI and Boston University's Framingham Heart Study, Framingham, Massachusetts, USA.
11. Department of Medicine, Boston University School of Medicine, Boston, Massachusetts, USA.
12. Department of Epidemiology, Boston University School of Public Health, Boston, Massachusetts, USA.
13. Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan 48109
14. National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA
15. The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY
16. Department of Biochemistry, Wake Forest School of Medicine, Winston-Salem, NC, USA
17. Center for Genomics and Personalized Medicine Research, Wake Forest School of Medicine, Winston-Salem, NC, USA
18. Center for Diabetes Research, Wake Forest School of Medicine, Winston-Salem, NC, USA
19. Department of Cardiovascular Sciences, University of Leicester, Leicester, UK
20. NIHR Leicester Biomedical Research Centre, Glenfield Hospital, Leicester, UK
21. Department of Epidemiology and Biostatistics, Imperial College London, London, UK
22. Department of Cardiology, Ealing Hospital NHS Trust, Southall, UK
23. Imperial College Healthcare NHS Trust, Imperial College London, London, UK
24. Department of Medicine and Therapeutics, The Chinese University of Hong Kong, Hong Kong, China.
25. Department of Medicine, Harvard Medical School, Boston, MA
26. Departments of Cardiovascular Medicine, Cellular and Molecular Medicine, Molecular Cardiology, and Quantitative Health Sciences, Cleveland Clinic, Cleveland, Ohio, USA.
27. McLean Hospital, Belmont, MA
28. Department of Medicine, University of Mississippi Medical Center, Jackson, Mississippi, USA
29. Department of Epidemiology, Colorado School of Public Health, Aurora, Colorado, USA.
30. Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA
31. Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts 02114, USA
32. Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA
33. Department of Medicine and Pharmacology, University of Illinois at Chicago
34. Department of Genetics, Texas Biomedical Research Institute, San Antonio, TX, USA
35. Department of Biostatistics, Boston University School of Public Health, Boston, MA 02118, USA
36. National Heart, Lung, and Blood Institute's Framingham Heart Study, Framingham, MA 01702, USA
37. Cardiac Arrhythmia Service and Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA
38. Cardiovascular Epidemiology and Genetics, Hospital del Mar Medical Research Institute (IMIM). Barcelona, Catalonia, Spain
39. CIBER CV, Barcelona, Catalonia, Spain
40. Departament of Medicine, Medical School, University of Vic-Central University of Catalonia. Vic, Catalonia, Spain
41. Institute for Cardiogenetics, University of Lübeck, Lübeck, Germany
42. 1. DZHK (German Research Centre for Cardiovascular Research), partner site Hamburg/Lübeck/Kiel, 23562 Lübeck, Germany

