nature genetics

Article

https://doi.org/10.1038/s41588-024-01820-9

Bayesian estimation of gene constraint from an evolutionary model with gene features

Received: 2 June 2023

Accepted: 29 May 2024

Published online: 08 July 2024

Check for updates

Tony Zeng[1,4] ✉, Jeffrey P. Spence[1,4] ✉, Hakhamanesh Mostafavi[1,2] & Jonathan K. Pritchard[1,3] ✉

Measures of selective constraint on genes have been used for many applications, including clinical interpretation of rare coding variants, disease gene discovery and studies of genome evolution. However, widely used metrics are severely underpowered at detecting constraints for the shortest ~25% of genes, potentially causing important pathogenic mutations to be overlooked. Here we developed a framework combining a population genetics model with machine learning on gene features to enable accurate inference of an interpretable constraint metric, s_{het}. Our estimates outperform existing metrics for prioritizing genes important for cell essentiality, human disease and other phenotypes, especially for short genes. Our estimates of selective constraint should have wide utility for characterizing genes relevant to human disease. Finally, our inference framework, GeneBayes, provides a flexible platform that can improve the estimation of many gene-level properties, such as rare variant burden or gene expression differences.

Identifying the genes important for disease and fitness is a central goal in human genetics. One particularly useful measure of importance is gene constraint, or how much natural selection limits the population frequencies of deleterious variants[1–4]. If a gene is constrained, then selection will act to remove variants that diminish gene function from the population, such as loss-of-function (LOF) variants. Specifically, LOFs in constrained genes reduce fitness, such that they decrease in frequency or vanish from the population over time.

In the last decade, large exome sequencing studies have made it possible to use LOFs, such as protein truncating or splice-disrupting variants, to calculate metrics of constraint for thousands of genes. Constraint has been used to prioritize de novo and rare variants for clinical follow-up[5,6], predict the toxicity of drugs[7], link genome-wide association studies (GWAS) hits to genes[8] and characterize transcriptional regulation[9,10], among many other applications.

Gene-level constraint metrics typically estimate the depletion in the number of LOFs observed per gene or estimate the fitness decrease from an LOF using a population genetics model that links fitness to the observed LOF frequencies. Specifically, in one line of research, the number of observed unique LOFs is compared to the expected number under a model of no selective constraint. This approach has led to the widely used metrics probability of being LOF intolerant (pLI)[11] and LOF observed/expected upper bound fraction (LOEUF)[12].

While pLI and LOEUF have proved useful for identifying genes intolerant to LOF mutations, they have important limitations[3]. First, they are uninterpretable in that they are only loosely related to the fitness consequences of LOFs. Their relationship with natural selection depends on the study's sample size and other technical factors[3]. Second, the lack of an explicit population genetics model makes it impossible to compare values of pLI or LOEUF to the strength of selection on variants other than LOFs[3,4].

Another line of research has solved these issues of interpretability by estimating the fitness reduction for heterozygous carriers of an LOF in any given gene[1,2,4]. Throughout, we will adopt the notation discussed in ref. 1 and refer to this reduction in fitness as s_{het}[2], although the same population genetic quantity has been referred to as hs[4,13]. In ref. 1, a deterministic approximation was used to estimate s_{het}, which was relaxed to incorporate the effects of genetic drift in ref. 2.

[1]Department of Genetics, Stanford University, Stanford, CA, USA. [2]Department of Population Health, New York University, New York, NY, USA. [3]Department of Biology, Stanford University, Stanford, CA, USA. [4]These authors contributed equally: Tony Zeng, Jeffrey P. Spence. ✉e-mail: tkzeng@stanford.edu; jspence@stanford.edu; pritch@stanford.edu

This model was subsequently extended in ref. 4, with a focus on uncertainty estimation and the interpretability of s_{het}.

A major issue for most previous methods is that thousands of genes have few expected unique LOFs under neutrality, as they have short protein-coding sequences. When LOEUF was introduced[12], it was underpowered for the ~25% of genes with fewer than ten expected unique LOFs. For the same reason, other methods are severely underpowered for this bottom quartile of genes, which we refer to as having 'few expected LOFs'.

Here we present an approach that can accurately estimate s_{het} even for genes with few expected LOFs while maintaining the interpretability of previous population-genetics-based estimates[1,2,4].

