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], Jessica Alföldi[1,2], Richard C Trembath[7,8], David A van Heel[8], Mark J Daly[1,2], Genome Aggregation Database Production Team*, Genome Aggregation Database Consortium*, Stuart L Schreiber[1,9], 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. Faculty of Life Sciences and Medicine, King's College London, London, WC2R 2LS, UK
8. Blizard Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, E1 2AT, UK
9. 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

Abstract

Human genetics has informed the clinical development of new drugs, and is beginning to influence the selection of new drug targets. Large-scale DNA sequencing studies have created a catalogue of naturally occurring genetic variants predicted to cause loss of function in human genes, which in principle should provide powerful *in vivo* models of human genetic “knockouts” to complement model organism knockout studies and inform drug development. Here, we consider the use of predicted loss-of-function (pLoF) variation catalogued in the Genome Aggregation Database (gnomAD) for the evaluation of genes as potential drug targets. Many drug targets, including the targets of highly successful inhibitors such as aspirin and statins, are under natural selection at least as extreme as known haploinsufficient genes, with pLoF variants almost completely depleted from the population. Thus, metrics of gene essentiality should not be used to eliminate genes from consideration as potential targets. The identification of individual humans harboring “knockouts” (biallelic gene inactivation), followed by individual recall and deep phenotyping, is highly valuable to study gene function. In most genes, pLoF alleles are sufficiently rare that ascertainment will be largely limited to heterozygous individuals in outbred populations. Sampling of diverse bottlenecked populations and consanguineous individuals will aid in identification of total “knockouts”. Careful filtering and curation of pLoF variants in a gene of interest is necessary in order to identify true LoF individuals for follow-up, and the positional distribution or frequency of true LoF variants may reveal important disease biology. Our analysis suggests that the value of pLoF variant data for drug discovery lies in deep curation informed by the nature of the drug and its indication, as well as the biology of the gene, followed by recall-by-genotype studies in targeted populations.

Main Text

Human genetics in drug discovery

Human genetics has inspired clinical development pathways for new drugs and shows promise in guiding the selection of new targets for drug discovery[1,2]. The majority of drug candidates that enter clinical trials eventually fail for lack of efficacy[3], 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 that target genes with human genetic evidence for causality in disease are more likely to become approved drugs[4,5].

An oft-cited example is the development of monoclonal antibodies to PCSK9. PCSK9 binds and causes degradation of the low-density lipoprotein (LDL) receptor, thus raising serum LDL and cardiovascular disease (CVD) risk[6]. Naturally occurring genetic variation in the *PCSK9* gene provided a full allelic series that correctly[7,8] predicted that pharmacological inhibition of PCSK9 would lower LDL and be protective against CVD. Gain-of-function variants in *PCSK9* raise LDL and CVD risk[9], whereas variants that reduce functionality lower LDL and CVD risk[10], and variants that result in a total loss of function lower LDL and CVD risk more strongly[11,12]. A human lacking any PCSK9 due to compound heterozygous inactivating mutations has very low LDL and no discernible adverse phenotype[13].

This story illustrates the potential for human genetics to inform on the phenotypic impact — both efficacy and tolerability — of a target's modulation and inactivation, thus providing dose-response and safety information even before any drug candidate has been identified[1]. This provides a powerful motivation to study human genetics when evaluating a potential drug target. At the same time, however, we will show in this article that the characteristics of *PCSK9* cannot be taken as a one-size-fits-all standard for what criteria a gene must meet in order to be a promising drug target. In contrast to *PCSK9*, many highly successful drugs target genes where pLoF variants are depleted by intense natural selection and are extremely rare in the general population. For many of these genes, pLoF variants are too rare for ascertainment of multiple double-null human “knockouts” to be a realistic goal. Nevertheless, the study of pLoF variants and the individuals harboring them, even if limited to heterozygotes, can be deeply informative for drug discovery, but only in the context of deep curation undertaken with awareness of gene and disease biology and of the potential drug and its indication.

Rationale and caveats for studying loss-of-function variants

Variants annotated as nonsense, frameshift, or essential splice site-disrupting are categorized as protein-truncating variants. Provided that there is rigorous filtering of false positives[14], such variants are generally expected to reduce gene function, and are referred to here as predicted loss-of-function (pLoF) variants. In the simplest case, a germline heterozygous loss-of-function allele may correspond to a 50% reduction in gene dosage compared to a wild-type individual, and germline double null (homozygous or compound heterozygous) genotypes may correspond to 0% of normal gene dosage, in all tissues, throughout life — though of course the reality may be more complex. While full dosage compensation appears to be rare, at least at the RNA level[15], a variety of mechanisms may cause heterozygous or even homozygous LoF to be phenotypically muted: for instance, factors other than gene dosage may be rate-limiting for the protein’s function[16], or paralogs may compensate[14]. In some cases, however, pLoF variants can phenocopy long-term pharmacological inhibition, and may be useful for predicting the effects of drugs that negatively impact their target's function, such as inhibitors, antagonists, and suppressors. Such drugs comprise a significant fraction of approved medicines (see below). While the effects of genetic inactivation of potential drug targets have been studied for decades using knockout mice[17], public databases of genetic variation such as the Genome Aggregation Database (gnomAD)[18], containing a total of 141,456 human genomes and exomes, now provide an opportunity to study the effects of gene knockout in the organism of most direct interest: humans.

As with any biological data, information from pLoF variants must be interpreted within a broader therapeutic context. Many drugs are not inhibitors, but rather confer neomorphic or hypermorphic gains of function on their targets[19], and are thus not well-modeled by pLoF variants at all. Even for drugs that antagonize their target's function, it is important to recognize several ways in which genetic knockout and pharmacological inhibition may have divergent effects. For example: a chemical probe may inhibit only one of a protein's two or more functional domains[20], may inhibit proteins encoded by two or more paralogous genes[21], or it may inhibit a target only when a particular complex is formed[22] or, alternatively, when a particular protein-protein interaction is absent[23]. Gene knockouts normally affect every tissue in which a gene is expressed, although in some cases variants may occur on tissue-specific isoforms[24–27], and meanwhile many drugs have tissue-restricted distribution. Genetic knockout is also lifelong, including embryonic phases, whereas pharmacological inhibition is generally temporary and age-restricted; this is important because dozens of approved drugs are known or suspected to cause fetal harm but are tolerated in adults[28]. Finally, as noted above, genetic knockout has a specific "dose", whereas the dosing of pharmacological inhibition can be titrated as needed.

While these caveats are important, the PCSK9 example illustrates that pLoF variants can nonetheless be predictive of the phenotypic effects of drugging a target, and other examples of protective LoF variants modeling therapeutic intervention have subsequently arisen[29–31]. In addition, some drug adverse events may have been predictable in light of human genetic data[32] — for instance, inhibition of DGAT1 resulted in gastrointestinal side effects which may phenocopy biallelic *DGAT1* loss-of-function mutations[33,34]. Currently, however, a systematic framework for applying human genetic data to the selection of drug targets and to the prediction of drug safety is lacking. In this article we lay the groundwork for such a framework, by analyzing the frequency, distribution, and signals of natural selection against pLoF variants in gnomAD, particularly in the targets of approved drugs.

Measuring natural selection in human genes by pLoF constraint

One natural question to ask of a gene is whether disruptive variants that arise in it are severely deleterious — that is, result in a severe disease state that would typically result in carriers having a high risk of being removed from the population by natural selection. In some cases such information is available directly from the observation of severe disease patients where pLoF mutations in that gene have been shown to be causal; however, for a substantial majority of human genes, no severe pLoF phenotype has yet been determined[35]. The lack of a known pLoF phenotype may arise for multiple reasons: (1) disruption of the gene may cause no discernible phenotype at all; (2) disruption may cause a phenotype that is evolutionarily deleterious but clinically mild, has effects only on reproductive fitness rather than individual health, or manifests only upon a certain environmental exposure; (3) the corresponding disease families may not yet have been sequenced or adequately analyzed, at least in sufficient numbers to convincingly demonstrate causation; or (4) pLoF variants may cause a phenotype so severe that human carriers are never observed (e.g. early embryonic lethality). Genes that fall in the latter three categories can be detected even if patients with the corresponding pLoF mutations have not yet been observed, by identifying a depletion of unique pLoF variants in the general population – a state known as *constraint*[36].

Identifying constraint requires comparing the number of pLoF variants observed in a gene in a large population with the number expected in the absence of natural selection. Determining the number of expected pLoF variants in the population relies on a mutation rate model, many of which are based on the rate at which mutations spontaneously arise. The model used here determines the rate of mutation by incorporating the exact nucleotide change (e.g. C to T) and the immediate sequence context, among other factors[18]. Thus, in any given reference population, such as the 125,748 human exomes in gnomAD[18], the expected number of unique genetic variants seen in at least one individual in a gene of interest, absent natural selection, is predicted based on mutation rates[36–38]. This expected number of variants can then be compared to the actual observed number of variants in the database in order to quantify the strength of purifying natural selection acting on that gene, for variants of each functional class — synonymous, missense, and pLoF[36]. Because true pLoF variants are very rare, annotation errors can account for a large fraction of apparent pLoF variants[14], and pLoF constraint is best assessed using rigorous filtering for known error modes and with transcript expression-aware annotation[18,24].

Constraint differs from evolutionary conservation in that constraint (1) informs on selection in humans, not other species; (2) primarily reflects selection against variants in a heterozygous state; and (3) can more finely discriminate strong versus weak selection signals[39,40]. In general, the degree of constraint observed across the genome varies dramatically between synonymous variants, which appear to be under almost no natural selection, missense variants, which show some weak selection, and pLoF variants, which show a strong signal of depletion genome-wide but with marked variation between genes (Figure 1). Various metrics have been developed to quantify constraint[39]; here, we focus on the ratio of observed to expected pLoF variants (obs/exp).

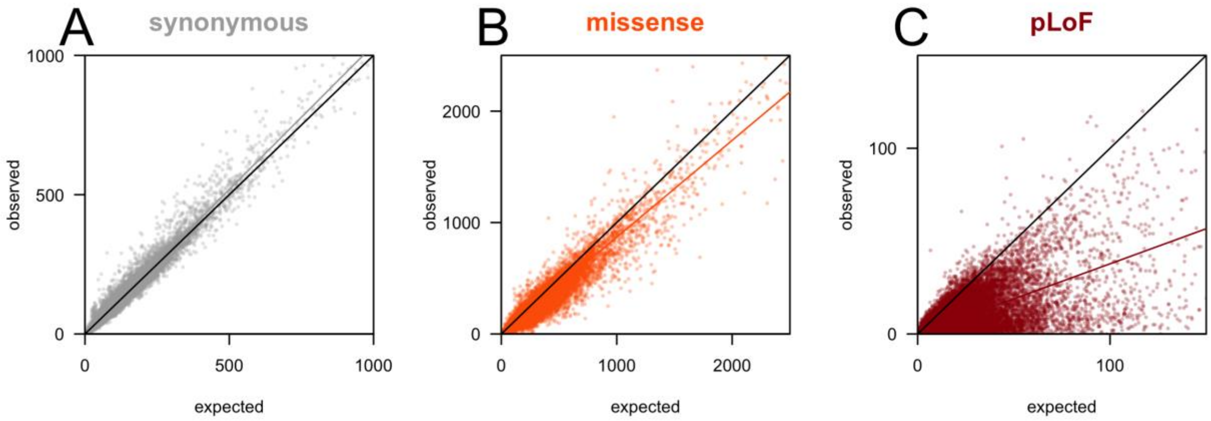

Figure 1. Gene constraint by functional class in gnomAD. *Each dot represents one gene, and its position on the plot represents the expected (x axis) and observed (y axis) number of variants in 125,748 human exomes in gnomAD. Black diagonals represent the expected relationship in the absence of natural selection; colored lines represent the actual best fit relationship. For synonymous variants, where we expect minimal natural selection, the correlation is excellent, with almost all dots lining up right on the diagonal. For missense variants, increased density below the diagonal indicates that some genes are intolerant of*

missense variation. For pLoF variants, most of the density lies below the diagonal, because most genes are at least somewhat intolerant of LoF variation, and some genes extremely so.

Comparison of pLoF constraint in drug targets versus other gene sets

As explained above, constraint allows us to quantify the degree of natural selection against loss-of-function variants in each gene in the human genome. One might expect that drug targets should be less constrained than other genes, since targeting genes that do not tolerate inactivation might result in more adverse events. Alternatively, however, one might expect that drug targets should be more constrained than other genes, since constraint partly reflects a gene's dosage sensitivity, and effective drugs should target genes where a change in gene dosage affects phenotype. We used the obs/exp constraint metric described above to assess the degree of natural selection against loss-of-function variants in the targets of approved drugs (extracted from DrugBank[41], *N*=383). The overall distribution of pLoF obs/exp values for drug targets was similar to that for all genes (Figure 2A). Drug targets include genes under no apparent natural selection against loss-of-function (obs/exp 100%) as well as genes under intense purifying selection (obs/exp 0%).

We compared the mean obs/exp value for drug targets to that of other gene lists (Figure 2B). As previously reported[18,42], the ranking of various gene lists aligns with expectation. Olfactory receptors, which are often dispensable in humans[43], have nearly 100% of their expected pLoF variation, and genes that tolerate homozygous inactivation in humans also have a higher proportion of their expected pLoF variants than the average gene. Recessive disease genes are close to the genome-wide average, possessing 59% of the expected number of pLoF variants, likely reflecting weak selection against heterozygous carriers of inactivating mutations in these genes. Dominant disease genes are more depleted for pLoF, and genes known to be essential in cell culture or associated with diseases of haploinsufficiency are even more severely depleted. Targets of approved drugs are on average more depleted for pLoF variation than the average gene (P = 0.0003), with only 44% of the expected amount of pLoF variation, versus 52% for all genes.