43. University Heart Center Lübeck, 23562 Lübeck, Germany
44. Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia
45. Helsinki University and Helsinki University Hospital, Clinic of Gastroenterology, Helsinki, Finland.
46. Diabetes Unit and Center for Genomic Medicine, Massachusetts General Hospital; Programs in Metabolism and Medical & Population Genetics, Broad Institute; Department of Medicine, Harvard Medical School
47. Institute of Clinical Molecular Biology (IKMB), Christian-Albrechts-University of Kiel, Kiel, Germany
48. Bioinformatics Program, MGH Cancer Center and Department of Pathology
49. Cancer Genome Computational Analysis, Broad Institute.
50. Endocrinology and Metabolism Department, Hadassah-Hebrew University Medical Center, Jerusalem, Israel
51. Department of Psychiatry and Behavioral Sciences; SUNY Upstate Medical University
52. Institute for Genomic Medicine, Columbia University Medical Center, Hammer Health Sciences, 1408, 701 West 168th Street, New York, New York 10032, USA.
53. Department of Genetics & Development, Columbia University Medical Center, Hammer Health Sciences, 1602, 701 West 168th Street, New York, New York 10032, USA.
54. Centro de Investigacion en Salud Poblacional. Instituto Nacional de Salud Publica MEXICO
55. Lund University, Sweden
56. Institute for Molecular Medicine Finland (FIMM), HiLIFE, University of Helsinki, Helsinki, Finland
57. Lund University Diabetes Centre
58. Human Genetics Center, University of Texas Health Science Center at Houston, Houston, TX 77030
59. Department of Neurology, Columbia University
60. Institute of Genomic Medicine, Columbia University
61. Institute of Biomedicine, University of Eastern Finland, Kuopio, Finland
62. Department of Psychiatry, PL 320, Helsinki University Central Hospital, Lapinlahdentie, 00 180 Helsinki, Finland
63. Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden
64. Icahn School of Medicine at Mount Sinai, New York, NY, USA
65. Department of Neurology, Helsinki University Central Hospital, Helsinki, Finland.
66. Department of Public Health, Faculty of Medicine, University of Helsinki, Finland
67. Center for Genomic Medicine, Massachusetts General Hospital, Boston, Massachusetts 02114, USA
68. Cardiovascular Disease Initiative and Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA
69. Center for Genome Science, Korea National Institute of Health, Chungcheongbuk-do, Republic of Korea.
70. MRC Centre for Neuropsychiatric Genetics & Genomics, Cardiff University School of Medicine, Hadyn Ellis Building, Maindy Road, Cardiff CF24 4HQ
71. National Heart and Lung Institute, Cardiovascular Sciences, Hammersmith Campus, Imperial College London, London, UK.
72. Department of Health, THL-National Institute for Health and Welfare, 00271 Helsinki, Finland.
73. Section of Cardiovascular Medicine, Department of Internal Medicine, Yale School of Medicine, New Haven, Connecticut3Center for Outcomes Research and Evaluation, Yale-New Haven Hospital, New Haven, Connecticut.
74. Division of Pediatric Gastroenterology, Emory University School of Medicine, Atlanta, Georgia, USA.
75. Department of Internal Medicine, Seoul National University Hospital, Seoul, Republic of Korea
76. The University of Eastern Finland, Institute of Clinical Medicine, Kuopio, Finland
77. Kuopio University Hospital, Kuopio, Finland
78. Department of Clinical Chemistry, Fimlab Laboratories and Finnish Cardiovascular Research Center-Tampere, Faculty of Medicine and Health Technology, Tampere University, Finland
79. The Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, NY
80. Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Hong Kong, China.
81. Hong Kong Institute of Diabetes and Obesity, The Chinese University of Hong Kong, Hong Kong, China.
82. Cardiovascular Research REGICOR Group, Hospital del Mar Medical Research Institute (IMIM). Barcelona, Catalonia.
83. Department of Genetics, Harvard Medical School, Boston, MA, USA
84. Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Churchill Hospital, Old Road, Headington, Oxford, OX3 7LJ UK
85. Wellcome Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK
86. Oxford NIHR Biomedical Research Centre, Oxford University Hospitals NHS Foundation Trust, John Radcliffe Hospital, Oxford OX3 9DU, UK
87. F Widjaja Foundation Inflammatory Bowel and Immunobiology Research Institute, Cedars-Sinai Medical Center, Los Angeles, CA, USA.
88. Atherogenomics Laboratory, University of Ottawa Heart Institute, Ottawa, Canada
89. Division of General Internal Medicine, Massachusetts General Hospital, Boston, MA, 02114

