Published in final edited form as:
Nat Genet. 2017 May ; 49(5): 806–810. doi:10.1038/ng.3831.

Estimating the Selective Effects of Heterozygous Protein Truncating Variants from Human Exome Data

Christopher A. Cassa[1,2,*], **Donate Weghorn**[1,*], **Daniel J. Balick**[1,*], **Daniel M. Jordan**[3,*], **David Nusinow**[1], **Kaitlin E. Samocha**[4,5], **Anne O'Donnell-Luria**[4,6], **Daniel G. MacArthur**[2,4], **Mark J. Daly**[2,4], **David R. Beier**[7,8,†], and **Shamil R. Sunyaev**[1,2,†]

[1]Brigham and Women's Hospital, Division of Genetics, Harvard Medical School, Boston, MA, USA

[2]Program in Medical and Population Genetics, Broad Institute, Cambridge, MA, USA

[3]Department of Genetic and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA

[4]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

[5]Program in Biological and Biomedical Sciences, Harvard Medical School, Boston, MA, USA

[6]Division of Genetics and Genomics, Boston Children's Hospital, Boston, MA, USA

[7]Center for Developmental Biology and Regenerative Medicine, Seattle Children's Research Institute, Seattle, WA, USA

[8]Dept. of Pediatrics, University of Washington School of Medicine, Seattle, WA, USA

Abstract

The dispensability of individual genes for viability has interested generations of geneticists. For some genes it is essential to maintain two functional chromosomal copies, while others may tolerate the loss of one or both copies. Exome sequence data from 60,706 individuals provide sufficient observations of rare protein truncating variants (PTVs) to make genome-wide estimates of selection against heterozygous loss of gene function. The cumulative frequency of rare deleterious PTVs is primarily determined by the balance between incoming mutations and purifying selection rather than genetic drift. This enables the estimation of the genome-wide

[†]Co-corresponding authors: Shamil Sunyaev, ssunyaev@rics.bwh.harvard.edu; David Beier, David.Beier@seattlechildrens.org.
[*]These authors contributed equally to this work.
Present Addresses: Shamil Sunyaev, 77 Avenue Louis Pasteur Room 466B, Boston, MA 02115
David Beier, 1900 Ninth Ave., Seattle, WA 98101 USA

Author Contributions
Overall concept and approach conceived and developed by C.C., D.B., and S.S. Implementation, data analysis and interpretation conducted by C.C., D.W., D.J.B., D.J., and D.N. Datasets and helpful advice were provided by D.M., M.D., K.S., and A.O. The article was written by C.C. and S.S. with contributions from D.W. and D.J.B. All authors read and discussed the manuscript.

Competing Financial Interests Statement
The authors have no competing interests as defined by Springer Nature, or other interests that might be perceived to influence the results and/or discussion reported in this paper.

distribution of selection coefficients for heterozygous PTVs and corresponding Bayesian estimates for individual genes. The strength of selection can discriminate the severity, age of onset, and mode of inheritance in Mendelian exome sequencing cases. We find that genes under the strongest selection are enriched in embryonic lethal mouse knockouts, putatively cell-essential genes, Mendelian disease genes, and regulators of transcription. Screening by essentiality, we find a large set of genes under strong selection that likely have critical function but have not yet been extensively annotated in published literature.

The evolutionary cost of gene loss is a central question in genetics and has been investigated in model organisms and human cell lines[1–3]. In humans, the question of dispensability and haploinsufficiency of individual genes is intimately related to their causal role in genetic disease. However, estimates of the selection and dominance coefficients in humans have proved elusive as inference techniques used in other sexual organisms generally require cross-breeding over several generations.

The analysis of patterns of natural genetic variation in humans provides an alternative approach to estimating selection intensity and dispensability of individual genes. Despite substantial methodological progress in the ascertainment and analysis of population sequence data[4–8], estimation of parameters of natural selection in humans has been complicated by genetic drift, complexities of human demographic history[4,5,7,9–12] and the role of non-additive genetic variation[13–15]. Additionally, naturally occurring PTVs are infrequent in the population, so datasets of thousands of individuals are underpowered for the estimation of gene dispensability in humans.

The Exome Aggregation Consortium (ExAC) dataset now provides a sufficiently powered sample to assess the selection that constrains the number of gene-specific PTVs in the general population[16]. We restrict our analysis to PTVs predicted to be consequential[17], which allows the assumption that all PTVs within a gene likely incur the same selective disadvantage. We can then treat each gene as a bi-allelic locus with a functional state and a loss-of-function state. In each gene, the cumulative frequency of rare deleterious PTVs (the sum of PTV allele frequencies throughout the gene) is then primarily determined by the balance between incoming mutations and selection rather than reassortment of alleles by stochastic drift. This makes our estimates robust to drift, population structure and historical changes in population size, which we evaluate analytically and with simulations (**Methods and** Supplementary Figure 1).

Using population frequency data from 60,706 individuals without severe Mendelian disorders, we estimate both the overall distribution of gene-based fitness effects and individual gene fitness cost in heterozygotes. Given gene-specific estimates of the *de novo* mutation rate[18,19], the observed number of PTV alleles throughout each gene, and number of chromosomes sampled, we estimate the genome-wide distribution of selective effects for heterozygous PTVs, s_{het}. We parameterize the distribution of selective effects using an inverse Gaussian, which is fit using maximum likelihood (Figure 1). We then estimate individual gene selection coefficients using the posterior probability for s_{het} given gene-specific values of the observed number of PTVs, number of chromosomes sampled and estimated mutation rate (Supplementary Table 1).

Although the distribution is broad, suggesting the effect of losing one copy of a gene is variable, the mode of the distribution corresponds to a fitness loss around 0.5% (s_{het} = 0.005). Despite the large sample size, resolution to distinguish between very high selective effects is limited. There are 2,984 genes with $s_{\text{het}} > 0.1$, a result concordant with previous estimates of loss of function intolerance derived from population data[16]. Even though some genes are heavily depleted of PTVs in ExAC as compared with mutational expectation, these values suggest that heterozygote PTVs in many genes are not necessarily responsible for observable, severe clinical consequences.

Unsurprisingly however, Mendelian diseases genes have higher s_{het} values. Among them, genes annotated exclusively as autosomal dominant (AD, N=867) have significantly higher s_{het} values than those annotated as autosomal recessive (AR, N=1,482)[20] [Mann-Whitney p-value 3.14×10^{-64}] (Figure 2[a,b]). This suggests it may be possible to prioritize candidate disease genes identified in clinical exome sequencing analysis using the observed mode of inheritance and s_{het} value.

In 504 clinical exome cases that resulted in Mendelian diagnosis[21], we find a similar enrichment of cases by MOI and selection value (Figure 2[c]). We find that 90.4% of novel, dominant variants are associated with heterozygous fitness loss greater than 0.04 (Figure 2[d]). Among disease variants, a cutoff of $s_{\text{het}} > 0.04$ provides a 96% positive predictive value for discriminating between AD and AR.