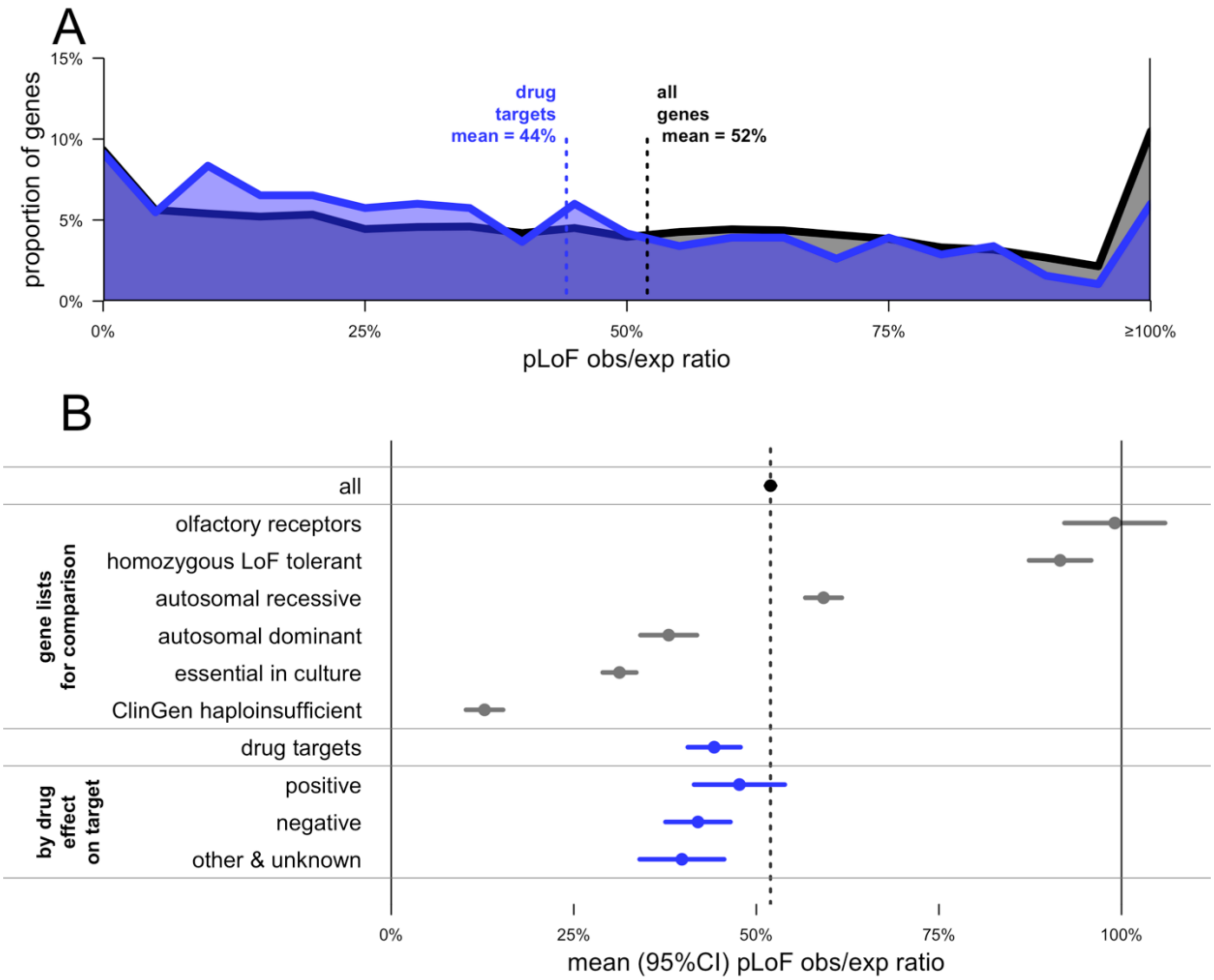

Figure 2. pLoF constraint for drug targets and other gene sets. *A) Histogram of pLoF obs/exp for all genes (black) versus drug targets (blue). B) Forest plot of means (dots) and 95% confidence intervals indicating our certainty about the mean (line segments), for pLoF obs/exp ratio in the indicated gene sets. Data sources for gene sets are listed in Methods.*

We then stratified by the drug's effect on its target: negative (inhibitors, antagonists, suppressors, etc., *N*=240), positive (activators, agonists, inducers, etc., *N*=142), and other/unknown (*N*=94). Although one might expect that targets of negative drugs would need to be more tolerant of pLoF variation, pLoF constraint did not differ significantly between these sets, and if anything, targets of negative drugs were more constrained than those of positive drugs (mean obs/exp 42% vs. 48%, *P*=0.31, Kolmogorov-Smirnov test, Figure 2B). We note that many drug targets are included in more than one of these three sets, and 50 genes are the targets of both positive and negative drugs.

Overall 19% of drug targets (*N*=73), including 53 targets of inhibitors or other negative drugs, have a pLoF obs/exp value less than the average (12.8%) for genes known to cause severe diseases of haploinsufficiency[44] (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). A few of the most heavily constrained drug targets are targets of cytotoxic chemotherapy agents such as topoisomerase inhibitors or cytoskeleton disruptors, a set of drugs intuitively expected to target essential genes. However, several genes with apparently complete or near-complete selection against pLoF variants are targets of highly successful, chronically used inhibitors including statins and aspirin.

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. Examples illustrating the variable degree of selection against pLoF variation in drug targets.

These examples demonstrate that even strong pLoF constraint does not preclude a gene from being a viable drug target. This mirrors the lesson from animal models that a lethal mouse knockout phenotype, such as that reported for *Hmgcr* or *Ptgs2*, does not rule out successful drug targeting[45–47]. The fact that pharmacological inhibition is apparently well-tolerated even in some genes where loss-of-function appears to be evolutionarily deleterious might reflect any of the issues raised above, including differences in effective “dosage”, tissue distribution, or the importance of the gene in embryonic versus adult life stages.

Potential confounding variables in the composition of drug targets

As noted above, drug targets are on average more depleted for pLoF variation than other genes, possessing on average just 44% as much pLoF variation as expected, compared to 52% for all genes (Figure 2), and the effect is similar or stronger when the analysis is limited to drugs with a negative effect on their target's function. From an efficacy perspective, one could argue that this makes sense: constrained genes should be enriched for dosage-sensitive genes, such that a pharmacological agent with less than 100% target engagement can still bring about a change in phenotype. But from a safety perspective, this result is counterintuitive: one would instead have expected that agents targeting more strongly constrained genes are more likely to cause adverse events and so less likely to become approved drugs. Before drawing any conclusions about whether pLoF constraint is predictive of drug success, we sought to identify potential confounding variables that could impact this analysis.

Drug targets are dominated by a few families or classes of proteins, including rhodopsin-like G-protein coupled receptors (GPCRs), nuclear receptors, voltage- and ligand-gated ion channels, and enzymes[48,49]. We asked whether controlling for these classes might affect the results shown in Figure 2. These four classes of genes are collectively enriched by 9.5-fold (95%CI: 7.7-11.7, $P < 1 \times 10^{-50}$, Fisher exact test) among approved drug targets and, in total, account for 54%

(207/386) of targets in our dataset. Each class has a mean pLoF obs/exp value significantly different from the set of all genes, with rhodopsin-like GPCRs being less constrained and the other three classes being more constrained (Figure 3). After controlling for membership in these four target classes as well as an "other" category, approved drug targets are still more constrained than other genes, with a mean pLoF obs/exp ratio lower by 10.0% (P = 6 × 10^{-6}, linear regression), mirroring the result in Figure 2.

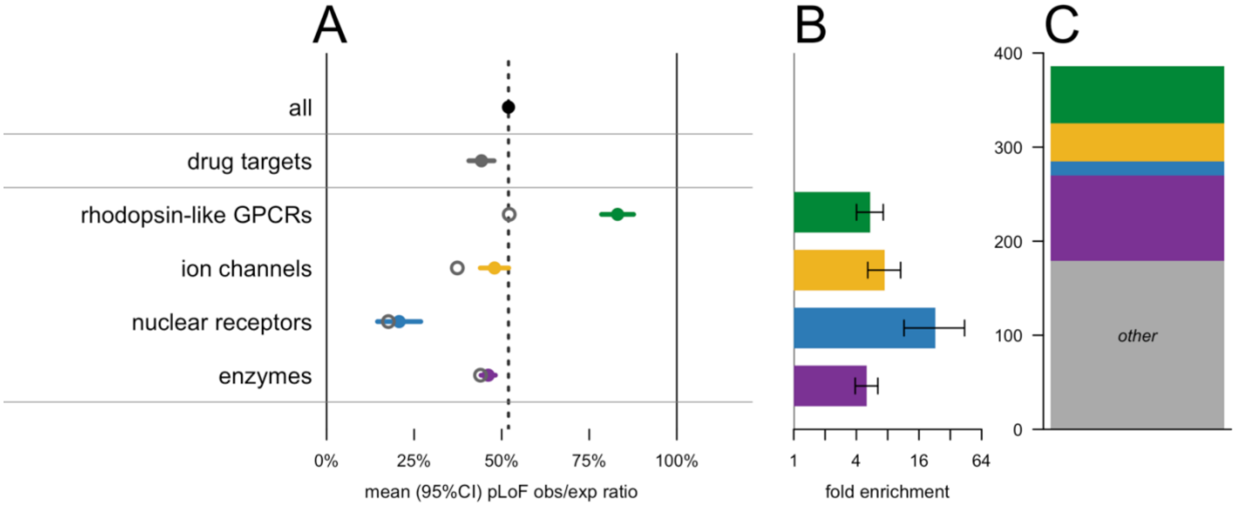

Figure 3. Drug target gene set confounding by target class. *A) Four major classes of proteins that are considered canonically "druggable" have LoF obs/exp ratios significantly different from the set of all genes. Within each class, the genes that are drug targets have a lower mean obs/exp ratio (hollow circles) than the class overall. B) Each of these classes is enriched several-fold among the set of drug targets. C) These classes cumulatively account for over half of drug targets.*

A second potential confounder is that many genes were chosen as drug targets because of their presumed relevance to human disease — targets with human genetic validation are reported to be four-fold enriched among approved drugs versus clinical candidates[4]. Genes adjacent to genome-wide association study (GWAS) hits — a proxy for involvement in human disease — are more constrained than the average gene[42] and are 2.2-fold enriched among drug targets (P = 2 × 10^{-14}, Fisher exact test), collectively accounting for 52% (200/386) of drug targets. However, even after controlling for GWAS hit adjacency, drug targets were still more constrained than other genes with a pLoF obs/exp ratio on average 6% lower (P = 0.003, linear regression).

A third confounder is the number of adult human tissues in which each gene is expressed (thresholding at a median of 1 transcript per million in GTEx v7)[50]. Broader expression across tissues is associated with more severe constraint, meaning inversely correlated with obs/exp (Spearman's correlation r = -0.31, P < 1 × 10^{-50}), and drug targets are on average expressed in fewer tissues than all genes (mean 32/53 vs. 37/53 tissues, P = 1 × 10^{-12}, Kolmogorov-Smirnov test). After controlling for this effect, however, drug targets are still more constrained than the average gene, with pLoF obs/exp 11% lower (P = 2 × 10^{-8}, linear regression).

All three observed variables considered above — protein family, disease association, and tissue expression — are confounded with a gene’s status as drug target. This suggests that many unobserved variables are likely to differ between drug target and non-drug target genes as well. Thus, although drug targets are more constrained than the average gene even after controlling for the variables considered here, it would not necessarily be appropriate to conclude that stronger pLoF constraint is associated with increased likelihood of drug target success. Instead, given the wide spectrum of constraint values observed in drug targets (Figure 2A) and the diverse examples form that spectrum (Table 1), the salient conclusion is that genes from the strongly constrained to the not at all constrained can make viable drug targets.

This analysis is limited in crucial ways by available annotations and gene lists. For instance, we only compared targets of successful drug candidates (those that reach approval) to all genes, whereas to gain insight into safety signals it might be more instructive to compare to the targets of drug candidates that failed early in development due to on-target toxicity; however, to our knowledge, no sufficiently large dataset of such targets currently exists. It is also possible that different trends would emerge if analysis could be limited to drugs taken chronically and systemically (as opposed to transiently and/or locally) and/or stratified by the severity of the indicated condition, as a proxy for the severity of side effects that can be tolerated. Such annotations exist but would require extensive manual curation, a direction for future research. Finally, as noted above, expression patterns during embryonic development may explain some differences between the phenotypic effects of genetic disruption and pharmacological inhibition, but data on human embryonic gene expression are lacking.

Prospects for ascertainment of heterozygous or homozygous “knockout” humans for target validation

The analyses above suggest that a simple statistical approach based solely on quantifying a gene’s constraint will not be sufficient to nominate or exclude good drug targets. Where humans with loss-of-function variants in a potential target can be ascertained and studied, however, their phenotypes are expected to be extremely valuable for predicting the phenotypic effects — both desired and undesired — of a drug against that target. The PCSK9 example illustrates the potential value of such “genotype-first” ascertainment, and has inspired many efforts to do the same for other potential targets of interest[51–54]. To date, however, it has generally been unclear, for any particular gene of interest, how best to go about finding null individuals. Likewise, in genes for which double null humans have not yet been identified, it is often unclear whether this is due to chance, or due to lethality of this genotype.

To explore these questions, we computed the cumulative allele frequency[18] (CAF, or p) of pLoF variants in each gene in gnomAD in order to assess how often heterozygous or homozygous null individuals might be identified for any given gene of interest. We first considered a random mating model, under which the expected frequency of pLoF heterozygotes is 2p(1-p) and the expected frequency of double null or total “knockout” individuals is p^2. Whereas gnomAD is now large enough to include at least one pLoF heterozygote for the majority of genes, ascertainment of total “knockout” individuals in outbred populations will require multiple orders of magnitude larger sample sizes for most genes (Figure 4A). For instance, consider a sample size of 14 million individuals from outbred populations, 100 times larger than gnomAD today. In this sample size, 75% of genes (*N*=14,340) would still be expected to have <1 double null individual, and 91% of genes (*N*=17,546) — including 92% of existing approved drug targets (*N*=357) — would have sufficiently few expected “knockouts” that observing zero of them would not represent a statistically significant departure from expectation. Indeed, for 38% of genes (*N*=7,546), even if all humans on Earth were sequenced, observing zero “knockouts” would still not be a statistically significant anomaly. Thus, for the vast majority of genes for the foreseeable future, examining outbred populations alone will not provide statistical evidence that a double null genotype is not tolerated in humans.