90. Program in Population and Medical Genetics, Broad Institute, Cambridge, MA
91. Department of Clinical Sciences, University Hospital Malmo Clinical Research Center, Lund University, Malmo, Sweden.
92. Lund University, Dept. Clinical Sciences, Skane University Hospital, Malmo, Sweden
93. Instituto Nacional de Medicina Genómica (INMEGEN), Mexico City, 14610, Mexico
94. Medical Research Institute, Ninewells Hospital and Medical School, University of Dundee, Dundee, UK.
95. Department of Molecular Medicine and Biopharmaceutical Sciences, Graduate School of Convergence Science and Technology, Seoul National University, Seoul, Republic of Korea
96. Department of Psychiatry, Keck School of Medicine at the University of Southern California, Los Angeles, California, USA.
97. Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA
98. Division of Genetics and Epidemiology, Institute of Cancer Research, London SM2 5NG
99. Medical Research Center, Oulu University Hospital, Oulu, Finland and Research Unit of Clinical Neuroscience, Neurology, University of Oulu, Oulu, Finland.
100. Research Center, Montreal Heart Institute, Montreal, Quebec, Canada, H1T 1C8
101. Department of Medicine, Faculty of Medicine, Université de Montréal, Québec, Canada
102. Broad Institute of MIT and Harvard, Cambridge MA, USA
103. Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee, USA.
104. Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, USA.
105. Department of Biostatistics and Epidemiology, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA
106. Department of Medicine, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA
107. Center for Non-Communicable Diseases, Karachi, Pakistan
108. National Institute for Health and Welfare, Helsinki, Finland
109. Deutsches Herzzentrum München, Germany
110. Technische Universität München
111. Division of Cardiovascular Medicine, Nashville VA Medical Center and Vanderbilt University, School of Medicine, Nashville, TN 37232-8802, USA.
112. Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA
113. Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA
114. Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY, USA
115. Institute of Clinical Medicine, neurology, University of Eastern Finad, Kuopio, Finland
116. Department of Twin Research and Genetic Epidemiology, King's College London, London UK
117. Departments of Genetics and Psychiatry, University of North Carolina, Chapel Hill, NC, USA
118. Saw Swee Hock School of Public Health, National University of Singapore, National University Health System, Singapore
119. Department of Medicine, Yong Loo Lin School of Medicine, National University of Singapore, Singapore
120. Duke-NUS Graduate Medical School, Singapore
121. Life Sciences Institute, National University of Singapore, Singapore.
122. Department of Statistics and Applied Probability, National University of Singapore, Singapore.
123. Folkhälsan Institute of Genetics, Folkhälsan Research Center, Helsinki, Finland
124. HUCH Abdominal Center, Helsinki University Hospital, Helsinki, Finland
125. Center for Behavioral Genomics, Department of Psychiatry, University of California, San Diego
126. Institute of Genomic Medicine, University of California, San Diego
127. Juliet Keidan Institute of Pediatric Gastroenterology, Shaare Zedek Medical Center, The Hebrew University of Jerusalem, Israel
128. Instituto de Investigaciones Biomédicas UNAM Mexico City
129. Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán Mexico City
130. National Heart & Lung Institute & MRC London Institute of Medical Sciences, Imperial College London, London UK
131. Cardiovascular Research Centre, Royal Brompton & Harefield Hospitals NHS Trust, London UK
132. Radcliffe Department of Medicine, University of Oxford, Oxford UK
133. Department of Gastroenterology and Hepatology, University of Groningen and University Medical Center Groningen, Groningen, the Netherlands
134. Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, MS 39216, USA
135. Program in Infectious Disease and Microbiome, Broad Institute of MIT and Harvard, Cambridge, MA, USA
136. Center for Computational and Integrative Biology, Massachusetts General Hospital

Supplementary Information

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

Supplementary Table 1 | Details of curated variants in neurodegenerative disease genes. *LRRK2 is not included here as curation is reported in detail in a separate publication[54]. We note that frameshift mutations in SOD1 at codons 126 or 127 have been reported to cause a pathogenic gain-of-function leading to ALS[122,123]. Both of these codons occur in the gene's fifth and final exon; all of the variants curated as leading to loss-of-function here are in exons 1-4.*

gene	variant	allele count	status	LOFTEE flags	manual curation result	comments
HTT	4-003076620-AGC-A	14	loftee	lcr		
HTT	4-003076623-AGCAG-A	14	loftee	lcr		
HTT	4-003076631-CAG-C	1	loftee	lcr		
HTT	4-003076632-AGC-A	11	loftee	lcr		
HTT	4-003076632-AGCAGCAGCAGCAGCAGCAGCAG-A	1	loftee	lcr		
HTT	4-003076635-AGCAGCAGCAGCAGCAGCAG-A	10	loftee	lcr		
HTT	4-003076635-AGCAGCAGCAGCAGCAGCAGCAGCAGCAACAG-A	1	loftee	lcr		
HTT	4-003076638-AGCAGCAGCAGCAGCAGCAG-A	1	loftee	lcr		
HTT	4-003076638-AGC-A	54	loftee	lcr		
HTT	4-003076640-CAG-C	116	loftee	lcr		
HTT	4-003076641-AGC-A	32	loftee	lcr		
HTT	4-003076641-AGCAGCAGCAGCAG-A	55	loftee	lcr		
HTT	4-003076644-AGC-A	31	loftee	lcr		
HTT	4-003076644-AGCAGCAGCAGCAG-A	1	loftee	lcr		
HTT	4-003076644-AGCAGCAGCAG-A	31	loftee	lcr		
HTT	4-003076644-AGCAGCAGCAGCAGCAGCAACAG-A	110	loftee	lcr		
HTT	4-003076646-CAG-C	2	loftee	lcr		
HTT	4-003076647-AGC-A	161	loftee	lcr		
HTT	4-003076647-AGCAGCAGCAGCAGCAACAG-A	3	loftee	lcr		
HTT	4-003076647-AGCAGCAGCAG-A	6	loftee	lcr		