Our approach has two main technical innovations. First, we use a flexible population genetics model of LOF allele frequencies. Previous methods have either only modeled the number of unique LOFs, throwing away frequency information[11,12,14], or considered the sum of LOF frequencies across the gene[1,2,4], an approach that is not robust to misannotated LOFs—variants that have been annotated as LOFs but do not abrogate gene function. In contrast to previous approaches, we model the frequencies of individual LOF variants, addressing both limitations. Our approach uses the computational machinery described in a companion paper[15] to accurately obtain the likelihood of observing an LOF at a given frequency.

Second, our approach uses thousands of gene features, including gene expression patterns, protein structure information and evolutionary constraint, to improve estimates for genes with few expected LOFs. By using these features, we can share information across similar genes. Intuitively, this allows us to improve estimates for genes with few expected LOFs by leveraging information from genes with similar features that do have sufficient LOF data. Our approach is similar to that of DeepLOF[14], which uses gene features in a deep learning model to improve the estimation of gene constraint, but DeepLOF scores face the same issues with interpretability as pLI and LOEUF.

We applied our method to gnomAD (v2.1), a large exome sequencing cohort[12]. Our estimates of s_{het} are substantially more predictive than previous metrics at prioritizing essential and disease-associated genes. We additionally use s_{het} to highlight differences in selection on different categories of genes and consider s_{het} in the context of selection on variants beyond LOFs.

Our approach, GeneBayes, is extremely flexible and can be applied to improve the estimation of numerous gene properties beyond s_{het}. Our implementation is available at https://github.com/tkzeng/GeneBayes.

Results

Model overview

Using LOF data to infer gene constraint is challenging for genes with few expected LOFs, with metrics like LOEUF interpreting nearly all such genes as unconstrained (Fig. 1a,b). We hypothesized that it would be possible to improve estimation by using auxiliary information that may be predictive of LOF constraint, including gene expression patterns across tissues, protein structure and evolutionary conservation. By pooling information across groups of similar genes, constraint estimated for genes with sufficient LOF data may help improve estimation for underpowered genes.

However, while the frequencies of LOFs can be related to s_{het} through models from population genetics[1,2,4], we lack an understanding of how other gene features relate to constraint a priori.

To address this problem, we developed a flexible empirical Bayes framework, GeneBayes, that learns the relationship between gene features and s_{het} (Fig. 1c). Our model consists of two main components. First, we model the prior on s_{het} for each gene as a function of its gene features (Fig. 1c, left). Specifically, we train gradient-boosted trees using NGBoost[16] to predict the parameters of each gene's prior distribution from its features, such as its expression level across tissues (Methods; see Supplementary Note for a full list).

Second, we use a model from population genetics to relate s_{het} to the observed LOF data (Fig. 1c, right), allowing us to fit the prior by maximizing the likelihood of the LOF data. Specifically, we use the discrete-time Wright–Fisher model with genic selection, a standard model in population genetics that accounts for mutation and genetic drift[13,17]. In our model, s_{het} is the reduction in fitness per copy of an LOF (Supplementary Note). We assume that the average number of offspring an individual has is proportional to 1, $1 - s_{het}$ or $1 - 2s_{het}$ if they carry zero, one or two copies of the LOF, respectively. Likelihoods are computed using methods described in a companion paper[15].

While previous methods used either the number of unique LOFs or the sum of the frequencies of all LOFs in a gene, our likelihood models the frequency of each individual LOF variant. We used LOF frequencies from the gnomAD consortium (v2.1), which consists of exome sequences from ~125,000 individuals for 19,071 protein-coding genes[12].

Combining these two components—the learned priors and the likelihood of the LOF data—we obtained posterior distributions over s_{het} for every gene (Data availability; Supplementary Table 1). Throughout, we use the posterior mean value of s_{het} for each gene as a point estimate.

Factors affecting the estimation of s_{het}

First, we explored how LOF frequency and mutation rate relate to s_{het} in our population genetics model (Fig. 2a). Invariant sites with high mutation rates are indicative of strong selection ($s_{het} > 10^{-2}$), consistent with ref. 18, while invariant sites with low mutation rates are consistent with essentially any value of s_{het} for the demographic model considered here. Regardless of mutation rate, singletons are consistent with most values of s_{het} but can rule out extremely strong selection, and variants observed at a frequency of >10% rule out even moderately strong selection ($s_{het} > 10^{-3}$).