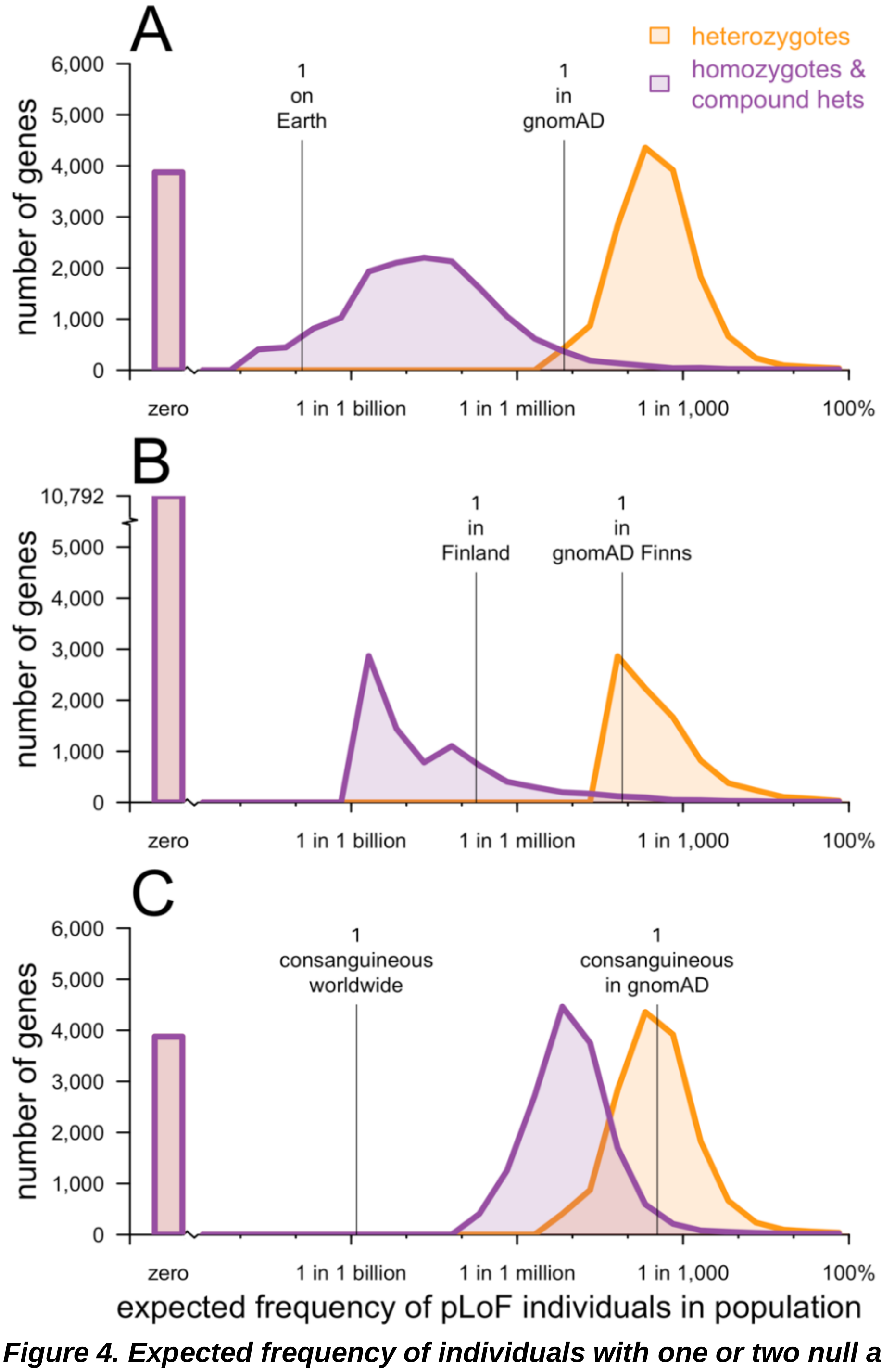

***Figure 4. Expected frequency of individuals with one or two null alleles for every protein-coding gene.** Each panel shows a histogram where the y axis is number of genes and the x axis shows the theoretically expected population frequency of heterozygotes (orange), versus homozygotes and compound heterozygotes (purple). Zero indicates the number of genes where no pLoF variants have been observed. **A)** Outbred populations, under random mating. **B)** Finnish individuals, an example of a bottlenecked population. **C)** Consanguineous individuals with 5.8% of their genome autozygous. See Methods for details.*

Some human populations, however, have demographic properties that render them substantially more likely to produce “knockout” individuals. One such category is bottlenecked populations, such as Finnish or Ashkenazi Jewish individuals, which are descended relatively recently (<100 generations) from a small number of historical founders that subsequently expanded rapidly in size to create large modern populations. This demographic history means that very rare LoF variants present in a founder — including both neutral and relatively deleterious variants — can rise to an unusually high allele frequency in the resulting population[51]. Thus, it is possible for a gene where LoF variants are ultra-rare in an outbred population, either due to chance or due to natural selection, to harbor relatively common LoF variants in a bottlenecked population. Examination of the full distribution of cumulative LoF allele frequencies for different genes (Figure 4B), however, reveals the double-edged sword of pLoF analysis in bottlenecked populations. In any given bottlenecked population, only a handful of genes have common pLoF variants, and meanwhile, rare pLoF variants that did not pass through the bottleneck have been effectively removed from these populations, resulting in a greater proportion of genes with few or no pLoF variants. If one’s gene of interest happens to be present at high frequency in such a population, then a large number of pLoF individuals can be ascertained, enabling association studies that would have been difficult or underpowered in a larger, outbred population[51,55]. But if one begins with a pre-determined gene or list of genes to study, any specific bottlenecked population is less likely to reveal interesting pLoF variants than are outbred populations. As such, any effort to use such populations for genome-wide target validation studies would be well-advised to draw samples from as diverse a set of bottlenecked populations as possible to maximize the probability of any specific gene being knocked out in at least one group.

The study of consanguineous individuals, by contrast, is much more likely to identify homozygous pLoF genotypes for a pre-determined gene of interest. The East London Genes & Health (ELGH) initiative[56] has recruited ~35,000 British Pakistani and Bangladeshi individuals, about 20% of whom report that their parents are related. On average, the *N*=2,912 individuals who reported that their parents were second cousins or closer had 5.8% (about 1/17th) of their genome in runs of autozygosity, meaning that both chromosomes are identical, inherited from the same recent ancestor. Consider, for example, a gene with pLoF allele frequency 1 in 3,000. This gene would be expected to have homozygous or compound heterozygous pLoF variants in $(1/3,000)^2$ = 1 in 9 million individuals in an outbred population, but 0.058 * 1/3,000 = 1 in 52,000 consanguineous individuals. Unlike in bottlenecked populations, where certain pLoF variants can be very common, the allele frequency of pLoF variants is not shifted in populations with elevated rates of consanguinity; only the homozygote frequency is dramatically shifted to the right (Figure 4C). These properties explain why the study of these populations has been highly fruitful to date[53,57,58] and justify ambitious plans to expand these cohorts in the coming years[56,59]. However, it is worth emphasizing that because the underlying variants are still rare, studying these populations may only identify a handful of individuals with a homozygous pLoF genotype in a specific gene of interest; such data may be adequate to address safety questions and identify stark phenotypic effects[53], but will often be highly underpowered for the study of subtle clinical phenotypes or the direct validation of disease-protective effects.

Ascertainment of double null “knockout” humans remains a desirable goal for establishing the phenotype associated with complete loss of the target gene. However, the data above demonstrate that discovery of substantial numbers of such individuals may be infeasible for many genes of interest in outbred populations. Even in consanguineous cohorts, for most genes, observing homozygous individuals will require orders of magnitude larger sample sizes than are available today (Figure 4C). At present, for most genes, we believe that well-powered studies of the phenotypic impact of human LoF alleles will be limited to heterozygous individuals, which often will provide valuable models of partial gene inhibition, as we describe in an accompanying manuscript exploring individuals heterozygous for *LRRK2* LoF[54].

Regardless of the study design, moving from pLoF genotypes to information about specific clinical outcomes will depend critically on the accuracy of pLoF identification. We thus next turn our attention to the careful curation required to filter for true LoF variants before embarking upon any genotype-based ascertainment effort.

Curation of pLoF variants in six neurodegenerative disease genes

To illustrate both the opportunities and the challenges associated with identifying true LoF individuals for further study, we manually curated the data from gnomAD as well as the scientific literature for six genes associated with gain-of-function (GoF) neurodegenerative diseases, for which inhibitors or suppressors are presently under development[60–68]: *HTT* (Huntington disease), *MAPT* (tauopathies), *PRNP* (prion disease), *SOD1* (amyotrophic lateral sclerosis), and *LRRK2* and *SNCA* (Parkinson disease). The results (Table 2 and Figure 5) illustrate four points about pLoF variant curation.

			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[69–71]
LRRK2	7,581	41%	0.23%	0.09%	1 in 500	1 in 3,300[72,73]
MAPT	2,328	0%*	14%	0%	not observed	1 in 5,000 – 31,000[74,75]
PRNP	759	99%**	0.0035%	0.0021%	1 in 18,000	1 in 50,000[76]
SNCA	420	0%	0.0012%	0%	not observed	1 in 360,000[72,77]
SOD1	462	18%	0.0060%	0.0038%	1 in 26,000	1 in 27,000-83,000[1–3]

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. Curation details and genetic prevalence calculations are included in the Supplement, except for LRRK2 which is described in detail in Whiffin et al[54]. *Constitutive brain-expressed exons only. **PRNP codons 1-144, see Figure 5C for rationale.*

First, other things being equal, genes with longer coding sequences have more opportunity for LoF variants to arise, and so are likely to have a higher cumulative frequency of LoF variants, unless they are heavily constrained. Thus, shorter and/or more constrained genes are more difficult targets for the follow-up of LoF individuals, even though constraint in and of itself does not rule out a gene being a good drug target (Table 1).

Second, many variants annotated as pLoF are in fact false positives, and this is particularly true of pLoF variants with higher allele frequencies, such that the true cumulative allele frequency of LoF is often much lower after manual curation than before. As such, studies of human pLoF variants that do not apply extremely stringent curation to their candidate variants can easily dilute their clinical studies with large numbers of false pLoF carriers or homozygotes, rendering the resulting data challenging or impossible to interpret. In the long term, we anticipate that high-throughput direct functional validation of candidate pLoF variants will become the standard for such studies in humans.

Third, even after careful curation, the cumulative frequency of LoF variants is sometimes sufficiently high to place certain bounds on what heterozygote phenotype might exist. For example, in *HTT*, *LRRK2*, *PRNP*, and *SOD1*, individuals with high-confidence heterozygous LoF variants are equally or more common in the population than people with gain-of-function variants that cause neurodegenerative disease. In each case, the gain-of-function disease has been well-characterized for decades. Thus, it seems unlikely that a comparably severe and penetrant heterozygous loss-of-function syndrome associated with the same gene could have gone unnoticed to the present day. Of course, this does not rule out the possibility that heterozygous loss-of-function could be associated with a less severe or less penetrant phenotype.

Finally, the positional distribution of pLoF variants often appears non-random, and careful curation of variants in such genes can often reveal a reason for the observed distribution, with resulting dramatic changes in the gene's constraint and/or cumulative LoF allele frequency. Three genes in our curation set — *HTT*, *MAPT*, and *PRNP* — are good examples of how different non-random positional distributions of pLoF variants in a gene's coding sequence can correspond to different error modes or disease biology (Figure 5).

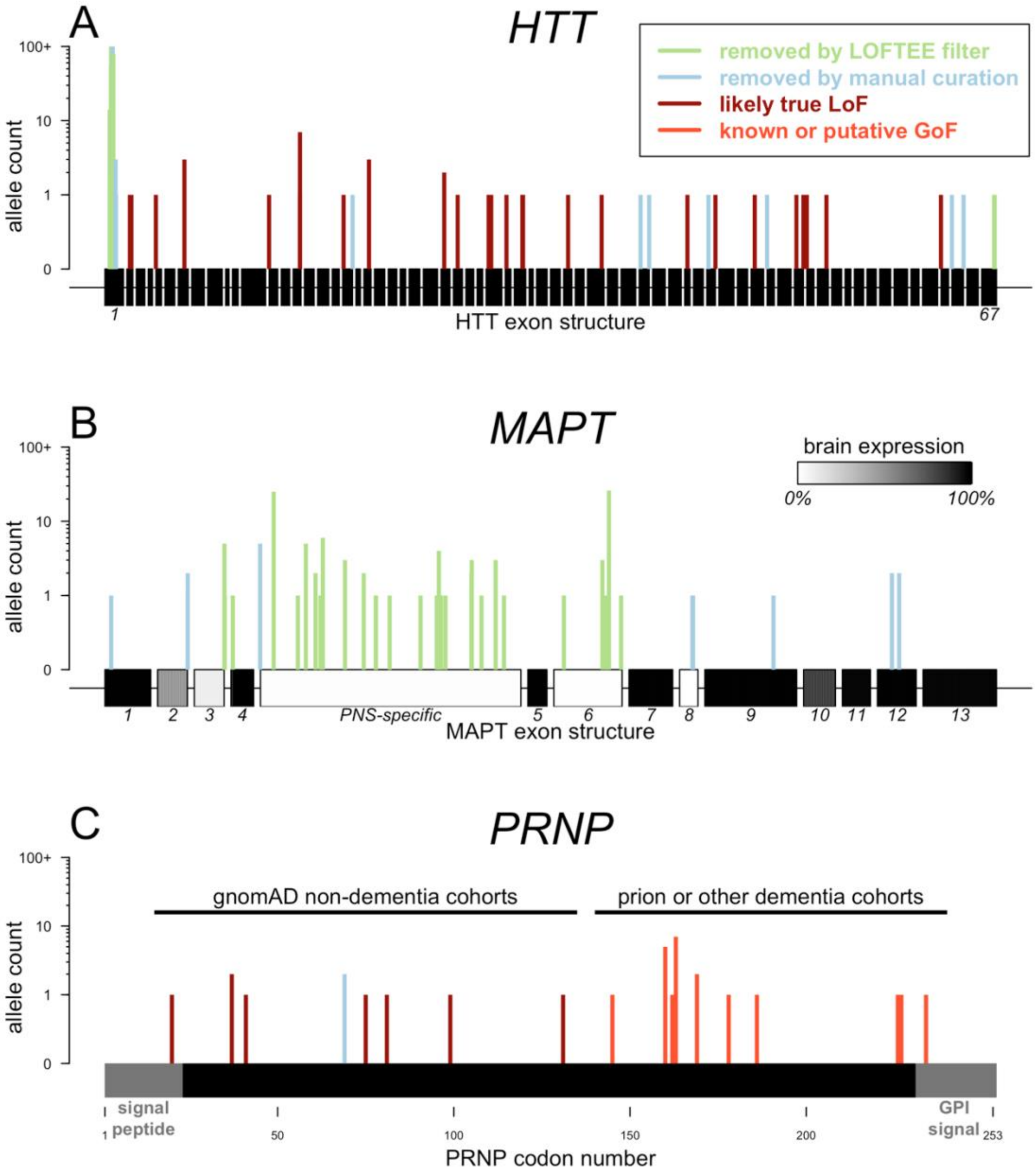

Figure 5. Non-random positional distributions of pLoF variants across a gene's coding sequence can reflect specific error modes or reveal disease biology. A) *HTT,* ***B)*** *MAPT, with exon numbering and annotation from Andreadis*[78] *and brain expression data from GTEx*[50]*, and* ***C)*** *PRNP, a single protein-coding exon with domains removed by post-translational modification in gray, showing previously reported variants*[52] *as well as those newly identified in gnomAD and in the literature*[79,80]*. See text for interpretation and Supplement for detailed curation results.*