HTT	4-003076649-CAG-C	8	loftee	lcr		
HTT	4-003076650-AGC-A	2673	loftee	lcr		
HTT	4-003076650-AGCAG-A	128	loftee	lcr		
HTT	4-003076650-AGCAGCAG-A	26	loftee	lcr		
HTT	4-003076650-AGCAGCAGCAGCAACAG-A	2	loftee	lcr		
HTT	4-003076653-AGC-A	80	loftee	lcr		
HTT	4-003076653-AG-A	1594	loftee	lcr		
HTT	4-003076653-AGCAG-A	1078	loftee	lcr		
HTT	4-003076654-G-GCCGC	2	loftee	lcr		
HTT	4-003076655-CAGCAGCAACA-C	2	loftee	lcr		
HTT	4-003076655-CAG-C	20	loftee	lcr		
HTT	4-003076656-AG-A	84	loftee	lcr		
HTT	4-003076656-A-ACC	2	loftee	lcr		
HTT	4-003076656-AGCAGCAACAG-A	4	loftee	lcr		
HTT	4-003076658-CAG-C	287	loftee	lcr		
HTT	4-003076658-CAGCAACA-C	2	loftee	lcr		
HTT	4-003076658-CA-C	1	loftee	lcr		
HTT	4-003076659-AG-A	1	loftee	lcr		
HTT	4-003076659-AGCAACAG-A	14	loftee	lcr		
HTT	4-003076659-A-ACC	2	loftee	lcr		
HTT	4-003076661-CAACA-C	8	loftee	lcr		
HTT	4-003076662-AACAG-A	251	curated		not_LoF	CAG repeat artifact
HTT	4-003076663-A-AGCAGCAGCAGCAGCAGCAG	2	loftee	lcr		
HTT	4-003076663-A-AGCAGCAGCAGCAGCAGCAGCAG	1	loftee	lcr		
HTT	4-003076663-A-AGCAGCAGCAG	1	loftee	lcr		
HTT	4-003076663-A-AGCAGCAGCAGCAGCAG	2	loftee	lcr		
HTT	4-003076665-A-ACCGCC	49	loftee	lcr		
HTT	4-003076669-GC-G	2	loftee	lcr		
HTT	4-003076670-C-CAGCAGCAG	1	loftee	lcr		
HTT	4-003076670-C-	1	loftee	lcr		