To assess how informative gene features are about s_{het}, we trained our model on a subset of genes and evaluated the model on held-out genes (Fig. 2b; Methods). We computed the Spearman correlation between s_{het} estimates from the prior and s_{het} estimates from the LOF data only. The correlation is high and comparable between train and test sets (Spearman $\rho = 0.80$ and 0.77, respectively), indicating the gene features alone are highly predictive of s_{het}. Furthermore, posteriors are substantially more concentrated for most genes when using gene features (Fig. 2c).

Some of our features, such as the degree of constraint estimated from missense variants[19], may correlate with LOF variation in ways that don't reflect differences in selection. However, these features do not majorly bias our results (Extended Data Fig. 1a and Supplementary Note). Given that demography has an important role in the likelihood, we further wanted to ensure that our results were robust to the misspecification of the demography. To do this, we trained models on the non-Finnish European (NFE) and the non-NFE subsets of gnomAD (~67,000 and ~56,000 individuals, respectively), and found the resulting s_{het} estimates to be highly concordant with estimates from the full gnomAD dataset (Extended Data Fig. 2 and Supplementary Note).

Next, we compared our estimates of s_{het} to LOEUF and to selection coefficients estimated in ref. 4 (Fig. 2d). To facilitate comparison, we use the posterior modes of s_{het} reported in ref. 4 as point estimates, but note that study in ref. 4 emphasizes the value of using full posterior distributions. While the correlation between our estimates is high for genes with sufficient LOFs (for genes with more LOFs than the median, Spearman ρ with LOEUF = 0.94; ρ with s_{het} (from ref. 4) = 0.87), it drops for genes with few expected LOFs (for genes with fewer LOFs than the median, Spearman ρ with LOEUF = 0.71; ρ with s_{het} (from ref. 4) = 0.69).

We found that many genes are considered constrained by s_{het} but not by LOEUF, which is designed to be highly conservative. In Table 1, we list 15 examples in the top ~15% most constrained genes by s_{het} but in the ~75% least constrained genes by LOEUF (Methods).