HTT, the gene encoding huntingtin, the cause of Huntington disease, appears at first glance to harbor several common LoF variants, with a cumulative allele frequency of 6%. This is surprising in view of this gene's strong constraint in humans (Table 2) and the known embryonic lethal phenotype of homozygous knockout in mice[81]. Inspection (Figure 5A) reveals that all of the common pLoF variants in *HTT* are sequencing read alignment artifacts within the polyglutamine and polyproline tracts of exon 1, some of which are removed by the automated annotation tool LOFTEE[18], and the rest of which can be identified quickly by visual inspection. True LoF variants in *HTT* are in fact rare, consisting mostly of singletons (variants seen only once in gnomAD's database of 141,456 individuals). Nonetheless, a total of 37 apparently real LoF alleles are observed in *HTT*, and these variants are positionally random and include nonsense, splice, and frameshift mutations. This suggests that ~1 in 3,800 people in the general population are heterozygous for genuine LoF of *HTT*, making this genotype about as common as the *HTT* CAG repeat expansion that causes Huntington's disease. While heterozygous *HTT* LoF variants do appear to be under negative selective pressure given the clear depletion of such variants in the population, the prevalence of this genotype makes it unlikely that such variants result in a penetrant, severe, syndromic illness. This conclusion is consistent with the lack of reported phenotype in a human with a heterozygous translocation disrupting *HTT*[82] and the heterozygous parents of children with a neurodevelopmental disorder due to compound heterozygous hypomorphic mutations in *HTT*[83,84]. Heterozygous knockout mice are likewise reported to have no obvious abnormality[81], although reduced body weight has been noted[85]. Functional studies to confirm that the observed variants in *HTT* are true LoF, and recall-by-genotype efforts to identify any phenotype in these individuals remain important future research directions. At present, the balance of evidence suggests that heterozygous *HTT* loss-of-function does not cause a severe, penetrant disease in humans.

MAPT, the gene encoding tau, the cause of tauopathies and an important protein in Alzheimer disease, appears at first glance to harbor a large number of LoF variants, some of which are common, leading to a cumulative LoF allele frequency of 14%. The positional distribution of variants is suspiciously non-random, however, with LoFs concentrated in a few exons. Plotting the variant data against brain RNA expression data[24] reveal the reason for this pattern (Figure 5B): almost all of the pLoF variants in *MAPT*, including all those with appreciable allele frequency, fall in exons that are not expressed in the brain. The few remaining pLoF variants that do fall in brain-expressed exons were all determined to be sequencing or annotation errors upon closer inspection, meaning that no true LoF variants are observed in *MAPT*. Heterozygous *MAPT* deletions in humans have been reported: a partial deletion of exons 6-9 is believed to result in pathogenic gain of function[86], while the 17q21.31 microdeletion syndrome[87] spanning *MAPT* and four other genes is associated with a neurodevelopmental disorder that has since been causally attributed to the loss of *KANSL1*[88]. Homozygous *Mapt* knockout mice are grossly normal[89,90]. Our data would be consistent with *MAPT* loss-of-function having some fitness effect in humans, but our sample size is insufficient to prove that *MAPT* loss-of-function is not tolerated (see Supplement). Even if heterozygous *MAPT* loss-of-function is pathogenic, this does not imply that *MAPT* is not a viable drug target, for the reasons explained above. However, this would mean that ascertaining and studying *MAPT* LoF individuals in order to determine whether reduced gene dosage is protective against tauopathies may prove difficult or impossible.

PRNP, the gene encoding prion protein, the cause of prion disease, is a single-exon gene, so truncating variants do not trigger nonsense-mediated decay and instead result in shortened proteins. *PRNP* appears at first glance to be modestly depleted for LoF variants, particularly in its C terminus. As previously reported[52], comparing gnomAD data to reported pathogenic variants in the literature (Figure 5C) reveals that truncating variants at codon 145 or higher are associated with a pathogenic gain-of-function leading to prion disease, apparently through removal of the protein's GPI anchor. All of the variants seen in non-dementia cohorts in gnomAD occur prior to codon 145 and appear to correspond to true LoF. An individual with a

G131X mutation was found to be neurologically healthy at age 77 with no family history of neurodegeneration (see Supplement), suggesting that stop codons up through at least codon 131 are benign. The sole C-terminal truncating variant observed in gnomAD, a frameshift at codon 234, at the beginning of the GPI signal, turns out to be an individual with dementia diagnosed clinically as Alzheimer disease (see Supplement). This is consistent with the slowly progressive dementia reported for some *PRNP* late truncating mutations[91], although we cannot exclude the possibility that this variant is benign and that the Alzheimer diagnosis is a coincidence. When only codons 1-144 are considered, *PRNP* is not constrained at all (Table 2). Because the gene is short, the cumulative frequency of LoF variants is still low: ~1 in 18,000 individuals are heterozygous for *PRNP* LoF, a frequency that has enabled phenotypic characterization of a small number of individuals (Supplement), although ascertainment of homozygotes will likely only ever be possible in consanguineous individuals.

The above examples illustrate only a few of the types of positional patterns and error modes that may appear upon manual curation. Additional examples have been reported previously[92,93], and a companion paper further illustrates the importance of transcript expression-aware annotation[24]. For anyone considering developing a drug against a target, the types of analyses described above are only a first step. Variants that appear to be true LoF after filtering and curation still occasionally turn out not to disrupt gene function, so RNA and/or protein studies are essential. Once true pLoF variants are identified, recontact efforts can be initiated where consents allow, and even when deep phenotype information is not available, examining the age distribution, study cohorts, and case/control status of pLoF individuals can be highly valuable. For an example of such a deeper analysis of one gene of interest, see our companion paper on pLoF variants in *LRRK2*[54].

Suggestions for assessing pLoF variation in potential drug targets

While there are many caveats, and pLoF variants in a gene will never be a perfect model of pharmacological inhibition of that gene's product, there are now many examples to illustrate that pLoF variants can have enormous predictive value for the phenotypic impact of drugging a target[1,2]. We therefore expect that many more sequencing, functional studies, recontact efforts, and association studies will be undertaken with the intent of characterizing the impact of pLoF variants on genes under consideration as potential drug targets. In view of the above analyses and findings, we suggest guidelines for how such approaches can be undertaken (Box 1).

- **Carefully filter and curate pLoF variants.** False positive pLoF variants abound, and are particularly enriched among common pLoF variants. Filtering using annotation tools such as LOFTEE[18], RNA expression data[24], and deep manual curation are critical before interpreting variants or initiating expensive downstream recontact or phenotyping efforts.
- **Consider the positional distribution of pLoF variants.** A non-random distribution of pLoF variants throughout a gene's coding sequence can reflect sequencing or annotation pitfalls, or can point to disease biology. Interpreting such patterns often requires careful analysis both of error modes and of gene-specific biology including transcript structure and expression.
- **Calculate cumulative allele frequency.** The sum of the frequency of all pLoF variants in a gene will predict how realistic it is to identify a sufficient number of heterozygous and double null individuals for follow-up studies, and can often be informative in itself. Identify any populations with higher pLoF frequencies, as these may be the most fruitful for follow-up studies. If ascertainment of homozygotes is desired, sequencing of populations with higher rates of consanguinity will often be the most realistic route.

- **Where possible, experimentally validate loss of function.** Even after careful filtering and curation, RNA or protein studies will sometimes reveal that a pLoF variant does not in fact disrupt gene function. For high-value target genes, developing high-throughput functional assays and using these to test all candidate pLoF variants will often be worthwhile before embarking on clinical follow-up studies.
- **Do not eliminate genes from consideration based solely on a lack of pLoF individuals.** Some genes, whether because they are short, and thus have few mutations expected *a priori*, or because they are under intense natural selection, have very few pLoF variants. Many successful approved drugs target such genes. Even when pLoF heterozygotes can be observed, double null individuals should not be expected for most genes at present sample sizes. While pLoF variation is valuable, lack thereof should not preclude a target from consideration.

Box 1. Suggested guidelines for studying pLoF variation in a candidate drug target.

Above all, we suggest that the study of pLoF variation should be informed by a full view of the biology of the gene, drug, and indication. Nothing about developing a drug is trivial, and that includes applying lessons from human genetics. But given the scale and expense of drug development, it is worth the effort to carefully read out, through human genetics, the valuable data from experiments that nature has already done.

Methods

Data sources

pLoF analyses used the gnomAD dataset of 141,456 individuals[18]. For data consistency, all genome-wide constraint and CAF analyses (Figures 1-4) 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 and December 2018 as shown in Table 3.

List	*N*	Description
All	19,194	HGNC protein-coding genes[94].
Olfactory receptors	371	As reported by Mainland et al[95].
Homozygous LoF tolerant	330	Genes with at least two different high-confidence pLoF variants 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[96].
Autosomal dominant	307	OMIM disease genes deemed to follow autosomal dominant inheritance according to extensive manual curation by the Przeworski group[96].
Essential in culture	683	Genes deemed essential in cultured cell lines based on CRISPR screens[97].

ClinGen haploinsufficient	294	Genes with sufficient evidence for dosage pathogenicity (level 3) as determined by the ClinGen Dosage Sensitivity Map[44]
Approved drug targets	386	Genes listed as the top-ranked mechanistic target of approved drugs in the DrugBank 5.0 XML release[41]. 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 targets	94	Action not listed in DrugBank, or any action other than those listed above for positive and negative targets.
Rhodopsin-like GPCRs	689	HGNC gene set 140: "G protein-coupled receptors, Class A rhodopsin-like"[94].
Ion channels	326	HGNC gene set 177: "Ion channels"[94].
Nuclear receptors	48	IUPHAR/BPS Guide to Pharmacology "Nuclear receptors" list[98] .
Enzymes	1,178	IUPHAR/BPS Guide to Pharmacology "Enzymes" list[98].
Genes adjacent GWAS hits	6,336	Closest gene to GWAS hits with *P* < 5-e8 in the EBI GWAS catalog (MAPPED_GENE column)[99].

Table 3. 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.*

Calculation of pLoF constraint

The calculation of constraint values for genes has been described in general elsewhere[36,42] and for this dataset specifically by Karczewski et al[18]. Constraint calculations were limited to single-nucleotide variants (which for pLoF means nonsense and essential splice site mutations) found in gnomAD exomes with minor allele frequency < 0.1% and categorized as high-confidence LoF by LOFTEE. 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 Figure 5B.

Calculation of pLoF heterozygote and homozygote/compound heterozygote frequencies

Cumulative pLoF allele frequency (CAF) was calculated as reported[18]. Briefly, 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 (Figure 4A), 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 (Figure 4B), 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 (Figure 4C), 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)[100].

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[42]. We naively group all populations together, even though the distribution of populations sampled in gnomAD does not reflect the world population[18]; 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[42]. It is important to note that the histograms shown in Figure 4 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 (Figures 4A and 4C) is larger than for only Finnish exomes (Figure 4B). For a version of Figure 4 with the global gnomAD population downsampled to the same sample size as the gnomAD Finnish population, see Figure S1.

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[42,52], 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 the Supplement.

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 is supported by the National Institutes of Health (F31 AI22592) and by an anonymous organization. gnomAD data aggregation was supported primarily by the Broad Institute, gnomAD analysis was funded by NIDDK U54 DK105566, and development of LOFTEE by NIGMS R01 GM104371. 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, Daniel Rhodes, 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/).