	CAGCAGCAGCAG					
HTT	4-003076672-ACC-A	79	loftee	lcr		
HTT	4-003076680-CG-C	3	loftee	lcr		
HTT	4-003076682-CCG-C	3	loftee	lcr		
HTT	4-003076703-CTTCCT-C	1	curated		not_LoF	repeat region
HTT	4-003076704-T-TCC	3	curated		likely_not_LoF	repeat region, nearby SNP
HTT	4-003076710-AG-A	1	curated			repeat region
HTT	4-003088708-TTGTC-T	1	true		LoF	true 4bp deletion
HTT	4-003088729-CAT-C	1	true		LoF	true 2bp deletion
HTT	4-003107083-G-A	1	true		LoF	essential splice acceptor lost. possible downstream rescue site is out-of-frame
HTT	4-003117118-C-T	3	true		LoF	true stop codon
HTT	4-003131650-G-A	1	true		LoF	true essential splice acceptor lost. 2 downstream splice rescue sites but both out of frame
HTT	4-003133110-CA-C	1	true		LoF	true 1bp deletion
HTT	4-003133110-CAG-C	7	true		LoF	true 2bp deletion
HTT	4-003136141-GTC-G	1	true		LoF	true 2bp deletion
HTT	4-003136269-T-G	1	curated		uncertain_LoF	raw reads not available. would be a true splice donor loss
HTT	4-003138025-C-T	3	true		likely_LoF	likely stop codon, though there is an outside chance it creates a splice donor that preserves frame
HTT	4-003156065-C-T	2	true		LoF	true stop codon
HTT	4-003158859-G-GT	1	true		LoF	true 1bp insertion
HTT	4-003174671-C-T	1	true		LoF	true stop codon
HTT	4-003174707-C-T	1	true		LoF	true stop codon
HTT	4-003176464-C-T	1	true		LoF	true stop codon
HTT	4-003176787-C-T	1	true		LoF	true stop codon
HTT	4-003176796-C-T	1	true		LoF	true stop codon
HTT	4-003184144-C-T	1	true		LoF	true stop codon
HTT	4-003189579-CAAAT-C	1	true		LoF	true 4bp deletion
HTT	4-003205754-CAA-C	1	curated		uncertain_LoF	raw reads not available
HTT	4-003205876-G-A	1	curated		likely_not_LoF	potential in-frame rescue site 3bp upstream
HTT	4-003209047-A-AT	1	true		LoF	true 1bp insertion
HTT	4-003211578-TC-T	1	curated		likely_not_LoF	1bp deletion could be avoided by using potential splice acceptor at subsequent codon
HTT	4-003211677-G-T	1	true		likely_LoF	MNP - D-1 and +1 site are both mutated to T. appears would still be true splice disruptor though
HTT	4-003215736-C-T	1	true		LoF	true stop codon

HTT	4-003216836-G-A	1	curated		likely_not_LoF	potential in-frame donor rescue 6bp downstream
HTT	4-003221937-CG-C	1	true		LoF	true 1bp deletion
HTT	4-003222036-G-A	1	true		LoF	true essential splice donor loss
HTT	4-003224113-G-T	1	true		LoF	true essential splice acceptor lost
HTT	4-003225261-CA-C	1	true		LoF	true 1bp deletion
HTT	4-003237874-A-T	1	true		LoF	true essential splice acceptor lost
HTT	4-003240172-A-G	1	curated		uncertain_LoF	raw reads not available
HTT	4-003240338-T-C	1	curated		likely_not_LoF	GC splice donor might still function, also alternate in-frame GT donor 9 bp upstream
HTT	4-003241749-C-CT	1	loftee	lc_lof		
HTT	4-003241757-C-T	1	loftee	lc_lof		
MAPT	17-044039722-G-T	1	curated		not_LoF	rescued by alternate start codon M11, with good Kozak context
MAPT	17-044049312-G-T	1	curated		not_LoF	non-constitutive exon
MAPT	17-044049312-G-A	2	curated		not_LoF	non-constitutive exon
MAPT	17-044049445-G-A	1	curated		not_LoF	not a real exon
MAPT	17-044051838-G-A	5	loftee	lc_lof		
MAPT	17-044051839-T-C	1	loftee	lc_lof		
MAPT	17-044055646-TA-T	1	loftee	lof_flag		
MAPT	17-044055647-A-T	39690	loftee	lc_lof,lof_flag		
MAPT	17-044055710-A-AC	2	loftee	lof_flag		
MAPT	17-044055746-G-A	1	loftee	lof_flag		
MAPT	17-044060543-G-C	5	curated		not_LoF	non-constitutive exon
MAPT	17-044060582-C-T	25	loftee	lof_flag		
MAPT	17-044060652-A-AG	1	loftee	lof_flag		
MAPT	17-044060675-C-T	5	loftee	lof_flag		
MAPT	17-044060703-CAG-C	2	loftee	lof_flag		
MAPT	17-044060717-C-CA	1	loftee	lof_flag		
MAPT	17-044060724-CT-C	6	loftee	lof_flag		
MAPT	17-044060788-AG-A	3	loftee	lof_flag		
MAPT	17-044060842-CG-C	2	loftee	lof_flag		
MAPT	17-044060877-A-AGGCCTCCCCAGCCCAAGATGGGC	1	loftee	lof_flag		
MAPT	17-044060877-AGGCCTCCCCAGCCCAAGATGGGC-A	1	loftee	lof_flag		
MAPT	17-044060917-C-CGCCAGAG	1	loftee	lof_flag		
MAPT	17-044061006-T-TCCCA	1	loftee	lof_flag		
MAPT	17-044061053-C-T	1	loftee	lof_flag		