https://doi.org/10.1038/s41588-024-01820-9

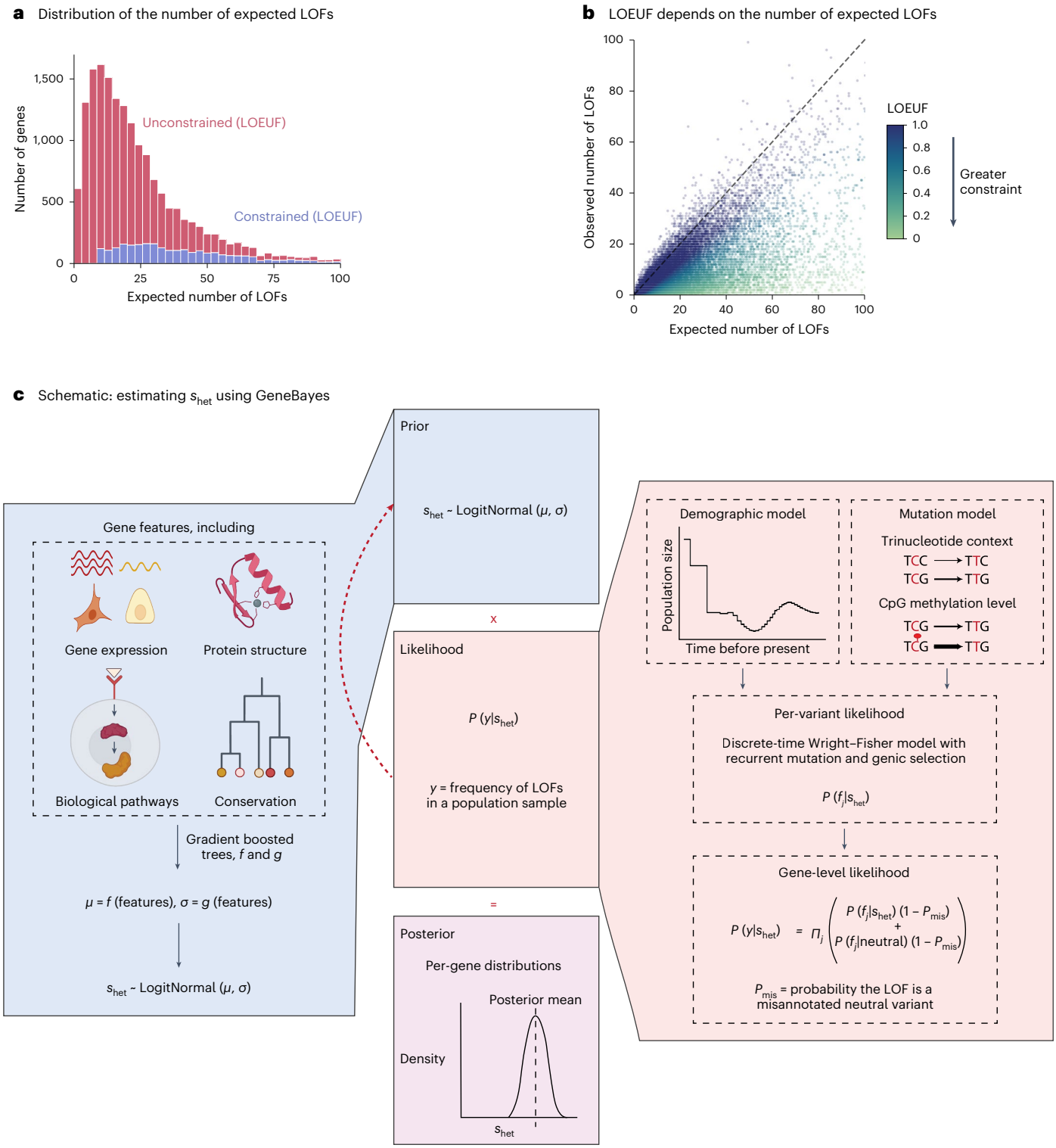


Fig. 1 | Limitations of LOEUF and schematic representation for inferring s_{het} using GeneBayes. a, Stacked histogram of the expected number of unique LOFs per gene, with the distribution for genes considered unconstrained by LOEUF colored in red and those considered constrained colored in blue. Genes with LOEUF < 0.35 are considered constrained, while all other genes are unconstrained (Methods). The plot is truncated on the *x* axis at 100 expected LOFs. **b**, Scatterplot of the observed against the expected number of unique LOFs per gene. The dashed line denotes observed = expected. Each point is a gene, colored by its LOEUF score; genes with LOEUF > 1 are colored as LOEUF = 1. **c**, Schematic representation for estimating s_{het} using GeneBayes, highlighting the major components of the model: prior (blue boxes) and likelihood (red boxes). Parameters of the prior are learned by maximizing the likelihood (red arrow). Combining the prior and likelihood produces posteriors over s_{het} (purple box). See Methods for details. The figure is created with BioRender.com.

One notable example is a set of 18 ribosomal protein genes for which heterozygous disruption causes Diamond–Blackfan anemia[20] (Supplementary Table 2). Sixteen of the genes are considered strongly constrained by s_{het}. In contrast, only six genes are considered constrained by LOEUF (LOEUF < 0.35), as many of these genes have few expected unique LOFs.