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

References

1. Plenge RM, Scolnick EM, Altshuler D. Validating therapeutic targets through human genetics. Nat Rev Drug Discov. 2013 Aug;12(8):581–594. PMID: 23868113
2. Kathiresan S. Developing medicines that mimic the natural successes of the human genome: lessons from NPC1L1, HMGCR, PCSK9, APOC3, and CETP. J Am Coll Cardiol. 2015 Apr 21;65(15):1562–1566. PMID: 25881938
3. Hay M, Thomas DW, Craighead JL, Economides C, Rosenthal J. Clinical development success rates for investigational drugs. Nat Biotechnol. 2014 Jan;32(1):40–51. PMID: 24406927
4. Nelson MR, Tipney H, Painter JL, Shen J, Nicoletti P, Shen Y, Floratos A, Sham PC, Li MJ, Wang J, Cardon LR, Whittaker JC, Sanseau P. The support of human genetic evidence for approved drug indications. Nat Genet. 2015 Aug;47(8):856–860. PMID: 26121088
5. King EA, Davis JW, Degner JF. 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. 2019 Jan 1;513945.
6. Horton JD, Cohen JC, Hobbs HH. Molecular biology of PCSK9: its role in LDL metabolism. Trends Biochem Sci. 2007 Feb;32(2):71–77. PMCID: PMC2711871
7. Sabatine MS, Giugliano RP, Wiviott SD, Raal FJ, Blom DJ, Robinson J, Ballantyne CM, Somaratne R, Legg J, Wasserman SM, Scott R, Koren MJ, Stein EA, Open-Label Study of Long-Term Evaluation against LDL Cholesterol (OSLER) Investigators. Efficacy and safety of evolocumab in reducing lipids and cardiovascular events. N Engl J Med. 2015 Apr 16;372(16):1500–1509. PMID: 25773607
8. Robinson JG, Farnier M, Krempf M, Bergeron J, Luc G, Averna M, Stroes ES, Langslet G, Raal FJ, El Shahawy M, Koren MJ, Lepor NE, Lorenzato C, Pordy R, Chaudhari U, Kastelein JJP, ODYSSEY LONG TERM Investigators. Efficacy and safety of alirocumab in reducing lipids and cardiovascular events. N Engl J Med. 2015 Apr 16;372(16):1489–1499. PMID: 25773378
9. Abifadel M, Varret M, Rabès J-P, Allard D, Ouguerram K, Devillers M, Cruaud C, Benjannet S, Wickham L, Erlich D, Derré A, Villéger L, Farnier M, Beucler I, Bruckert E, Chambaz J, Chanu B, Lecerf J-M, Luc G, Moulin P, Weissenbach J, Prat A, Krempf M, Junien C, Seidah NG, Boileau C. Mutations in PCSK9 cause autosomal dominant hypercholesterolemia. Nat Genet. 2003 Jun;34(2):154–156. PMID: 12730697
10. Kotowski IK, Pertsemlidis A, Luke A, Cooper RS, Vega GL, Cohen JC, Hobbs HH. A spectrum of PCSK9 alleles contributes to plasma levels of low-density lipoprotein cholesterol. Am J Hum Genet. 2006 Mar;78(3):410–422. PMCID: PMC1380285
11. Cohen J, Pertsemlidis A, Kotowski IK, Graham R, Garcia CK, Hobbs HH. Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. Nat Genet. 2005 Feb;37(2):161–165. PMID: 15654334
12. Cohen JC, Boerwinkle E, Mosley TH, Hobbs HH. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. N Engl J Med. 2006 Mar 23;354(12):1264–1272. PMID: 16554528
13. Zhao Z, Tuakli-Wosornu Y, Lagace TA, Kinch L, Grishin NV, Horton JD, Cohen JC, Hobbs HH. Molecular characterization of loss-of-function mutations in PCSK9 and identification of a compound heterozygote. Am J Hum Genet. 2006 Sep;79(3):514–523. PMCID: PMC1559532
14. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, Jostins L, Habegger L, Pickrell JK, Montgomery SB, Albers CA, Zhang ZD, Conrad DF, Lunter G, Zheng H, Ayub Q, DePristo MA, Banks E, Hu M, Handsaker RE, Rosenfeld JA, Fromer M, Jin M, Mu XJ, Khurana E, Ye K, Kay M, Saunders GI, Suner M-M, Hunt T, Barnes IHA, Amid C, Carvalho-Silva DR, Bignell AH, Snow C, Yngvadottir B, Bumpstead S, Cooper DN, Xue Y, Romero IG, Wang J, Li Y, Gibbs RA, McCarroll SA, Dermitzakis ET, Pritchard JK, Barrett JC, Harrow J, Hurles ME, Gerstein MB, Tyler-Smith C. A systematic survey of loss-of-function variants in human protein-coding genes. Science. 2012 Feb 17;335(6070):823–828. PMID: 22344438
15. Rivas MA, Pirinen M, Conrad DF, Lek M, Tsang EK, Karczewski KJ, Maller JB, Kukurba KR, DeLuca DS, Fromer M, Ferreira PG, Smith KS, Zhang R, Zhao F, Banks E, Poplin R, Ruderfer DM, Purcell SM, Tukiainen T, Minikel EV, Stenson PD, Cooper DN, Huang KH, Sullivan TJ, Nedzel J, GTEx Consortium, Geuvadis Consortium, Bustamante CD, Li JB, Daly MJ, Guigo R,

Donnelly P, Ardlie K, Sammeth M, Dermitzakis ET, McCarthy MI, Montgomery SB, Lappalainen T, MacArthur DG. Human genomics. Effect of predicted protein-truncating genetic variants on the human transcriptome. Science. 2015 May 8;348(6235):666–669. PMCID: PMC4537935

16. Högenauer C, Santa Ana CA, Porter JL, Millard M, Gelfand A, Rosenblatt RL, Prestidge CB, Fordtran JS. Active intestinal chloride secretion in human carriers of cystic fibrosis mutations: an evaluation of the hypothesis that heterozygotes have subnormal active intestinal chloride secretion. Am J Hum Genet. 2000 Dec;67(6):1422–1427. PMCID: PMC1287919
17. Zambrowicz BP, Sands AT. Knockouts model the 100 best-selling drugs--will they model the next 100? Nat Rev Drug Discov. 2003;2(1):38–51. PMID: 12509758
18. Karczewski KJ, Francioli LC, Genome Aggregation Database Consortium, Neale BM, Daly MJ, MacArthur DG. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance of human protein-coding genes. In preparation.
19. Schreiber S. A chemical biology view of bioactive small molecules and a binder-based approach to connect biology to precision medicines. bioRxiv [Internet]. 2018 Jan 1; Available from: http://biorxiv.org/content/early/2018/08/08/386904.abstract
20. Haggarty SJ, Koeller KM, Wong JC, Grozinger CM, Schreiber SL. Domain-selective small-molecule inhibitor of histone deacetylase 6 (HDAC6)-mediated tubulin deacetylation. Proc Natl Acad Sci U S A. 2003 Apr 15;100(8):4389–4394. PMCID: PMC153564
21. Zhang BW, Zimmer G, Chen J, Ladd D, Li E, Alt FW, Wiederrecht G, Cryan J, O'Neill EA, Seidman CE, Abbas AK, Seidman JG. T cell responses in calcineurin A alpha-deficient mice. J Exp Med. 1996 Feb 1;183(2):413–420. PMCID: PMC2192457
22. Jacinto E, Loewith R, Schmidt A, Lin S, Rüegg MA, Hall A, Hall MN. Mammalian TOR complex 2 controls the actin cytoskeleton and is rapamycin insensitive. Nat Cell Biol. 2004 Nov;6(11):1122–1128. PMID: 15467718
23. Hoshi N, Langeberg LK, Gould CM, Newton AC, Scott JD. Interaction with AKAP79 modifies the cellular pharmacology of PKC. Mol Cell. 2010 Feb 26;37(4):541–550. PMCID: PMC3014287
24. Cummings BB. Transcript expression-aware annotation increases power for rare variant discovery in Mendelian and complex disease. In preparation.
25. Nousbeck J, Burger B, Fuchs-Telem D, Pavlovsky M, Fenig S, Sarig O, Itin P, Sprecher E. A mutation in a skin-specific isoform of SMARCAD1 causes autosomal-dominant adermatoglyphia. Am J Hum Genet. 2011 Aug 12;89(2):302–307. PMCID: PMC3155166
26. Guven A, Tolun A. TBC1D24 truncating mutation resulting in severe neurodegeneration. J Med Genet. 2013 Mar;50(3):199–202. PMID: 23343562
27. Liao P, Soong TW. CaV1.2 channelopathies: from arrhythmias to autism, bipolar disorder, and immunodeficiency. Pflugers Arch. 2010 Jul;460(2):353–359. PMID: 19916019
28. Uhl K, Kennedy DL, Kweder SL. Risk management strategies in the Physicians' Desk Reference product labels for pregnancy category X drugs. Drug Saf. 2002;25(12):885–892. PMID: 12241129
29. TG and HDL Working Group of the Exome Sequencing Project, National Heart, Lung, and Blood Institute, Crosby J, Peloso GM, Auer PL, Crosslin DR, Stitziel NO, Lange LA, Lu Y, Tang Z, Zhang H, Hindy G, Masca N, Stirrups K, Kanoni S, Do R, Jun G, Hu Y, Kang HM, Xue C, Goel A, Farrall M, Duga S, Merlini PA, Asselta R, Girelli D, Olivieri O, Martinelli N, Yin W, Reilly D, Speliotes E, Fox CS, Hveem K, Holmen OL, Nikpay M, Farlow DN, Assimes TL, Franceschini N, Robinson J, North KE, Martin LW, DePristo M, Gupta N, Escher SA, Jansson J-H, Van Zuydam N, Palmer CNA, Wareham N, Koch W, Meitinger T, Peters A, Lieb W, Erbel R, Konig IR, Kruppa J, Degenhardt F, Gottesman O, Bottinger EP, O'Donnell CJ, Psaty BM, Ballantyne CM, Abecasis G, Ordovas JM, Melander O, Watkins H, Orho-Melander M, Ardissino D, Loos RJF, McPherson R, Willer CJ, Erdmann J, Hall AS, Samani NJ, Deloukas P, Schunkert H, Wilson JG, Kooperberg C, Rich SS, Tracy RP, Lin D-Y, Altshuler D, Gabriel S, Nickerson DA, Jarvik GP, Cupples LA, Reiner AP, Boerwinkle E, Kathiresan S. Loss-of-function mutations in APOC3, triglycerides, and coronary disease. N Engl J Med. 2014 Jul 3;371(1):22–31. PMCID: PMC4180269
30. Myocardial Infarction Genetics Consortium Investigators, Stitziel NO, Won H-H, Morrison AC, Peloso GM, Do R, Lange LA, Fontanillas P, Gupta N, Duga S, Goel A, Farrall M, Saleheen D, Ferrario P, König I, Asselta R, Merlini PA, Marziliano N, Notarangelo MF, Schick U, Auer P, Assimes TL, Reilly M, Wilensky R, Rader DJ, Hovingh GK, Meitinger T, Kessler T, Kastrati A, Laugwitz K-L, Siscovick D, Rotter JI, Hazen SL, Tracy R, Cresci S, Spertus J, Jackson R, Schwartz SM, Natarajan P, Crosby J, Muzny D, Ballantyne C, Rich SS, O'Donnell CJ, Abecasis

G, Sunyaev S, Nickerson DA, Buring JE, Ridker PM, Chasman DI, Austin E, Ye Z, Kullo IJ, Weeke PE, Shaffer CM, Bastarache LA, Denny JC, Roden DM, Palmer C, Deloukas P, Lin D-Y, Tang Z, Erdmann J, Schunkert H, Danesh J, Marrugat J, Elosua R, Ardissino D, McPherson R, Watkins H, Reiner AP, Wilson JG, Altshuler D, Gibbs RA, Lander ES, Boerwinkle E, Gabriel S, Kathiresan S. Inactivating mutations in NPC1L1 and protection from coronary heart disease. N Engl J Med. 2014 Nov 27;371(22):2072–2082. PMCID: PMC4335708

31. Emdin CA, Khera AV, Natarajan P, Klarin D, Won H-H, Peloso GM, Stitziel NO, Nomura A, Zekavat SM, Bick AG, Gupta N, Asselta R, Duga S, Merlini PA, Correa A, Kessler T, Wilson JG, Bown MJ, Hall AS, Braund PS, Samani NJ, Schunkert H, Marrugat J, Elosua R, McPherson R, Farrall M, Watkins H, Willer C, Abecasis GR, Felix JF, Vasan RS, Lander E, Rader DJ, Danesh J, Ardissino D, Gabriel S, Saleheen D, Kathiresan S, CHARGE–Heart Failure Consortium, CARDIoGRAM Exome Consortium. Phenotypic Characterization of Genetically Lowered Human Lipoprotein(a) Levels. J Am Coll Cardiol. 2016 Dec 27;68(25):2761–2772. PMCID: PMC5328146
32. Nguyen PA, Born DA, Deaton AM, Nioi P, Ward LD. Phenotypes associated with genes encoding drug targets are predictive of clinical trial side effects. bioRxiv. 2018 Jan 1;285858.
33. Haas JT, Winter HS, Lim E, Kirby A, Blumenstiel B, DeFelice M, Gabriel S, Jalas C, Branski D, Grueter CA, Toporovski MS, Walther TC, Daly MJ, Farese RV. DGAT1 mutation is linked to a congenital diarrheal disorder. J Clin Invest. 2012 Dec;122(12):4680–4684. PMCID: PMC3533555
34. Denison H, Nilsson C, Kujacic M, Löfgren L, Karlsson C, Knutsson M, Eriksson JW. Proof of mechanism for the DGAT1 inhibitor AZD7687: results from a first-time-in-human single-dose study. Diabetes Obes Metab. 2013 Feb;15(2):136–143. PMID: 22950654
35. Chong JX, Buckingham KJ, Jhangiani SN, Boehm C, Sobreira N, Smith JD, Harrell TM, McMillin MJ, Wiszniewski W, Gambin T, Coban Akdemir ZH, Doheny K, Scott AF, Avramopoulos D, Chakravarti A, Hoover-Fong J, Mathews D, Witmer PD, Ling H, Hetrick K, Watkins L, Patterson KE, Reinier F, Blue E, Muzny D, Kircher M, Bilguvar K, López-Giráldez F, Sutton VR, Tabor HK, Leal SM, Gunel M, Mane S, Gibbs RA, Boerwinkle E, Hamosh A, Shendure J, Lupski JR, Lifton RP, Valle D, Nickerson DA, Centers for Mendelian Genomics, Bamshad MJ. The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. Am J Hum Genet. 2015 Aug 6;97(2):199–215. PMID: 26166479
36. Samocha KE, Robinson EB, Sanders SJ, Stevens C, Sabo A, McGrath LM, Kosmicki JA, Rehnström K, Mallick S, Kirby A, Wall DP, MacArthur DG, Gabriel SB, DePristo M, Purcell SM, Palotie A, Boerwinkle E, Buxbaum JD, Cook EH, Gibbs RA, Schellenberg GD, Sutcliffe JS, Devlin B, Roeder K, Neale BM, Daly MJ. A framework for the interpretation of de novo mutation in human disease. Nat Genet. 2014 Sep;46(9):944–950. PMCID: PMC4222185
37. Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. Genic intolerance to functional variation and the interpretation of personal genomes. PLoS Genet. 2013;9(8):e1003709. PMCID: PMC3749936
38. Aggarwala V, Voight BF. An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. Nat Genet. 2016 Apr;48(4):349–355. PMCID: PMC4811712
39. Fuller Z, Berg JJ, Mostafavi H, Sella G, Przeworski M. Measuring intolerance to mutation in human genetics. bioRxiv [Internet]. 2018 Jan 1; Available from: http://biorxiv.org/content/early/2018/08/01/382481.abstract
40. Weghorn D, Balick DJ, Cassa C, Kosmicki J, Daly MJ, Beier DR, Sunyaev SR. Applicability of the mutation-selection balance model to population genetics of heterozygous protein-truncating variants in humans. bioRxiv [Internet]. 2018 Jan 1; Available from: http://biorxiv.org/content/early/2018/10/03/433961.abstract
41. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, Assempour N, Iynkkaran I, Liu Y, Maciejewski A, Gale N, Wilson A, Chin L, Cummings R, Le D, Pon A, Knox C, Wilson M. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2018 Jan 4;46(D1):D1074–D1082. PMCID: PMC5753335
42. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, Tukiainen T, Birnbaum DP, Kosmicki JA, Duncan LE, Estrada K, Zhao F, Zou J, Pierce-Hoffman E, Berghout J, Cooper DN, Deflaux N, DePristo M, Do R, Flannick J, Fromer M, Gauthier L, Goldstein J, Gupta N, Howrigan D, Kiezun A, Kurki MI, Moonshine AL, Natarajan P, Orozco L, Peloso GM, Poplin R, Rivas MA, Ruano-Rubio V, Rose SA, Ruderfer DM,