To test the generalizable utility of prioritizing candidate genes in Mendelian sequencing studies using s_{het}, we compared the overall prevalence of genes with $s_{\text{het}} > 0.04$ to the corresponding fraction in an independently ascertained dataset of new dominant Mendelian diagnoses (Figure 2[e])[22]. This analysis suggests that restricting to genes with $s_{\text{het}} > 0.04$ would provide a three-fold reduction of candidate variants, given the overall distribution of s_{het} values. Thus, initial effort in clinical cases can be focused on just a few genes for functional validation, familial segregation studies, and patient matching. We summarize the classification accuracy (AUC 0.9312) and generate mode of inheritance probabilities for each gene using the full set of clinical sequencing cases (Supplementary Figure 2 and Supplementary Table 2).

Beyond mode of inheritance, we find that s_{het} helps predict phenotypic severity, age of onset, penetrance, and the fraction of *de novo* variants in a set of high-confidence haploinsufficient disease genes (Figure 3). In broader sets of known disease genes, s_{het} estimates significantly correlate with the number of references in OMIM MorbidMap and the number of HGMD disease “DM” variants (Supplementary Figure 3).

Gene-specific fitness loss values allow us to plot the distribution of selective effects for different disorders. This provides information about the breadth and severity of selection associated with various disorder groups using both well-established genes (Figure 4[a]) and findings from Mendelian exome cases (Figure 4[b]). Overall, genes involved in neurologic phenotypes and congenital heart disease appear to be under more intense selection compared with other disorder groups, or tolerated knockouts from a consanguineous cohort (Figure 4[c,d])[23]. Interestingly, genes recessive for these disorders appear to have only partially recessive effects on fitness, so selection on heterozygotes is not negligible in these genes (Figure 4).

In germline cancer predisposition, genes under stronger selection are enriched in individuals with cancer over those in ExAC (Supplementary Figure 4). This suggests that genes with low s_{het} values should not be prioritized in prospective genetic screening for cancer predisposition. Consistent with previous studies[18], we find *de novo* mutations in patients with autism spectrum disorder are significantly enriched in genes under stronger selection than those identified in controls (Supplementary Figure 5 and Supplementary Table 3).

Next, we analyze s_{het} in the context of developmental and functional assays. In a large set of neutrally-ascertained mouse knockouts (N=2,179)[24], mice that are null mutant for orthologous genes with higher s_{het} estimates are enriched for embryonic lethality or sub-viability, while those with the lowest s_{het} estimates are depleted for embryonic lethality [Mann-Whitney $p=2.95\times10^{-28}$] (Figure 5[a,b]).

It is well known that mutations that are haploinsufficient in humans can often be well-tolerated when heterozygous in mice[25]. A classic example is *SHH*; heterozygous null mutations in this important developmental signaling gene result in holoprosencephaly[26]. Haploinsufficiency for other genes in this signaling pathway also results in developmental defects; e.g. *GLI3* (Pallister-Hall syndrome and Greig cephalopolysyndactyly syndrome)[27–29] and *GLI2* (Holoprosencephaly 9)[30]. Interestingly, haploinsufficiency for these genes is tolerated in mouse models; mice heterozygous for null variation in the *SHH* signaling pathway are phenotypically normal, while homozygous mutant mice have defects that recapitulate features of the human syndrome[31–33]. This extends to many other human developmental disorders, enabling the experimental characterization of the molecular consequences of these mutations. Thus, it is notable that homozygous null mice in orthologous genes with higher s_{het} values are enriched for lethality.

High-throughput genetic analysis of cell-essentiality provides an orthogonal dataset for comparison with s_{het}. In genes putatively essential for human cell proliferation using CRISPR-based inactivation (Figure 5[c]) and gene trap inactivation assays[3] (Figure 5[d]), we find that essential genes are heavily enriched with high s_{het} values [p-values 5.13×10^{-16}, 4.90×10^{-18}, respectively].

Key developmental pathways are dramatically enriched in genes under strong selection (Figure 6[a]). We also find a significant positive correlation between the number of protein-protein interactions for each gene and its s_{het} value (Figure 6[b,c]), identified from high-throughput mass spectrometry data. In the context of molecular and cellular function, a set of genes with very high selective effects ($s_{\text{het}} > 0.15$, 2,072 genes) is enriched in biological process categories "transcription regulation" (Bonferroni $p=1.8\times10^{-39}$), "transcription" (7.5×10^{-36}), and "negative regulators of biosynthetic processes" (Supplementary Material)[34]. Nucleus was the most enriched cellular compartment for these genes (4.8×10^{-76}). The enrichment of transcription factors in these genes is consistent with literature that describes dosage dependence for enzymatic proteins and haploinsufficiency for transcriptional regulators[35].

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

We inspect genes that lack disease annotations and publications but that have high s_{het} values to determine whether they share functional and genetic features reminiscent of known genes with central roles in cell housekeeping and developmental biology. We measure the relative knowledge about each gene in the primary literature from Entrez and PubMed[36] using the number of gene reports connected with each manuscript, and sum the weighted contributions across all available manuscripts[37] (PubMed score, **Methods**). While the PubMed score is positively correlated with s_{het} values, a substantial number of understudied genes fall in the highest s_{het} decile (Supplementary Figure 6).

We selected the 250 most cited and least cited genes within the top s_{het} decile, and compared their frequency of protein-protein interactions, viability of orthologous mouse knockouts and cell essentiality assays. Genes with the fewest publications (no more than one individual citation) have nearly the same number of embryonic lethal mouse knockouts as genes with the most publications. Other assays are only slightly depleted in genes with the fewest publications (Supplementary Figure 7). These findings suggest there may be additional essential developmental pathways yet to be uncovered in genes under strong selection that lack functional or disease annotations, and provides a promising gene set for further exploration. We have created a prioritized list of genes using developed from functional evidence to indicate the most promising candidates for future functional screening (Supplementary Table 4).

To place our inferences in the broader evolutionary context, we use comparable estimates from model organisms including flies and yeast, based on knockout competition with wild type or explicit crosses. In yeast, the analysis of a library of PTV knockouts provides a mean estimate of $s_{\text{het}} \approx 0.013$, which is close to our inferred results ($s_{\text{het}} \approx 0.059$) in humans[38], given that the functional experiments excluded genes with very high s, and we have excluded genes with high cumulative allele frequency. Estimates in flies derived from homozygote lethal mutations which reduce viability in heterozygotes (rather than only PTVs) suggest values of s_{het} on the order of 1–3%, which is also in broad agreement with our estimates in humans[1,39]. While values of s in this range have a small impact in each generation, they may have dramatic evolutionary consequences[40].

In conclusion, we use the genome-wide distribution of PTVs to estimate fitness loss due to heterozygous loss of each gene. Unlike recent work on genic intolerance[18,41], we explicitly estimate the distribution of selection coefficients for PTVs. Our estimates are also distinct from earlier work on the estimation of fitness effects of allelic variants in humans[42] as the large sample size coupled with the assumption of strong selection makes our approach robust with respect to complexities of demographic history and dominance, and allows gene-based inferences. Conversely, our assumptions are justified for many but not all genes, as the method has limited resolution for genes under the strongest and weakest selection. These results may be useful in Mendelian disease gene discovery efforts and provide clinical utility in the inference of severity and mode of inheritance underlying Mendelian disease.