https://doi.org/10.1038/s41588-024-01820-9

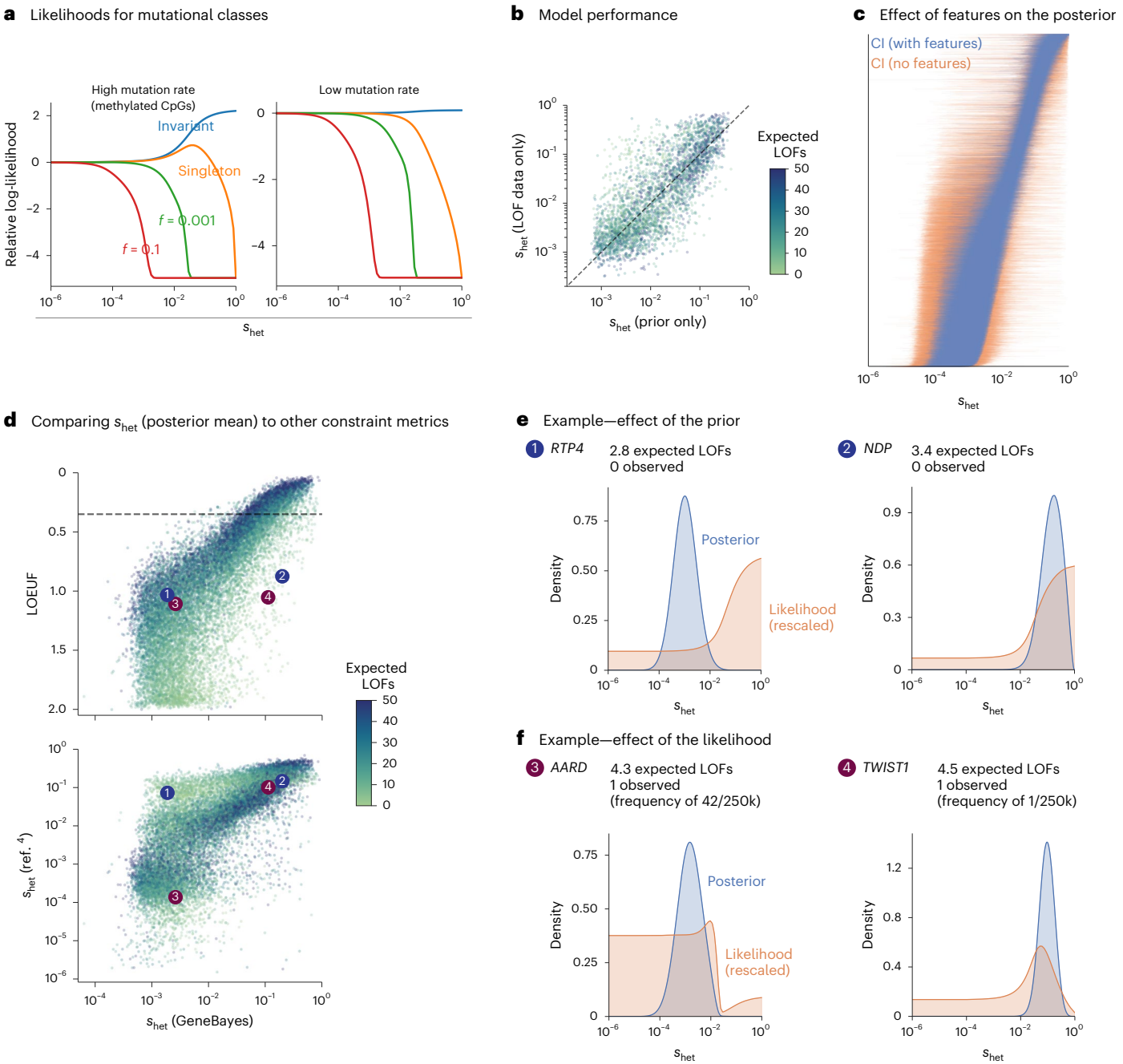


Fig. 2 | Factors that contribute to our estimates of s_{het}. **a**, Likelihood curves for different allele frequencies (f) at sites with high mutation rates (typical of methylated CpGs; left) and low mutation rates (typical of transversions; right). Blue, orange, green and red lines correspond to invariant, singleton, $f = 0.001$ and $f = 0.1$ sites, respectively. **b**, Scatterplot of s_{het} estimated from LOF data (y axis; posterior mean from a model without features) against the prior's predictions of s_{het} (x axis; mean of learned prior). Dashed line denotes $y = x$. Each point is a gene, colored by the expected number of LOFs. **c**, Comparison of posterior distributions of s_{het} (95% credible intervals) from a model with (blue lines) and without (orange lines) gene features. Genes are ordered by their posterior mean in the model with gene features. **d**, Top: scatterplot of LOEUF (y axis) and our s_{het} estimates (x axis; posterior mean). Each point is a gene, colored by the expected number of LOFs. Bottom: scatterplot of s_{het} estimates from ref. 4 (y axis; posterior mode) and our s_{het} estimates (x axis; posterior mean). Numbered points refer to genes in **e** and **f**. **e**, *RTP4* and *NDP* are two examples of genes where the gene features substantially affect the posterior. We plot their posterior distributions (blue) and likelihoods (orange; rescaled so that the area under the curve = 1). **f**, *AARD* and *TWIST1* are two examples of genes with the same LOEUF but different s_{het}. Posteriors and likelihoods are plotted as in **e**.