Shakir K, Stenson PD, Stevens C, Thomas BP, Tiao G, Tusie-Luna MT, Weisburd B, Won H-H, Yu D, Altshuler DM, Ardissino D, Boehnke M, Danesh J, Donnelly S, Elosua R, Florez JC, Gabriel SB, Getz G, Glatt SJ, Hultman CM, Kathiresan S, Laakso M, McCarroll S, McCarthy MI, McGovern D, McPherson R, Neale BM, Palotie A, Purcell SM, Saleheen D, Scharf JM, Sklar P, Sullivan PF, Tuomilehto J, Tsuang MT, Watkins HC, Wilson JG, Daly MJ, MacArthur DG, Exome Aggregation Consortium. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016 Aug 18;536(7616):285–291. PMCID: PMC5018207

43. Nozawa M, Kawahara Y, Nei M. Genomic drift and copy number variation of sensory receptor genes in humans. Proc Natl Acad Sci U S A. 2007 Dec 18;104(51):20421–20426. PMCID: PMC2154446

44. Rehm HL, Berg JS, Brooks LD, Bustamante CD, Evans JP, Landrum MJ, Ledbetter DH, Maglott DR, Martin CL, Nussbaum RL, Plon SE, Ramos EM, Sherry ST, Watson MS, ClinGen. ClinGen--the Clinical Genome Resource. N Engl J Med. 2015 04;372(23):2235–2242. PMCID: PMC4474187

45. Morham SG, Langenbach R, Loftin CD, Tiano HF, Vouloumanos N, Jennette JC, Mahler JF, Kluckman KD, Ledford A, Lee CA, Smithies O. Prostaglandin synthase 2 gene disruption causes severe renal pathology in the mouse. Cell. 1995 Nov 3;83(3):473–482. PMID: 8521477

46. Ohashi K, Osuga J, Tozawa R, Kitamine T, Yagyu H, Sekiya M, Tomita S, Okazaki H, Tamura Y, Yahagi N, Iizuka Y, Harada K, Gotoda T, Shimano H, Yamada N, Ishibashi S. Early embryonic lethality caused by targeted disruption of the 3-hydroxy-3-methylglutaryl-CoA reductase gene. J Biol Chem. 2003 Oct 31;278(44):42936–42941. PMID: 12920113

47. Nagashima S, Yagyu H, Ohashi K, Tazoe F, Takahashi M, Ohshiro T, Bayasgalan T, Okada K, Sekiya M, Osuga J, Ishibashi S. Liver-specific deletion of 3-hydroxy-3-methylglutaryl coenzyme A reductase causes hepatic steatosis and death. Arterioscler Thromb Vasc Biol. 2012 Aug;32(8):1824–1831. PMID: 22701022

48. Overington JP, Al-Lazikani B, Hopkins AL. How many drug targets are there? Nat Rev Drug Discov. 2006 Dec;5(12):993–996. PMID: 17139284

49. Imming P, Sinning C, Meyer A. Drugs, their targets and the nature and number of drug targets. Nat Rev Drug Discov. 2006 Oct;5(10):821–834. PMID: 17016423

50. GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, NIH/NIMH, NIH/NIDA, Biospecimen Collection Source Site—NDRI, Biospecimen Collection Source Site—RPCI, Biospecimen Core Resource—VARI, Brain Bank Repository—University of Miami Brain Endowment Bank, Leidos Biomedical—Project Management, ELSI Study, Genome Browser Data Integration &Visualization—EBI, Genome Browser Data Integration &Visualization—UCSC Genomics Institute, University of California Santa Cruz, Lead analysts:, Laboratory, Data Analysis &Coordinating Center (LDACC):, NIH program management:, Biospecimen collection:, Pathology:, eQTL manuscript working group:, Battle A, Brown CD, Engelhardt BE, Montgomery SB. Genetic effects on gene expression across human tissues. Nature. 2017 11;550(7675):204–213. PMCID: PMC5776756

51. Lim ET, Würtz P, Havulinna AS, Palta P, Tukiainen T, Rehnström K, Esko T, Mägi R, Inouye M, Lappalainen T, Chan Y, Salem RM, Lek M, Flannick J, Sim X, Manning A, Ladenvall C, Bumpstead S, Hämäläinen E, Aalto K, Maksimow M, Salmi M, Blankenberg S, Ardissino D, Shah S, Horne B, McPherson R, Hovingh GK, Reilly MP, Watkins H, Goel A, Farrall M, Girelli D, Reiner AP, Stitziel NO, Kathiresan S, Gabriel S, Barrett JC, Lehtimäki T, Laakso M, Groop L, Kaprio J, Perola M, McCarthy MI, Boehnke M, Altshuler DM, Lindgren CM, Hirschhorn JN, Metspalu A, Freimer NB, Zeller T, Jalkanen S, Koskinen S, Raitakari O, Durbin R, MacArthur DG, Salomaa V, Ripatti S, Daly MJ, Palotie A, Sequencing Initiative Suomi (SISu) Project. Distribution and medical impact of loss-of-function variants in the Finnish founder population. PLoS Genet. 2014 Jul;10(7):e1004494. PMCID: PMC4117444

52. Minikel EV, Vallabh SM, Lek M, Estrada K, Samocha KE, Sathirapongsasuti JF, McLean CY, Tung JY, Yu LPC, Gambetti P, Blevins J, Zhang S, Cohen Y, Chen W, Yamada M, Hamaguchi T, Sanjo N, Mizusawa H, Nakamura Y, Kitamoto T, Collins SJ, Boyd A, Will RG, Knight R, Ponto C, Zerr I, Kraus TFJ, Eigenbrod S, Giese A, Calero M, de Pedro-Cuesta J, Haïk S, Laplanche J-L, Bouaziz-Amar E, Brandel J-P, Capellari S, Parchi P, Poleggi A, Ladogana A, O’Donnell-Luria AH, Karczewski KJ, Marshall JL, Boehnke M, Laakso M, Mohlke KL, Kähler A, Chambert K, McCarroll

S, Sullivan PF, Hultman CM, Purcell SM, Sklar P, van der Lee SJ, Rozemuller A, Jansen C, Hofman A, Kraaij R, van Rooij JGJ, Ikram MA, Uitterlinden AG, van Duijn CM, Exome Aggregation Consortium (ExAC), Daly MJ, MacArthur DG. Quantifying prion disease penetrance using large population control cohorts. Sci Transl Med. 2016 Jan 20;8(322):322ra9. PMCID: PMC4774245

53. Saleheen D, Natarajan P, Armean IM, Zhao W, Rasheed A, Khetarpal SA, Won H-H, Karczewski KJ, O'Donnell-Luria AH, Samocha KE, Weisburd B, Gupta N, Zaidi M, Samuel M, Imran A, Abbas S, Majeed F, Ishaq M, Akhtar S, Trindade K, Mucksavage M, Qamar N, Zaman KS, Yaqoob Z, Saghir T, Rizvi SNH, Memon A, Hayyat Mallick N, Ishaq M, Rasheed SZ, Memon F-U-R, Mahmood K, Ahmed N, Do R, Krauss RM, MacArthur DG, Gabriel S, Lander ES, Daly MJ, Frossard P, Danesh J, Rader DJ, Kathiresan S. Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. Nature. 2017 12;544(7649):235–239. PMCID: PMC5600291

54. Whiffin N. Human loss-of-function variants suggest that partial LRRK2 inhibition is a safe therapeutic strategy for Parkinson's disease. In preparation.

55. Sulem P, Helgason H, Oddson A, Stefansson H, Gudjonsson SA, Zink F, Hjartarson E, Sigurdsson GT, Jonasdottir A, Jonasdottir A, Sigurdsson A, Magnusson OT, Kong A, Helgason A, Holm H, Thorsteinsdottir U, Masson G, Gudbjartsson DF, Stefansson K. Identification of a large set of rare complete human knockouts. Nat Genet. 2015 May;47(5):448–452. PMID: 25807282

56. Finer S, Martin H, Hunt KA, MacLaughlin B, Ashcroft R, Khan A, McCarthy MI, Robson J, MacArthur D, Griffiths C, Wright J, Trembath RC, van Heel D. Cohort Profile: East London Genes &amp; Health (ELGH), a community based population genomics and health study in people of British-Bangladeshi and -Pakistani heritage. bioRxiv. 2018 Jan 1;426163.

57. Narasimhan VM, Hunt KA, Mason D, Baker CL, Karczewski KJ, Barnes MR, Barnett AH, Bates C, Bellary S, Bockett NA, Giorda K, Griffiths CJ, Hemingway H, Jia Z, Kelly MA, Khawaja HA, Lek M, McCarthy S, McEachan R, O'Donnell-Luria A, Paigen K, Parisinos CA, Sheridan E, Southgate L, Tee L, Thomas M, Xue Y, Schnall-Levin M, Petkov PM, Tyler-Smith C, Maher ER, Trembath RC, MacArthur DG, Wright J, Durbin R, van Heel DA. Health and population effects of rare gene knockouts in adult humans with related parents. Science. 2016 Apr 22;352(6284):474–477. PMCID: PMC4985238

58. MacGregor T et al. Deep phenotyping of a healthy human HAO1 knockout informs therapeutic development for primary hyperoxaluria. In preparation.

59. Saleheen D, Zaidi M, Rasheed A, Ahmad U, Hakeem A, Murtaza M, Kayani W, Faruqui A, Kundi A, Zaman KS, Yaqoob Z, Cheema LA, Samad A, Rasheed SZ, Mallick NH, Azhar M, Jooma R, Gardezi AR, Memon N, Ghaffar A, Fazal-ur-Rehman null, Khan N, Shah N, Ali Shah A, Samuel M, Hanif F, Yameen M, Naz S, Sultana A, Nazir A, Raza S, Shazad M, Nasim S, Javed MA, Ali SS, Jafree M, Nisar MI, Daood MS, Hussain A, Sarwar N, Kamal A, Deloukas P, Ishaq M, Frossard P, Danesh J. The Pakistan Risk of Myocardial Infarction Study: a resource for the study of genetic, lifestyle and other determinants of myocardial infarction in South Asia. Eur J Epidemiol. 2009;24(6):329–338. PMCID: PMC2697028

60. Wild EJ, Tabrizi SJ. Therapies targeting DNA and RNA in Huntington's disease. Lancet Neurol. 2017 Oct;16(10):837–847. PMCID: PMC5604739

61. Deng X, Dzamko N, Prescott A, Davies P, Liu Q, Yang Q, Lee J-D, Patricelli MP, Nomanbhoy TK, Alessi DR, Gray NS. Characterization of a selective inhibitor of the Parkinson's disease kinase LRRK2. Nat Chem Biol. 2011 Apr;7(4):203–205. PMCID: PMC3287420

62. DeVos SL, Miller RL, Schoch KM, Holmes BB, Kebodeaux CS, Wegener AJ, Chen G, Shen T, Tran H, Nichols B, Zanardi TA, Kordasiewicz HB, Swayze EE, Bennett CF, Diamond MI, Miller TM. Tau reduction prevents neuronal loss and reverses pathological tau deposition and seeding in mice with tauopathy. Sci Transl Med. 2017 Jan 25;9(374). PMID: 28123067

63. Vallabh S, Minikel EV, Schreiber SL, Lander ES. A path to prevention of genetic prion disease. In preparation.

64. Bennett CF, Freier SM, Mallajosyula J. Modulation of alpha synuclein expression [Internet]. US20140005252A1, 2014 [cited 2018 Dec 6]. Available from: https://patents.google.com/patent/US20140005252A1/en/

65. McCampbell A, Cole T, Wegener AJ, Tomassy GS, Setnicka A, Farley BJ, Schoch KM, Hoye ML, Shabsovich M, Sun L, Luo Y, Zhang M, Thankamony S, Salzman DW, Cudkowicz M, Graham DL,

Bennett CF, Kordasiewicz HB, Swayze EE, Miller TM, Comfort N, Wang B, Amacker J. Antisense oligonucleotides extend survival and reverse decrement in muscle response in ALS models. J Clin Invest. 2018 01;128(8):3558–3567. PMCID: PMC6063493

66. Zhao HT, John N, Delic V, Ikeda-Lee K, Kim A, Weihofen A, Swayze EE, Kordasiewicz HB, West AB, Volpicelli-Daley LA. LRRK2 Antisense Oligonucleotides Ameliorate α-Synuclein Inclusion Formation in a Parkinson's Disease Mouse Model. Mol Ther Nucleic Acids. 2017 Sep 15;8:508–519. PMCID: PMC5573879

67. Estrada AA, Liu X, Baker-Glenn C, Beresford A, Burdick DJ, Chambers M, Chan BK, Chen H, Ding X, DiPasquale AG, Dominguez SL, Dotson J, Drummond J, Flagella M, Flynn S, Fuji R, Gill A, Gunzner-Toste J, Harris SF, Heffron TP, Kleinheinz T, Lee DW, Le Pichon CE, Lyssikatos JP, Medhurst AD, Moffat JG, Mukund S, Nash K, Scearce-Levie K, Sheng Z, Shore DG, Tran T, Trivedi N, Wang S, Zhang S, Zhang X, Zhao G, Zhu H, Sweeney ZK. Discovery of highly potent, selective, and brain-penetrable leucine-rich repeat kinase 2 (LRRK2) small molecule inhibitors. J Med Chem. 2012 Nov 26;55(22):9416–9433. PMID: 22985112