Data Availability

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

Online Methods

Model of deterministic mutation-selection balance

For most genes, protein-truncating alleles are both individually and collectively rare. For genes where they are collectively rare, estimation of the selective effect against heterozygous PTVs (s_{het}) can be greatly simplified. We model each gene as a single bi-allelic locus with cumulative frequency $X = \Sigma_j x_j$, where the sum is over PTVs in gene i for PTV sites j. This is motivated by the simplifying assumption of identical selection coefficients for all PTVs within a gene, and the observation that the frequency of the vast majority of PTVs is extremely low such that the occurrence of multiple variable sites within a gene on a single haplotype is also extremely low ($2Nx_{ij}x_{ik} < 1$ for sample size N). Moreover, multiple PTVs in a gene in an individual would be functionally equivalent to a single PTV resulting in a loss of function state.

Then for each gene, the cumulative allele frequency X is influenced by incoming mutation, selection and the random reassortment of alleles (genetic drift). When selection is strong, $s \gg 2.5\times10^{-5}$ (i.e. when $4N_e s \gg 1$, with effective population size N_e 10^4), drift is much smaller than the contribution of selection. Furthermore, the strength of genetic drift is weakest for genes at low frequencies: for a variant with cumulative frequency of $X = 0.001$ the expected frequency change due to drift is only $\langle \Delta X^2 \rangle \sim X/4N_e = 2.5\times10^{-8}$ per generation. Notably, at the locus level assuming $X \ll 1$ the drift contribution is also much smaller than the mutational influx. Hence under strong selection and for small allele frequencies the expected cumulative frequency of PTVs is determined by the equilibrium between the influx of *de novo* mutations (estimated to increase the cumulative frequency by an average 1.4×10^{-6} per locus per generation by mutational model) and the outflux due to natural selection.

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

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

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

$$\partial_t X \approx -s_{\text{het}} X - s_{\text{hom}} X^2 + U \qquad (2)$$

Because $X \ll 1$, selection against heterozygotes (the linear term) generally also dominates over selection against homozygotes (the quadratic term), provided $s_{\text{het}/} s_{\text{hom}} \gg$ X. This is only violated in cases of extreme recessivity (where the dominance coefficient $h \ll 0.001$), but even in that case the expected cumulative frequency of PTVs in essential genes is unlikely to exceed 0.001 (the characteristic X in the completely recessive case is $\sqrt{U/s} \sim 10^{-3}$ when $s \sim 1$, see simulations in Supplementary Figure 1). The strong selection regime thus corresponds to mutation-selection balance in the heterozygote state of a PTV mutation. In our model, we do not assume that selection acts exclusively on heterozygotes, but aim at estimating only fitness loss due to the lack of one functional copy of a gene. Even in the case of strong selection against homozygotes, the population frequency is primarily controlled by efficient selection against heterozygotes.

Notably, there is no dependence on the demography or population size in this regime, as the contribution from drift vanishes because selection drives alleles out of the population efficiently and on very short time scales. Classic papers by Li[43,44] and Maruyama[45,46] showed that relevant time scales are short, even in the case of exponential expansion, because individual deleterious alleles are predominantly recent, with an allelic age on the order of 1/s. Current estimates of recent population histories for most of populations included in the ExAC dataset suggest that $4Ns$ safely exceeds 1. Even if individual alleles are subject to stochastic drift, this effect is mitigated by the aggregation of variants on the gene level. One possible concern is the inclusion of individuals with Finnish ancestry, as this population underwent an intense, relatively recent bottleneck. We address this population explicitly using forward simulations and by removing them from our analysis to show no significant deviation from our initial estimates in their absence (below).

From Eq. 2 follows that for a population sample of size N chromosomes, sample allele counts $n = N\hat{X} = N\Sigma_j \hat{x}_j$ are expected to be Poisson distributed around the expectation given by:

$$E(n) \equiv \frac{NU}{s_{\text{het}}} \qquad (3)$$

Generally, genes under the strongest and weakest selection are expected to have greater estimation uncertainty, as the resolution to estimate s_{het} deteriorates when variants are so common that they may not only be controlled by heterozygote selection, but also by drift or complex demography. However, the overwhelming majority of genes conform to our assumptions of cumulative PTV allele frequency not exceeding 0.001. Despite issues such as the admixture of populations, consanguineous samples in ExAC[23], and the Wahlund effect, very few genes (1,201 of 17,199 covered genes) have higher estimated cumulative allele frequencies $\hat{X}$, which we restrict from the estimation procedure. On the other end of the spectrum, genes under strong selection may lack PTVs by chance alone in ExAC, which limits the ability to distinguish between large selective effects.

Population genetics simulations of model assumptions

To validate the assumption that estimates of selection can be made under mutation-selection balance independent of demography or population size for variants under sufficiently strong selection, we used SLiM 2.0 to conduct forward population genetics simulations[47]. We ran 10,000 replicates each of simulations with selection coefficients of -5×10^{-1}, -5×10^{-2}, -5×10^{-3}, -5×10^{-4}, and -5×10^{-5} through a realistic demography derived from previously published histories for African, Non-Finnish European, and Finnish populations[48,49] (Supplementary Figure 1). We compare the theoretical expectation of cumulative allele frequency (U/s_{het} [Equation 3]) with the simulated cumulative allele frequency. We do this in three populations (African, Non-Finnish European and Finnish), plus a “Combined” population which includes pooled site frequency spectra from all three populations in proportions represented in the ExAC dataset. The simulations support our assumption of mutation-selection balance in the strong selection regime ($|s_{het}| >= 1\times10^{-3}$), which appears to be appropriate for PTVs. This is true for all three populations examined and for the combined population, demonstrating that this assumption is robust to differences in the strength of drift due the distinct demographic histories of included human populations.

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

Dataset for s_{het} estimation

In this analysis, we use Exome Aggregation Consortium (ExAC) dataset version 0.3, a set of jointly-called exomes from 60,706 individuals ascertained with no known severe, early-onset Mendelian disorders. The mean coverage depth was calculated for each gene (canonical transcript from Ensembl v75, GENCODE v19) in the ExAC dataset (mean 57.75; s.d. 20.96). Genes with average coverage depth of at least 30x were used in further analysis (N=17,199). Single nucleotide substitution variants annotated as PASS quality with predicted functional effects in the canonical transcript of “stop_gained”, “splice_donor”, or “splice_acceptor” (as annotated by Variant Effect Predictor) were included in the analysis. Variants such as indels, in-frame mutations, and frameshift variants were excluded from this analysis, as many of these variants may have annotation issues or may not be functionally impactful. Along the same lines, we are mindful that not all PTVs will result in complete loss of gene function, due to alternative transcripts or nonsense mediated decay. To address this, variants were filtered using LOFTEE[50] and restricted to those predicted with high confidence to have consequences in the canonical transcript.