MAPT	17-044061059-GC-G	4	loftee	lof_flag		
MAPT	17-044061065-CT-C	1	loftee	lof_flag		
MAPT	17-044061078-TTCACGTGGAAA-T	1	loftee	lof_flag		
MAPT	17-044061153-CAGGGGCCCCTGGAGAGGGGCCAG-C	2	loftee	lof_flag		
MAPT	17-044061154-AGGGGCCCCTGGAGAGGGGCCAGAGGCCC-A	3	loftee	lof_flag		
MAPT	17-044061182-CGG-C	1	loftee	lof_flag		
MAPT	17-044061223-TC-T	3	loftee	lof_flag		
MAPT	17-044061247-TG-T	1	loftee	lof_flag		
MAPT	17-044067273-G-GA	1	loftee	lof_flag		
MAPT	17-044067384-C-G	3	loftee	lof_flag		
MAPT	17-044067395-TC-T	1	loftee	lof_flag		
MAPT	17-044067403-C-T	26	loftee	lof_flag		
MAPT	17-044067438-C-CA	1	loftee	lof_flag		
MAPT	17-044071327-GCC-G	1	curated		not_LoF	non-constitutive exon
MAPT	17-044071329-C-CGGGTA	1	curated		not_LoF	non-constitutive exon
MAPT	17-044073963-A-ACC	1	curated		uncertain_LoF	raw reads not available
MAPT	17-044096026-AGGACAGAGTCCAGTCGAAG-A	2	curated		not_LoF	actually in-frame. starting at K682 it becomes AAATGGT preserving frame
MAPT	17-044096047-TTGGGTCCCTGGACAATATCACCCACGTCCCTGGCGGAGGAAATAAAAAGGTAAAGGG-T	2	curated		not_LoF	actually in-frame. this is the same exact variant as the previous one
PRNP	20-004679975-C-T	2	true			R37X
PRNP	20-004679987-C-T	1	true			Q41X
PRNP	20-004680069-CT-C	1	curated		not_LoF	false variant call, apparent alignment artifact at octapeptide repeat region
PRNP	20-004680071-CATGGTGGTGGCTGGGGGCAGCCCCATGGTGGTGGCTGGGACAGCCT-C	1	curated		not_LoF	false variant call, apparent alignment artifact at octapeptide repeat region
PRNP	20-004680089-C-T	1	true			Q75X
PRNP	20-004680108-G-A	1	true			W81X
PRNP	20-004680162-G-A	1	true			W99X
PRNP	20-004680257-G-T	1	true			G131X
PRNP	20-004680566-CT-C	1	curated		not_LoF	L234Pfs7X; possible pathogenic gain-of-function in dementia case. see Table S1 for details

SNCA	4-090743391-C-TRUE	3	loftee	lc_lof		
SOD1	21-033032095-GC-G	1	true		LoF	true early frameshift, no rescue
SOD1	21-033036098-TAAAGG-T	1	true		likely_LoF	true 5bp frameshift deletion, splice site may be rescued by downstream AG but resulting frame is shifted.
SOD1	21-033036178-GA-G	4	true		LoF	
SOD1	21-033038788-AATCCTCT-A	2	true		LoF	
SOD1	21-033038833-T-C	1	true		LoF	
SOD1	21-033039619-CG-C	2	true		LoF	
SOD1	21-033039689-G-T	2	curated		not_LoF	alternative GT donor 3 bases upstream, in-frame
SOD1	21-033039689-G-GT	2	curated		not_LoF	splice donor D +1 site G->GT insertion creates its own new splice donor