68. Estrada AA, Chan BK, Baker-Glenn C, Beresford A, Burdick DJ, Chambers M, Chen H, Dominguez SL, Dotson J, Drummond J, Flagella M, Fuji R, Gill A, Halladay J, Harris SF, Heffron TP, Kleinheinz T, Lee DW, Le Pichon CE, Liu X, Lyssikatos JP, Medhurst AD, Moffat JG, Nash K, Scearce-Levie K, Sheng Z, Shore DG, Wong S, Zhang S, Zhang X, Zhu H, Sweeney ZK. Discovery of highly potent, selective, and brain-penetrant aminopyrazole leucine-rich repeat kinase 2 (LRRK2) small molecule inhibitors. J Med Chem. 2014 Feb 13;57(3):921–936. PMID: 24354345

69. Pringsheim T, Wiltshire K, Day L, Dykeman J, Steeves T, Jette N. The incidence and prevalence of Huntington's disease: a systematic review and meta-analysis. Mov Disord Off J Mov Disord Soc. 2012 Aug;27(9):1083–1091. PMID: 22692795

70. Fisher ER, Hayden MR. Multisource ascertainment of Huntington disease in Canada: prevalence and population at risk. Mov Disord Off J Mov Disord Soc. 2014 Jan;29(1):105–114. PMID: 24151181

71. Kay C, Collins JA, Miedzybrodzka Z, Madore SJ, Gordon ES, Gerry N, Davidson M, Slama RA, Hayden MR. Huntington disease reduced penetrance alleles occur at high frequency in the general population. Neurology. 2016 Jul 19;87(3):282–288. PMCID: PMC4955276

72. Pringsheim T, Jette N, Frolkis A, Steeves TDL. The prevalence of Parkinson's disease: a systematic review and meta-analysis. Mov Disord Off J Mov Disord Soc. 2014 Nov;29(13):1583–1590. PMID: 24976103

73. Healy DG, Falchi M, O'Sullivan SS, Bonifati V, Durr A, Bressman S, Brice A, Aasly J, Zabetian CP, Goldwurm S, Ferreira JJ, Tolosa E, Kay DM, Klein C, Williams DR, Marras C, Lang AE, Wszolek ZK, Berciano J, Schapira AHV, Lynch T, Bhatia KP, Gasser T, Lees AJ, Wood NW, International LRRK2 Consortium. Phenotype, genotype, and worldwide genetic penetrance of LRRK2-associated Parkinson's disease: a case-control study. Lancet Neurol. 2008 Jul;7(7):583–590. PMCID: PMC2832754

74. Onyike CU, Diehl-Schmid J. The epidemiology of frontotemporal dementia. Int Rev Psychiatry Abingdon Engl. 2013 Apr;25(2):130–137. PMCID: PMC3932112

75. Bang J, Spina S, Miller BL. Frontotemporal dementia. Lancet Lond Engl. 2015 Oct 24;386(10004):1672–1682. PMCID: PMC5970949

76. Minikel EV, Vallabh SM, Orseth MC, Brandel J-P, Haik S, Laplanche J-L, Zerr I, Parchi P, Capellari S, Safar J, Kenny J, Fong JC, Takada LT, Ponto C, Hermann P, Knipper T, Stehmann C, Kitamoto T, Ae R, Hamaguchi T, Sanjo N, Tsukamoto T, Mizusawa H, Collins SJ, Chiesa R, Roiter I, de Pedro-Cuesta J, Calero M, Geschwind MD, Yamada M, Nakamura Y, Mead S. Age of onset in genetic prion disease and the design of preventive clinical trials. bioRxiv [Internet]. 2018 Aug 29; Available from: http://biorxiv.org/content/early/2018/08/29/401406.abstract

77. Trinh J, Guella I, Farrer MJ. Disease penetrance of late-onset parkinsonism: a meta-analysis. JAMA Neurol. 2014 Dec;71(12):1535–1539. PMID: 25330418

78. Andreadis A. Tau splicing and the intricacies of dementia. J Cell Physiol. 2012 Mar;227(3):1220–1225. PMCID: PMC3177961

79. Bommarito G, Cellerino M, Prada V, Venturi C, Capellari S, Cortelli P, Mancardi GL, Parchi P, Schenone A. A novel prion protein gene-truncating mutation causing autonomic neuropathy and diarrhea. Eur J Neurol. 2018 Aug;25(8):e91–e92. PMID: 29984897

80. Capellari S, Baiardi S, Rinaldi R, Bartoletti-Stella A, Graziano C, Piras S, Calandra-Buonaura G, D'Angelo R, Terziotti C, Lodi R, Donadio V, Pironi L, Cortelli P, Parchi P. Two novel PRNP truncating mutations broaden the spectrum of prion amyloidosis. Ann Clin Transl Neurol. 2018 Jun;5(6):777–783. PMCID: PMC5989776
81. Duyao MP, Auerbach AB, Ryan A, Persichetti F, Barnes GT, McNeil SM, Ge P, Vonsattel JP, Gusella JF, Joyner AL. Inactivation of the mouse Huntington's disease gene homolog Hdh. Science. 1995 Jul 21;269(5222):407–410. PMID: 7618107
82. Ambrose CM, Duyao MP, Barnes G, Bates GP, Lin CS, Srinidhi J, Baxendale S, Hummerich H, Lehrach H, Altherr M. Structure and expression of the Huntington's disease gene: evidence against simple inactivation due to an expanded CAG repeat. Somat Cell Mol Genet. 1994 Jan;20(1):27–38. PMID: 8197474
83. Rodan LH, Cohen J, Fatemi A, Gillis T, Lucente D, Gusella J, Picker JD. A novel neurodevelopmental disorder associated with compound heterozygous variants in the huntingtin gene. Eur J Hum Genet EJHG. 2016 Jun 22; PMID: 27329733
84. Lopes F, Barbosa M, Ameur A, Soares G, de Sá J, Dias AI, Oliveira G, Cabral P, Temudo T, Calado E, Cruz IF, Vieira JP, Oliveira R, Esteves S, Sauer S, Jonasson I, Syvänen A-C, Gyllensten U, Pinto D, Maciel P. Identification of novel genetic causes of Rett syndrome-like phenotypes. J Med Genet. 2016 Mar;53(3):190–199. PMID: 26740508
85. Van Raamsdonk JM, Gibson WT, Pearson J, Murphy Z, Lu G, Leavitt BR, Hayden MR. Body weight is modulated by levels of full-length huntingtin. Hum Mol Genet. 2006 May 1;15(9):1513–1523. PMID: 16571604
86. Rovelet-Lecrux A, Lecourtois M, Thomas-Anterion C, Le Ber I, Brice A, Frebourg T, Hannequin D, Campion D. Partial deletion of the MAPT gene: a novel mechanism of FTDP-17. Hum Mutat. 2009 Apr;30(4):E591-602. PMID: 19263483
87. Shaw-Smith C, Pittman AM, Willatt L, Martin H, Rickman L, Gribble S, Curley R, Cumming S, Dunn C, Kalaitzopoulos D, Porter K, Prigmore E, Krepischi-Santos ACV, Varela MC, Koiffmann CP, Lees AJ, Rosenberg C, Firth HV, de Silva R, Carter NP. Microdeletion encompassing MAPT at chromosome 17q21.3 is associated with developmental delay and learning disability. Nat Genet. 2006 Sep;38(9):1032–1037. PMID: 16906163
88. Koolen DA, Kramer JM, Neveling K, Nillesen WM, Moore-Barton HL, Elmslie FV, Toutain A, Amiel J, Malan V, Tsai AC-H, Cheung SW, Gilissen C, Verwiel ETP, Martens S, Feuth T, Bongers EMHF, de Vries P, Scheffer H, Vissers LELM, de Brouwer APM, Brunner HG, Veltman JA, Schenck A, Yntema HG, de Vries BBA. Mutations in the chromatin modifier gene KANSL1 cause the 17q21.31 microdeletion syndrome. Nat Genet. 2012 Apr 29;44(6):639–641. PMID: 22544363
89. Harada A, Oguchi K, Okabe S, Kuno J, Terada S, Ohshima T, Sato-Yoshitake R, Takei Y, Noda T, Hirokawa N. Altered microtubule organization in small-calibre axons of mice lacking tau protein. Nature. 1994 Jun 9;369(6480):488–491. PMID: 8202139
90. Tan DCS, Yao S, Ittner A, Bertz J, Ke YD, Ittner LM, Delerue F. Generation of a New Tau Knockout (tauΔex1) Line Using CRISPR/Cas9 Genome Editing in Mice. J Alzheimers Dis JAD. 2018;62(2):571–578. PMID: 29480201
91. Mead S, Reilly MM. A new prion disease: relationship with central and peripheral amyloidoses. Nat Rev Neurol. 2015 Jan 27; PMID: 25623792
92. Carlston CM, O'Donnell-Luria AH, Underhill HR, Cummings BB, Weisburd B, Minikel EV, Birnbaum DP, Exome Aggregation Consortium, Tvrdik T, MacArthur DG, Mao R. Pathogenic ASXL1 somatic variants in reference databases complicate germline variant interpretation for Bohring-Opitz Syndrome. Hum Mutat. 2017;38(5):517–523. PMCID: PMC5487276
93. Shaw ND, Brand H, Kupchinsky ZA, Bengani H, Plummer L, Jones TI, Erdin S, Williamson KA, Rainger J, Stortchevoi A, Samocha K, Currall BB, Dunican DS, Collins RL, Willer JR, Lek A, Lek M, Nassan M, Pereira S, Kammin T, Lucente D, Silva A, Seabra CM, Chiang C, An Y, Ansari M, Rainger JK, Joss S, Smith JC, Lippincott MF, Singh SS, Patel N, Jing JW, Law JR, Ferraro N, Verloes A, Rauch A, Steindl K, Zweier M, Scheer I, Sato D, Okamoto N, Jacobsen C, Tryggestad J, Chernausek S, Schimmenti LA, Brasseur B, Cesaretti C, García-Ortiz JE, Buitrago TP, Silva OP, Hoffman JD, Mühlbauer W, Ruprecht KW, Loeys BL, Shino M, Kaindl AM, Cho C-H, Morton CC, Meehan RR, van Heyningen V, Liao EC, Balasubramanian R, Hall JE, Seminara SB, Macarthur D, Moore SA, Yoshiura K-I, Gusella JF, Marsh JA, Graham JM, Lin AE, Katsanis N, Jones PL, Crowley WF, Davis EE, FitzPatrick DR, Talkowski ME. SMCHD1 mutations associated

with a rare muscular dystrophy can also cause isolated arhinia and Bosma arhinia microphthalmia syndrome. Nat Genet. 2017 Feb;49(2):238–248. PMCID: PMC5473428

94. Yates B, Braschi B, Gray KA, Seal RL, Tweedie S, Bruford EA. Genenames.org: the HGNC and VGNC resources in 2017. Nucleic Acids Res. 2017 04;45(D1):D619–D625. PMCID: PMC5210531

95. Mainland JD, Li YR, Zhou T, Liu WLL, Matsunami H. Human olfactory receptor responses to odorants. Sci Data. 2015;2:150002. PMCID: PMC4412152

96. Blekhman R, Man O, Herrmann L, Boyko AR, Indap A, Kosiol C, Bustamante CD, Teshima KM, Przeworski M. Natural selection on genes that underlie human disease susceptibility. Curr Biol CB. 2008 Jun 24;18(12):883–889. PMCID: PMC2474766

97. Hart T, Tong AHY, Chan K, Van Leeuwen J, Seetharaman A, Aregger M, Chandrashekhar M, Hustedt N, Seth S, Noonan A, Habsid A, Sizova O, Nedyalkova L, Climie R, Tworzyanski L, Lawson K, Sartori MA, Alibeh S, Tieu D, Masud S, Mero P, Weiss A, Brown KR, Usaj M, Billmann M, Rahman M, Constanzo M, Myers CL, Andrews BJ, Boone C, Durocher D, Moffat J. Evaluation and Design of Genome-Wide CRISPR/SpCas9 Knockout Screens. G3 Bethesda Md. 2017 07;7(8):2719–2727. PMCID: PMC5555476

98. Harding SD, Sharman JL, Faccenda E, Southan C, Pawson AJ, Ireland S, Gray AJG, Bruce L, Alexander SPH, Anderton S, Bryant C, Davenport AP, Doerig C, Fabbro D, Levi-Schaffer F, Spedding M, Davies JA, NC-IUPHAR. The IUPHAR/BPS Guide to PHARMACOLOGY in 2018: updates and expansion to encompass the new guide to IMMUNOPHARMACOLOGY. Nucleic Acids Res. 2018 Jan 4;46(D1):D1091–D1106. PMCID: PMC5753190

99. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, Junkins H, McMahon A, Milano A, Morales J, Pendlington ZM, Welter D, Burdett T, Hindorff L, Flicek P, Cunningham F, Parkinson H. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 2017 04;45(D1):D896–D901. PMCID: PMC5210590

100. Bittles AH, Black ML. Evolution in health and medicine Sackler colloquium: Consanguinity, human evolution, and complex diseases. Proc Natl Acad Sci U S A. 2010 Jan 26;107 Suppl 1:1779–1786. PMCID: PMC2868287

101. Keum JW, Shin A, Gillis T, Mysore JS, Abu Elneel K, Lucente D, Hadzi T, Holmans P, Jones L, Orth M, Kwak S, MacDonald ME, Gusella JF, Lee J-M. The HTT CAG-Expansion Mutation Determines Age at Death but Not Disease Duration in Huntington Disease. Am J Hum Genet. 2016 Feb 4;98(2):287–298. PMCID: PMC4746370

102. Hernandez DG, Reed X, Singleton AB. Genetics in Parkinson disease: Mendelian versus non-Mendelian inheritance. J Neurochem. 2016;139 Suppl 1:59–74. PMCID: PMC5155439

103. Funayama M, Hasegawa K, Kowa H, Saito M, Tsuji S, Obata F. A new locus for Parkinson's disease (PARK8) maps to chromosome 12p11.2-q13.1. Ann Neurol. 2002 Mar;51(3):296–301. PMID: 11891824

104. Zimprich A, Biskup S, Leitner P, Lichtner P, Farrer M, Lincoln S, Kachergus J, Hulihan M, Uitti RJ, Calne DB, Stoessl AJ, Pfeiffer RF, Patenge N, Carbajal IC, Vieregge P, Asmus F, Müller-Myhsok B, Dickson DW, Meitinger T, Strom TM, Wszolek ZK, Gasser T. Mutations in LRRK2 cause autosomal-dominant parkinsonism with pleomorphic pathology. Neuron. 2004 Nov 18;44(4):601–607. PMID: 15541309