For each of the 17,199 genes we have observable values for (n, U, N), where n denotes the total number of observed PTV alleles in the population sample of N chromosomes covered in the gene, and U the PTV mutation rate across the canonical gene transcript from a mutational model[18,19]. Values of U for each gene from Samocha *et al.* were used along with the number of well-covered chromosomes N in each gene to generate the null mutational expectation of neutral evolution, NU. Incorrectly specified values from this mutational model could alter estimates of selection for individual genes, as higher estimates of selection are made in genes with greater depletions from the null expectation model. Our inference of selection coefficients relies on the assumption that the cumulative population frequency of PTV mutations, X, is small due to strong negative selection, so genes with $\hat{X} = n/N > 0.001$ are omitted from the analysis, leaving 15,998 genes.

Estimation of $P(s_{\mathrm{het}})$

A genome-wide ensemble of observed (n) and expected ($NU \equiv \nu$) genic PTV counts enables the inference of the distribution of heterozygous loss-of-function fitness effects, $P(s_{\mathrm{het}})$, which underlies the evolutionary dynamics of this class of mutations. We estimate the parameters (α,β) of this distribution by fitting the observed distribution of PTV counts across genes:

$$P(n|\alpha, \beta;\nu)=\int P(n|s_{\mathrm{het}};\nu)\, P(s_{\mathrm{het}};\alpha, \beta)\, \mathrm{d}s_{\mathrm{het}}. \qquad (4)$$

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

$$\begin{aligned}P(n|\alpha, \beta;\nu)&=\int \mathrm{Poiss}(n;\lambda=\nu/s_{\mathrm{het}})\, \mathrm{IG}(s_{\mathrm{het}};\alpha, \beta)\, \mathrm{d}s_{\mathrm{het}}\\&=\tfrac{1}{n!}e^{\frac{\beta}{\alpha}}\sqrt{\tfrac{2\beta}{\pi\alpha}}\left(\tfrac{\nu}{\alpha}\right)^{n}\left(\tfrac{\beta}{\beta+2\nu}\right)^{\frac{1+2n}{4}}K_{\frac{1}{2}+n}\left(\tfrac{\sqrt{\beta(\beta+2\nu)}}{\alpha}\right),\end{aligned} \qquad (5)$$

where $K_n(z)$ is the modified Bessel function of the second kind. To estimate parameters of the distribution of selection coefficients, $P(s_{\mathrm{het}};\alpha,\beta)$, we fit Eq. 5 to the observed distribution of PTV counts, $Q(n)$ by maximizing the log-likelihood

$$\log[\mathscr{L}(\alpha, \beta|\{n\})]=\log\sum_{i=1}^{G} P(n_i|\alpha, \beta;\nu_i) \qquad (6)$$

on the regime $\alpha \in [10^{-2}, 2]$ and $\beta \in [10^{-4}, 2]$, where G is the number of genes. In order to account for a slight positive correlation between the mutation rate and selection strength (Supplementary Figure 8), we separately perform the fit on U terciles of the data set and combine the results in a mixture distribution with equal weights. The mean mutation rates in the three terciles are $\bar{U}_1 = 4.6 \cdot 10^{-7}$, $\bar{U}_2 = 1.1 \cdot 10^{-6}$, and $\bar{U}_2 = 2.6 \cdot 10^{-6}$. We estimate ($\hat{\alpha}_1$, $\hat{\beta}_1$) = (0.057±0.010,0.0052±0.0003), ($\hat{\alpha}_2$, $\hat{\beta}_2$) = (0.046±0.005,0.0087±0.0004), and ($\hat{\alpha}_2$, $\hat{\beta}_2$) = (0.074±0.005,0.0160±0.0005), with error margins denoting two s.d. from 100 bootstrapping replicates of the set of ~5,333 genes in each tercile. This error estimate is intended to quantify the effect of the sampling noise in the data set on the parameter inference while local mutation rate estimates are assumed fixed. The resulting fitted distributions of counts are shown in Supplementary Figure 9 together with the corresponding $Q(n)$, while Figure 1 shows the inferred $P(s_{\text{het}}; \hat{\alpha}, \hat{\beta})$ = (IG(s_{het}; $\hat{\alpha}_1$, $\hat{\beta}_1$) + (IG(s_{het}; $\hat{\alpha}_2$, $\hat{\beta}_2$) + (IG(s_{het}; $\hat{\alpha}_3$, $\hat{\beta}_3$))/3. The choice for the functional form of $P(s_{\text{het}})$ is motivated by the shape of the empirical distribution of the naïve estimator ν/n (given by a simple inversion of Eq. 3). We also compared the log-likelihood of the fit to $Q(n)$ obtained with this model to that obtained from two other two-parameter distributions, s_{het} ~ Gamma and s_{het} ~ InvGamma, and chose the model with the highest likelihood, which is s_{het} ~ IG.

To assess the relative change in the distribution of heterozygote selection coefficients when different population subsets are included, we first estimated the distribution of s_{het} using only non-Finnish Europeans (NFE) in Supplementary Figure 10. We find high concordance between the overall distribution generated using all ExAC samples and NFE specific estimates. We also separately removed Finnish individuals from the estimation of the distribution of selection coefficients, and find very high concordance between estimates made using all ExAC samples and ExAC without Finnish individuals (Supplementary Figure 11). These analyses demonstrate that the model is robust to concerns about recent demographic history in Finnish individuals, supporting the validity of the deterministic approximation. We cannot completely rule out the possibility that other included populations may have issues related to complexities of their recent demographic history.

Inference of s_{het} on individual genes

From the inferred distributions $P(s_{\text{het}}; \hat{\alpha}_t, \hat{\beta}_t)$ in each tercile t of the mutation rate U, we construct a per-gene estimator of s_{het} for genes in the tercile using the posterior probability given n, which mitigates the stochasticity of the observed PTV count:

$$P\left(s_{\text{het},i} \middle| n_i;\nu_i\right) = \frac{P\left(n_i | s_{\text{het},i};\nu_i\right) P\left(s_{\text{het},i};\hat{\alpha}_t,\hat{\beta}_t\right)}{\int P\left(n_i|s;\nu_i\right) P\left(s;\hat{\alpha}_t,\hat{\beta}_t\right)\,\mathrm{d}s}, \qquad (7)$$

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

Predicted mode of inheritance in clinical exome cases

We trained a Naïve Bayes classifier to predict the mode of inheritance in a set of solved clinical exome sequencing cases from Baylor College of Medicine (N=283 cases)[21] and UCLA[22] (N=176 cases). Using data from UCLA as the training dataset, we are able to cross-predict the mode of inheritance in separately ascertained Baylor cases with classification accuracy of 88.0%, sensitivity of 86.1%, specificity of 90.2%, and an AUC of

0.931. Genes that were related to diagnosis in both clinics (overlapping genes) were removed from the larger Baylor set (Supplementary Figure 2).