105. Goldwurm S, Zini M, Mariani L, Tesei S, Miceli R, Sironi F, Clementi M, Bonifati V, Pezzoli G. Evaluation of LRRK2 G2019S penetrance: relevance for genetic counseling in Parkinson disease. Neurology. 2007 Apr 3;68(14):1141–1143. PMID: 17215492

106. Do CB, Tung JY, Dorfman E, Kiefer AK, Drabant EM, Francke U, Mountain JL, Goldman SM, Tanner CM, Langston JW, Wojcicki A, Eriksson N. Web-based genome-wide association study identifies two novel loci and a substantial genetic component for Parkinson's disease. PLoS Genet. 2011 Jun;7(6):e1002141. PMCID: PMC3121750

107. Kinoshita T, Fujita M. Biosynthesis of GPI-anchored proteins: special emphasis on GPI lipid remodeling. J Lipid Res. 2016 Jan;57(1):6–24. PMCID: PMC4689344

108. Kitamoto T, Iizuka R, Tateishi J. An amber mutation of prion protein in Gerstmann-Sträussler syndrome with mutant PrP plaques. Biochem Biophys Res Commun. 1993 Apr 30;192(2):525–531. PMID: 8097911

109. Finckh U, Müller-Thomsen T, Mann U, Eggers C, Marksteiner J, Meins W, Binetti G, Alberici A, Hock C, Nitsch RM, Gal A. High prevalence of pathogenic mutations in patients with early-onset

dementia detected by sequence analyses of four different genes. Am J Hum Genet. 2000 Jan;66(1):110–117. PMCID: PMC1288316

110. Jayadev S, Nochlin D, Poorkaj P, Steinbart EJ, Mastrianni JA, Montine TJ, Ghetti B, Schellenberg GD, Bird TD, Leverenz JB. Familial prion disease with Alzheimer disease-like tau pathology and clinical phenotype. Ann Neurol. 2011 Apr;69(4):712–720. PMCID: PMC3114566
111. Fong JC, Rojas JC, Bang J, Legati A, Rankin KP, Forner S, Miller ZA, Karydas AM, Coppola G, Grouse CK, Ralph J, Miller BL, Geschwind MD. Genetic Prion Disease Caused by PRNP Q160X Mutation Presenting with an Orbitofrontal Syndrome, Cyclic Diarrhea, and Peripheral Neuropathy. J Alzheimers Dis JAD. 2017;55(1):249–258. PMCID: PMC5149415
112. Mead S, Gandhi S, Beck J, Caine D, Gajulapalli D, Gallujipali D, Carswell C, Hyare H, Joiner S, Ayling H, Lashley T, Linehan JM, Al-Doujaily H, Sharps B, Revesz T, Sandberg MK, Reilly MM, Koltzenburg M, Forbes A, Rudge P, Brandner S, Warren JD, Wadsworth JDF, Wood NW, Holton JL, Collinge J. A novel prion disease associated with diarrhea and autonomic neuropathy. N Engl J Med. 2013 Nov 14;369(20):1904–1914. PMCID: PMC3863770
113. Matsuzono K, Ikeda Y, Liu W, Kurata T, Deguchi S, Deguchi K, Abe K. A novel familial prion disease causing pan-autonomic-sensory neuropathy and cognitive impairment. Eur J Neurol Off J Eur Fed Neurol Soc. 2013 May;20(5):e67-69. PMID: 23577609
114. Jansen C, Parchi P, Capellari S, Vermeij AJ, Corrado P, Baas F, Strammiello R, van Gool WA, van Swieten JC, Rozemuller AJM. Prion protein amyloidosis with divergent phenotype associated with two novel nonsense mutations in PRNP. Acta Neuropathol (Berl). 2010 Feb;119(2):189–197. PMCID: PMC2808512
115. Chiò A, Traynor BJ, Lombardo F, Fimognari M, Calvo A, Ghiglione P, Mutani R, Restagno G. Prevalence of SOD1 mutations in the Italian ALS population. Neurology. 2008 Feb 12;70(7):533–537. PMID: 18268245
116. Cudkowicz ME, McKenna-Yasek D, Sapp PE, Chin W, Geller B, Hayden DL, Schoenfeld DA, Hosler BA, Horvitz HR, Brown RH. Epidemiology of mutations in superoxide dismutase in amyotrophic lateral sclerosis. Ann Neurol. 1997 Feb;41(2):210–221. PMID: 9029070
117. Renton AE, Chiò A, Traynor BJ. State of play in amyotrophic lateral sclerosis genetics. Nat Neurosci. 2014 Jan;17(1):17–23. PMCID: PMC4544832
118. Byrne S, Walsh C, Lynch C, Bede P, Elamin M, Kenna K, McLaughlin R, Hardiman O. Rate of familial amyotrophic lateral sclerosis: a systematic review and meta-analysis. J Neurol Neurosurg Psychiatry. 2011 Jun;82(6):623–627. PMID: 21047878
119. Rowland LP, Shneider NA. Amyotrophic lateral sclerosis. N Engl J Med. 2001 May 31;344(22):1688–1700. PMID: 11386269
120. Hirtz D, Thurman DJ, Gwinn-Hardy K, Mohamed M, Chaudhuri AR, Zalutsky R. How common are the “common” neurologic disorders? Neurology. 2007 Jan 30;68(5):326–337. PMID: 17261678
121. Logroscino G, Traynor BJ, Hardiman O, Chiò A, Mitchell D, Swingler RJ, Millul A, Benn E, Beghi E, EURALS. Incidence of amyotrophic lateral sclerosis in Europe. J Neurol Neurosurg Psychiatry. 2010 Apr;81(4):385–390. PMCID: PMC2850819
122. Bruijn LI, Houseweart MK, Kato S, Anderson KL, Anderson SD, Ohama E, Reaume AG, Scott RW, Cleveland DW. Aggregation and motor neuron toxicity of an ALS-linked SOD1 mutant independent from wild-type SOD1. Science. 1998 Sep 18;281(5384):1851–1854. PMID: 9743498
123. Liu J, Lillo C, Jonsson PA, Vande Velde C, Ward CM, Miller TM, Subramaniam JR, Rothstein JD, Marklund S, Andersen PM, Brännström T, Gredal O, Wong PC, Williams DS, Cleveland DW. Toxicity of familial ALS-linked SOD1 mutants from selective recruitment to spinal mitochondria. Neuron. 2004 Jul 8;43(1):5–17. PMID: 15233913

Supplement

Downsampling of cumulative allele frequency analysis

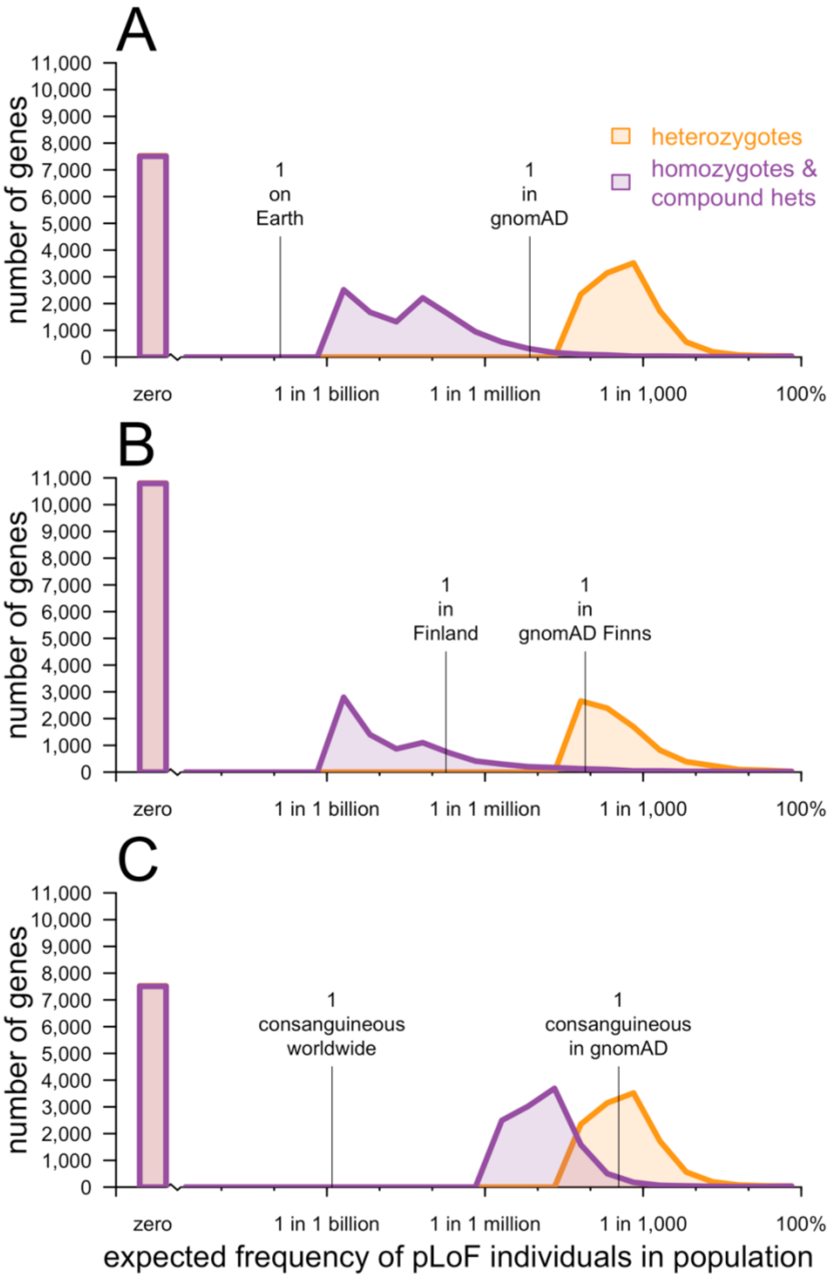

Figure S1. Expected frequency of individuals with one or two null alleles for every protein-coding gene across different population models, with sample size held constant. *This is identical to Figure 4 except for the differences described in the text below.*

As noted in Methods, one caveat about Figure 4 is that the sample size is larger for the plots using all gnomAD exomes (Figure 4A and 4C) than for Finnish exomes (Figure 4B). Figure S1 shows the same analysis as Figure 4, 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. We also used a constant y axis with no axis breaks in Figure S1 to make this difference more clearly visible.

HTT

We considered several approaches to estimating the genetic prevalence of Huntington's disease (HD). A reported HD incidence of 0.38 cases per 100,000 per year based on meta-analysis[69] multiplied by an average age at death of ~60 for the most common CAG lengths[101] gives a genetic prevalence of 1 in 4,386. One exhaustively ascertained study of HD[70] found a prevalence of 13.7 per 100,000 symptomatic plus 81.6 per 100,000 at 25-50% risk. 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, equal to 1 in 2,451. Finally, a genetic screen of a general population sample[71] 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.

LRRK2

Based on meta-analysis[72], 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[73], 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, or 1 in 3,300.

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[102]. some variants exhibit Mendelian segregation with disease[103,104], implying high risk; the G2019S variant is estimated to have ~32% penetrance[105]; and other common variants are risk factors with odds ratios of only ~1.2 estimated through genome-wide association studies (GWAS)[106]. The GWAS-implicated common variants were not included in the case series on which our estimate is based[73], 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%).

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[74] 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[75]. Multiplying all these figures results in range of 0.0032% to 0.020%, or 1 in 5,000 – 31,000.

As noted in the main text, our sample size is not sufficient to prove that *MAPT* loss-of-function is not tolerated. When we restrict to constitutive, brain-expressed exons (Ensembl transcript ENST00000334239), we expect 12.6 pLoF variants and observe 0. The 95% confidence interval on *MAPT* constraint is thus (0%, 23.7%). The upper bound of 23.7% implies that our data do not rule out a true pLoF obs/exp value of up to 3.0/12.6, or in other words, we cannot rule out that another population sample as large as gnomAD might yield up to 3 genuine pLoF variants.

PRNP

We have recently considered the lifetime risk of genetic prion disease in detail[76]. 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[52,76], 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.

Figure 5C displays variants from gnomAD plus the literature, including those previously reported[52], and Table S1 shows details for each variant. Allele count for variants from the literature in Figure 5C is the total number of definite or probable cases with sequencing performed in the studies cited in Table S1. 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[107]: 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.

Table S1. Details of PRNP truncating variants.

variant	allele count	neurological phenotype	comments	reference
G20Gfs84X	1	healthy	As previously reported.	[52]
R37X	2	healthy, unknown	One previously reported, one new.	[52]
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	[52]
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.	[52], this work
Y145X	1	dementia		[108]
Q160X	5	dementia		[109–111]
Y162X	1	dementia		[79]
Y163X	7	dementia		[80,112]
Y169X	2	dementia		[80]
D178Efs25X	1	dementia		[113]
Q186X	1	dementia		[52]
Y226X	1	dementia		[114]
Q227X	1	dementia		[114]
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

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[77]. Thus, we estimate that *SNCA* mutations account for a 1.9% * 10% * 15% = 0.00028% risk of PD in the general population, or 1 in 360,000.

SOD1

SOD1 mutations are believed to account for ~12% to 24% of familial ALS[115,116] and 1% of sporadic ALS[115,117]. One a meta-analysis found that ~4.6% of ALS is familial[118], although a figure of 10% is also often used[119]. 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[120,121], 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[116], implying a 1.2 – 3.7 per 100,000 population prevalence of pathogenic *SOD1* mutations, or a range of 1 in 27,000-83,000.

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.

Table S2. Details of curated variants in neurodegenerative disease genes. *LRRK2 is not included here as curation is reported in detail in a separate publication[54].*

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-AGCAGCAGCAGCAGCAGCAG-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-	1	loftee	lcr		

	AGCAGCAGCAGCAG-A					
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-AGCAGCAGCAGCAGCAG	2	loftee	lcr		

HTT	4-003076663-A-AGCAGCAGCAGCAGCAGCAGCAGCAG	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-CAGCAGCAGCAG	1	loftee	lcr		
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-CATGGTGGTGGCTGGGGGCAGCCCCATGGTGGTGGCTGGGGACAGCCT-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