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

Mouse knockout comparative analysis

We reviewed mouse knockout enrichments from two datasets: the full set of mouse knockouts from a neutrally-ascertained mouse knockout screen (N=2,179 genes) generated by the International Mouse Phenotyping Consortium[24]. Genes were classified as 'Viable', 'Sub-Viable', or 'Lethal' based on the results for the assay.

PubMed gene score and enrichment analysis

We developed a score to estimate the relative importance of each gene in the published medical and scientific literature. First, we connected literature from Entrez which included both PubMed citations and references to Entrez genes. We assigned a weight to each article referencing a gene of $1/a_i$, where a_i was the number of genes referred to by article *i*. For example, an article referring to four genes would receive a weight of 1/4. Finally, we assigned each gene a score which was the sum of the weighted article scores. These scores ranged from 4,672 articles per gene (p53) to 0.0001 articles/gene.

Next, we focused on genes that are estimated to be under very strong selection but that lack functional or clinical annotations. In the top decile of s_{het} values, we separated the top 250 and bottom 250 genes by PubMed score. We then annotated each of these with unbiased genome-wide assays, including the number of protein-protein interactions (as determined by a genome-wide mass spectrometry assay)[51], whether each gene is determined to be cell-essential in genome-wide CRISPR and gene trap assays[3], and whether there is a mouse knockout in the neutrally-ascertained orthologous nonviable mouse knockout[52]. To limit the number of genes with incorrect s_{het} estimates in this set of 500 genes, we pre-filtered any genes with only a single exon, as they may be enriched for recent pseudogenes, and also removed any olfactory, mucin, and zinc finger proteins.

Functional enrichment analysis

We inspected the functional annotations related to approximately the top 10% of selectively disadvantageous genes (with $s_{\text{het}} > 0.15$, N=2,072 genes) that were successfully mapped using Database for Annotation, Visualization, and Integrated Discovery (DAVID) version 6.7[34], DAVID. Separately, two other cutoffs ($s_{\text{het}} > 0.25$, N=897 genes and $s_{\text{het}} > 0.5$, N=32 genes) were also tested and similar results were identified.

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

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

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by National Institutes of Health Grant HG007229 (C.C.), GM078598 (S.S., D.J., D.J.B.), and MH101244 (S.S., D.W.). We thank I. Adzhubei, K. Karczewski, and A. Kondrashov for helpful advice.

References

1. Mukai T, Chigusa SI, Mettler LE, Crow JF. Mutation rate and dominance of genes affecting viability in Drosophila melanogaster. Genetics. 1972; 72:335–55. [PubMed: 4630587]
2. Deng HW, Lynch M. Estimation of deleterious-mutation parameters in natural populations. Genetics. 1996; 144:349–360. [PubMed: 8878698]
3. Wang T, et al. Identification and characterization of essential genes in the human genome. Science (80-). 2015; 350:1096–1101.
4. Williamson SH, et al. Simultaneous inference of selection and population growth from patterns of variation in the human genome. Proc Natl Acad Sci U S A. 2005; 102:7882–7. [PubMed: 15905331]
5. Boyko AR, et al. Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet. 2008; 4:e1000083. [PubMed: 18516229]
6. Kryukov GV, Pennacchio LA, Sunyaev SR. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. Am J Hum Genet. 2007; 80:727–39. [PubMed: 17357078]
7. Kryukov GV, Shpunt A, Stamatoyannopoulos JA, Sunyaev SR. Power of deep, all-exon resequencing for discovery of human trait genes. Proc Natl Acad Sci U S A. 2009; 106:3871–3876. [PubMed: 19202052]
8. Eyre-Walker A, Keightley PD. The distribution of fitness effects of new mutations. Nat Rev Genet. 2007; 8:610–8. [PubMed: 17637733]
9. Do R, et al. No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. Nat Genet. 2015; 47:126–131. [PubMed: 25581429]
10. Fu W, Gittelman RM, Bamshad MJ, Akey JM. Characteristics of neutral and deleterious protein-coding variation among individuals and populations. Am J Hum Genet. 2014; 95:421–36. [PubMed: 25279984]
11. Lohmueller KE. The distribution of deleterious genetic variation in human populations. Curr Opin Genet Dev. 2014; 29:139–46. [PubMed: 25461617]
12. Gravel S. When Is Selection Effective? Genetics. 2016; 203:451–62. [PubMed: 27010021]
13. Williamson S, Fledel-Alon A, Bustamante CD. Population genetics of polymorphism and divergence for diploid selection models with arbitrary dominance. Genetics. 2004; 168:463–75. [PubMed: 15454557]
14. Balick DJ, Do R, Cassa CA, Reich D, Sunyaev SR. Dominance of Deleterious Alleles Controls the Response to a Population Bottleneck. PLoS Genet. 2015; 11:e1005436. [PubMed: 26317225]

15. Simons YB, Turchin MC, Pritchard JK, Sella G. The deleterious mutation load is insensitive to recent population history. Nat Genet. 2014; 46:220–224. [PubMed: 24509481]
16. Lek M, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016; 536:285–291. [PubMed: 27535533]
17. MacArthur DG, et al. A systematic survey of loss-of-function variants in human protein-coding genes. Science (80-). 2012; 335:823–828.
18. Samocha KE, et al. A framework for the interpretation of de novo mutation in human disease. Nat Genet. 2014; 46:944–50. [PubMed: 25086666]
19. Francioli LC, et al. Genome-wide patterns and properties of de novo mutations in humans. Nat Genet. 2015; 47:822–6. [PubMed: 25985141]
20. Solomon BD, Nguyen AD, Bear KA, Wolfsberg TG. Clinical genomic database. Proc Natl Acad Sci U S A. 2013; 110:9851–5. [PubMed: 23696674]
21. Yang Y, et al. Molecular Findings Among Patients Referred for Clinical Whole-Exome Sequencing. JAMA. 2014; doi: 10.1001/jama.2014.14601
22. Lee H, et al. Clinical Exome Sequencing for Genetic Identification of Rare Mendelian Disorders. JAMA. 2014; doi: 10.1001/jama.2014.14604
23. Saleheen D, et al. Human knockouts in a cohort with a high rate of consanguinity. 2015; doi: 10.1101/031518
24. Koscielny G, et al. The International Mouse Phenotyping Consortium Web Portal, a unified point of access for knockout mice and related phenotyping data. Nucleic Acids Res. 2014; 42:D802–9. [PubMed: 24194600]
25. Georgi B, Voight BF, Bućan M. From Mouse to Human: Evolutionary Genomics Analysis of Human Orthologs of Essential Genes. PLoS Genet. 2013; 9:e1003484. [PubMed: 23675308]
26. Roessler E, et al. Mutations in the human Sonic Hedgehog gene cause holoprosencephaly. Nat Genet. 1996; 14:357–60. [PubMed: 8896572]
27. Kang S, Graham JM, Olney AH, Biesecker LG. GLI3 frameshift mutations cause autosomal dominant Pallister-Hall syndrome. Nat Genet. 1997; 15:266–8. [PubMed: 9054938]
28. Vortkamp A, Gessler M, Grzeschik KH. GLI3 zinc-finger gene interrupted by translocations in Greig syndrome families. Nature. 1991; 352:539–40. [PubMed: 1650914]
29. Wild A, et al. Point mutations in human GLI3 cause Greig syndrome. Hum Mol Genet. 1997; 6:1979–84. [PubMed: 9302279]
30. Roessler E, et al. Loss-of-function mutations in the human GLI2 gene are associated with pituitary anomalies and holoprosencephaly-like features. Proc Natl Acad Sci U S A. 2003; 100:13424–9. [PubMed: 14581620]
31. Chiang C, et al. Cyclopia and defective axial patterning in mice lacking Sonic hedgehog gene function. Nature. 1996; 383:407–13. [PubMed: 8837770]
32. Hui CC, Joyner AL. A mouse model of greig cephalopolysyndactyly syndrome: the extra-toesJ mutation contains an intragenic deletion of the Gli3 gene. Nat Genet. 1993; 3:241–6. [PubMed: 8387379]
33. Mo R, et al. Specific and redundant functions of Gli2 and Gli3 zinc finger genes in skeletal patterning and development. Development. 1997; 124:113–23. [PubMed: 9006072]
34. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2008; 4:44–57.
35. Seidman JG, Seidman C. Transcription factor haploinsufficiency: when half a loaf is not enough. J Clin Invest. 2002; 109:451–455. [PubMed: 11854316]
36. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2013; 41:D8–D20. [PubMed: 23193264]
37. Raychaudhuri S, et al. Identifying relationships among genomic disease regions: Predicting genes at pathogenic SNP associations and rare deletions. PLoS Genet. 2009; 5
38. Agrawal AF, Whitlock MC. Inferences about the distribution of dominance drawn from yeast gene knockout data. Genetics. 2011; 187:553–566. [PubMed: 21098719]
39. Simmons MJ, Crow JF. Mutations Affecting Fitness in Drosophila Populations. Annu Rev Genet. 1977; 11:49–78. [PubMed: 413473]

40. Wright S. Evolution in Mendelian populations. 1931. Bull Math Biol. 1990; 52 241-95–7.

41. Petrovski S, et al. The Intolerance of Regulatory Sequence to Genetic Variation Predicts Gene Dosage Sensitivity. PLoS Genet. 2015; 11:e1005492. [PubMed: 26332131]

42. Kiezun A, et al. Exome sequencing and the genetic basis of complex traits. Nat Genet. 2012; 44:623–30. [PubMed: 22641211]

43. Li WH, Nei M. Total number of individuals affected by a single deleterious mutation in a finite population. Am J Hum Genet. 1972; 24:667–79. [PubMed: 5082917]

44. Li WH. The first arrival time and mean age of a deleterious mutant gene in a finite population. Am J Hum Genet. 1975; 27:274–86. [PubMed: 803010]

45. Maruyama T. The age of a rare mutant gene in a large population. Am J Hum Genet. 1974; 26:669–73. [PubMed: 4440678]

46. Maruyama T. The age of an allele in a finite population. Genet Res. 1974; 23:137–43. [PubMed: 4417585]

47. Messer PW. SLiM: simulating evolution with selection and linkage. Genetics. 2013; 194:1037–9. [PubMed: 23709637]

48. Tennessen JA, et al. Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. Science (80-). 2012; 337:64–69.

49. Wang SR, et al. Simulation of Finnish population history, guided by empirical genetic data, to assess power of rare-variant tests in Finland. Am J Hum Genet. 2014; 94:710–20. [PubMed: 24768551]

50. Karczewski, K. LOFTEE (Loss-Of-Function Transcript Effect Estimator). 2015. at <https://github.com/konradjk/loftee>

51. Huttlin EL, et al. The BioPlex Network: A Systematic Exploration of the Human Interactome. Cell. 2015; 162:425–40. [PubMed: 26186194]

52. Ayadi A, et al. Mouse large-scale phenotyping initiatives: overview of the European Mouse Disease Clinic (EUMODIC) and of the Wellcome Trust Sanger Institute Mouse Genetics Project. Mamm Genome. 2012; 23:600–10. [PubMed: 22961258]

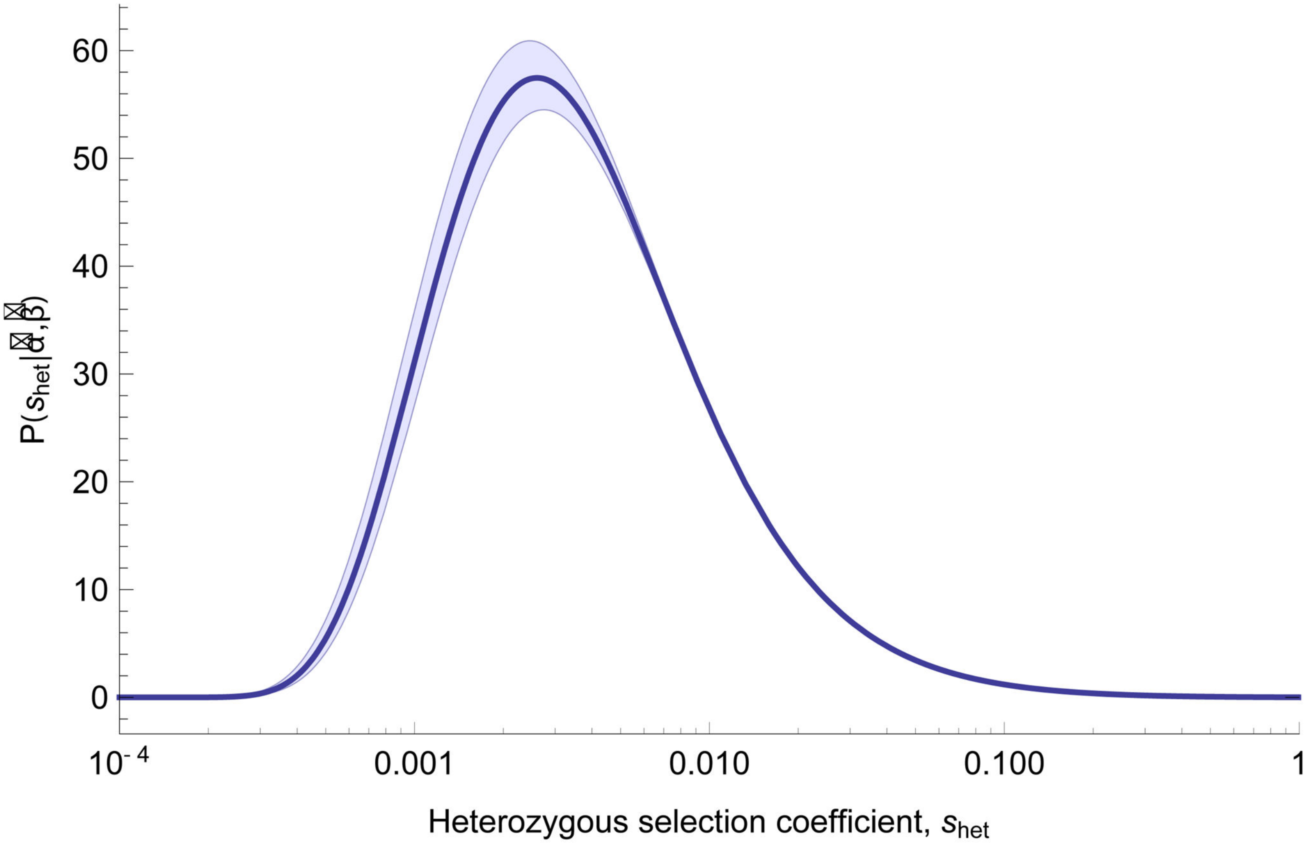

Figure 1.

Inferred distribution of fitness effects for heterozygous loss of gene function. Estimates of parameters ($\hat{\alpha}$, $\hat{\beta}$) from maximum likelihood fit to the observed distribution of PTV counts across 15,998 genes in terciles of mutation rate, assuming $s_{het} \sim \mathrm{IG}(\alpha, \beta)$. Shaded areas show 95% CI obtained from 100 bootstrapping replicates, intended to quantify the influence of sampling noise in the data set on parameter inference, with fixed estimates of local mutation rate.

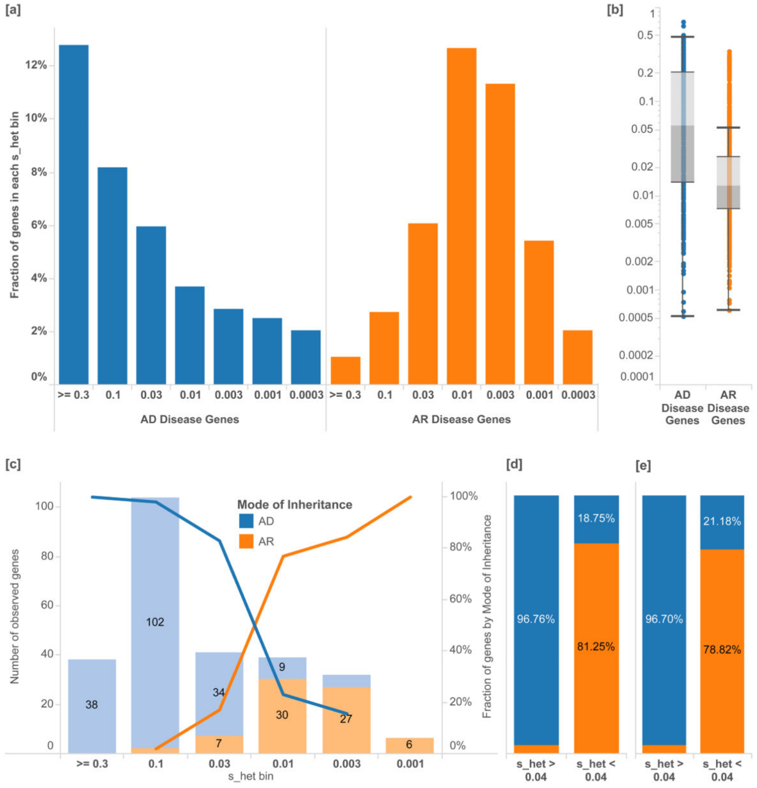

Figure 2.

Separation of disease genes and clinical cases by mode of inheritance. [a] The percentage of genes associated with exclusively autosomal dominant (AD, N=867) disorders versus autosomal recessive (AR, N=1,482) disorders as annotated by the Clinical Genomics Database (CGD) in each s_{het} bin. Logarithmic bins are ordered from greatest to smallest s_{het} values. [b] Overall, AD genes have significantly higher s_{het} values than AR genes [Mann-Whitney U p-value 3.14×10^{-64}]. [c] Similarly, in solved Mendelian clinical exome sequencing cases (Baylor)[21], s_{het} values can help discriminate between AR and AD disease genes, as annotated by clinical geneticists. [d] A s_{het} value of 0.04 can be used as a simple classification threshold for AD genes with a PPV of 96%. [e] This finding is replicated in a separately ascertained sample from UCLA. Box plots range from 25th–75th percentile values and whiskers include 1.5 times the interquartile range.

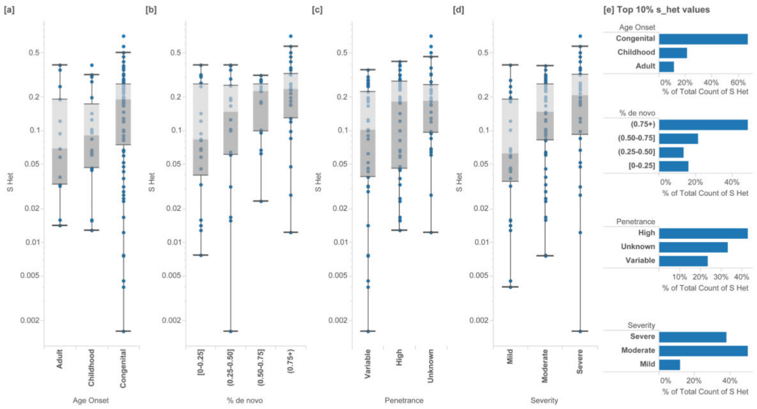

Figure 3.

Enrichments of s_{het} in known haploinsufficient disease genes of high confidence (ClinGen Dosage Sensitivity Project). In (N=127) autosomal genes, we annotate the s_{het} scores of genes associated with each disease category and classification. Higher s_{het} values are associated with [a] earlier age of onset (Mann-Whitney U $p=1.46 \times 10^{-2}$), [b] a larger fraction of *de novo* variants ($p=8\times10^{-5}$), [c] high or unspecified penetrance ($p=1.79 \times10^{-2}$) and [d] increased phenotypic severity ($p=4.87\times10^{-3}$). Box plots range from 25th–75th percentile values and whiskers include 1.5 times the interquartile range. [e] Genes with the 10% highest s_{het} values are also similarly enriched with more severe clinical annotations.

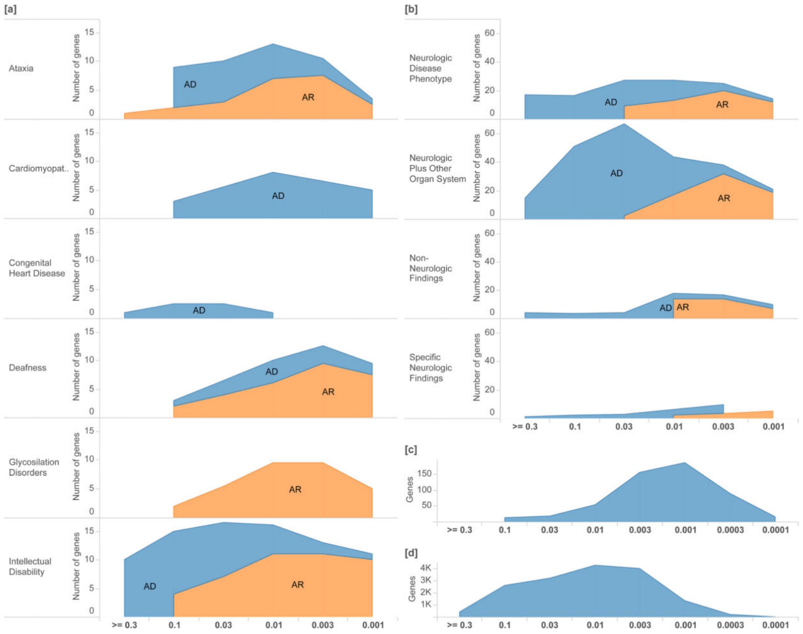

Figure 4.

Distribution of s_{het} values for phenotypes in known disease genes and clinical cases. We plot the distribution of selective effects for different disorder groups, providing information about the breadth and severity of selection associated with each group. [a] We include known Mendelian disease genes (Clinical Genomic Database) annotated as either Autosomal Recessive or Autosomal Dominant and [b] clinical exome sequencing cases[21]. We contrast these with [c] all tolerated knockouts in a consanguineous cohort (PROMIS)[23] and [d] the distribution of selective effects in all scored genes. Logarithmic bins are ordered from greatest to smallest s_{het} values.

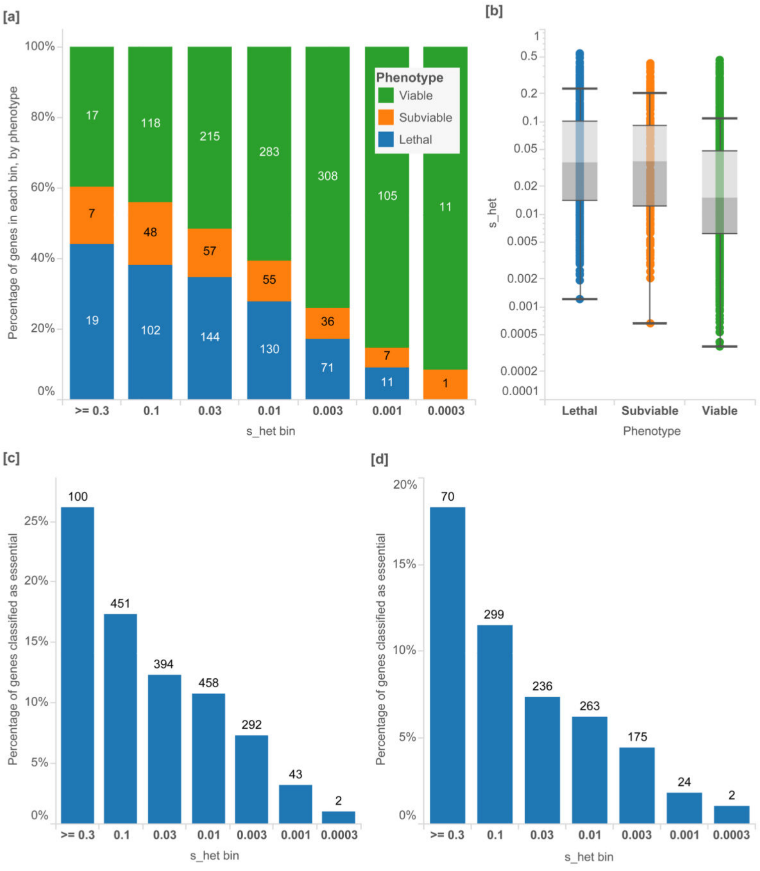

Figure 5.

High-throughput screens of gene essentiality in mice and cell assays, as a percentage of all genes in each s_{het} bin. [a] Proportion of orthologous mouse knockout genes by phenotype, from a neutrally-ascertained set of genes generated by the International Mouse Phenotyping Consortium (IMCP). Logarithmic bins are ordered from greatest to smallest s_{het} values. [b] ICMP mice are separated into viable (N=1,057), sub-viable (N=211) and lethal knockouts (N=477), and lethal knockouts have significantly higher s_{het} values than viable [Mann-Whitney U p-value 2.95×10^{-28}]. [c] Cell-essential genes as reported by Wang *et al.* [3] from genome-wide KBM-7 tumor cell CRISPR assay (N=1,740) have significantly higher s_{het} values [p-value 5.13×10^{-16}] [d] as do genes that were characterized as essential in a gene trap assay (N= 1,081) [p-value = 4.90×10^{-18}]. In the CRISPR assay, all genes with adjusted p-values < 0.05 and negative assay scores are included, and genes with gene trap scores < 0.4 or lower are included. Box plots range from 25th–75th percentile values and whiskers include 1.5 times the interquartile range.

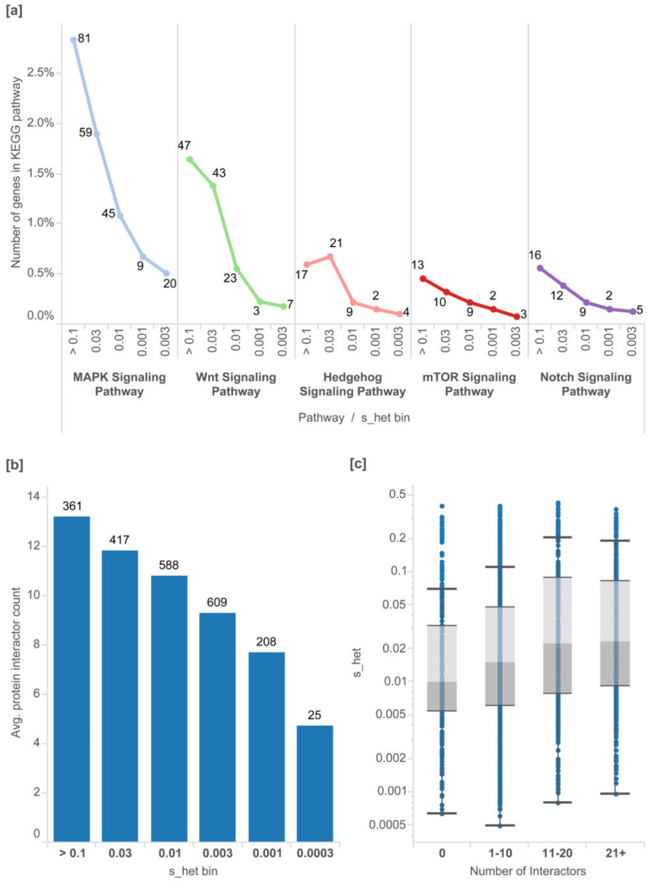

Figure 6.

Protein pathways and protein-protein interactions, as a percentage of the associated developmental genes in each s_{het} bin. [a] In key developmental pathways in KEGG, we find that genes with higher s_{het} values are enriched in genes important to development. [b] We plot the distribution of the number of protein-protein interactions for each gene, as determined by a genome-wide mass spectrometry assay[51] versus s_{het} value. [c] We find that s_{het} values are positively correlated with the number of observed interactors for each gene. Box plots range from 25th–75th percentile values and whiskers include 1.5 times the interquartile range.