Next, we explored a few examples to understand the differences between our s_{het} estimates and other measures of constraint. *RTP4* and *NDP* have few expected LOFs, and their likelihoods are consistent with any level of constraint (Fig. 2e). Due to the high degree of uncertainty, LOEUF and the s_{het} point estimates from ref. 4 are uninformative, providing similar estimates for the two genes (Fig. 2d). In contrast, by using gene features, our posterior distributions of s_{het} indicate that *NDP* is strongly constrained but *RTP4* is not, consistent with the observation that hemizygous LOFs in *NDP* cause Norrie disease[21].

Unlike estimates of s_{het}, LOEUF further ignores information about allele frequencies by considering only the number of unique LOFs. For example, *AARD* and *TWIST1* have almost the same number

https://doi.org/10.1038/s41588-024-01820-9

Table 1 | OMIM genes constrained by s_{het} but not by LOEUF

Genes	s_{het}	LOEUF	Obs.	Exp.	Conditions and references
RPS15A[a]	0.68	0.56	0	5.4	Diamond–Blackfan anemia: red blood cell aplasia resulting in growth, craniofacial and other congenital defects[20]
DCX	0.28	0.62	3	12.6	Lissencephaly: migrational arrest of neurons resulting in mental retardation and seizures[58]
UBE2A	0.28	0.54	0	5.6	Intellectual disorder, Nascimento type: intellectual disability characterized by dysmorphic features[59]
PQBP1	0.28	0.50	1	9.5	Renpenning syndrome: mental retardation with short stature and a small head size[60]
NAA10	0.28	0.52	1	9.1	Syndromic microphthalmia: missing or abnormally small eyes from birth[61]
SOX3	0.22	0.86	1	5.5	Intellectual disorder and isolated growth hormone deficiency: impaired fetal growth and intellectual development[62]
NDP	0.20	0.88	0	3.4	Norrie disease: retinal dystrophy resulting in early childhood blindness, mental disorders and deafness[21]
EIF5A	0.19	0.54	1	8.7	Faundes–Banka syndrome: developmental delay, microcephaly and facial dysmorphisms[63]
CDKN1C	0.19	0.53	0	5.7	Beckwith–Wiedemann syndrome: pediatric overgrowth with predisposition to tumor development[64]
BCAP31	0.15	0.65	2	9.7	Deafness, dystonia and cerebral hypomyelination: motor and intellectual disabilities, with deafness and involuntary muscle contraction[65]
SOX2	0.14	0.57	1	8.3	Syndromic microphthalmia: missing or abnormally small eyes from birth[66]
SH2D1A	0.14	0.96	1	4.9	Lymphoproliferative syndrome: immunodeficiency characterized by severe immune dysregulation after viral infection[67]
GATA4	0.12	0.53	3	14.7	Atrial septal defect: congenital heart defect resulting in a hole between the atria[68]
TWIST1	0.11	1.1	1	4.5	Saethre–Chotzen syndrome: craniosynostosis, facial dysmorphism and hand and foot abnormalities[22,23]
TAFAZZIN	0.11	0.49	2	13.0	Barth syndrome: disorder in lipid metabolism characterized by heart, muscle, immune and growth defects[69]

Mutations that disrupt the functions of these genes are associated with Mendelian diseases in the OMIM database[70]. Genes are ordered by s_{het} (posterior mean). Obs. and Exp. are the unique number of observed and expected LOFs, respectively, in the gnomAD (v2.1) release we analyzed. These genes were chosen from 301 genes that had $s_{het} > 0.1$ but were not in the most constrained LOEUF quartile. This includes 71 of 3,045 genes with pathogenic ClinVar variants that fall outside the most constrained LOEUF quartile[12]. [a]*RPS15A* is associated with Diamond–Blackfan anemia along with 12 other genes considered constrained by s_{het} but not by LOEUF (Supplementary Table 2), with 9 of the 12 genes falling outside the most constrained quartile by LOEUF.

of observed and expected unique LOFs, so LOEUF is similar for both. However, *AARD*'s observed LOF is ~40× more frequent than that of *TWIST1*. Consequently, the likelihood rules out the possibility of strong constraint for *AARD* (Fig. 2f), causing the two genes to differ in their estimated s_{het} (Fig. 2d).

In contrast to *AARD*, *TWIST1* has a posterior mean s_{het} of 0.11 when using gene features, indicating very strong selection. Consistent with this, *TWIST1* encodes a transcription factor critical for the specification of the cranial mesoderm, and heterozygous LOFs in the gene are associated with Saethre–Chotzen syndrome[22,23].

We provide additional examples of genes with varying numbers of expected LOFs in Extended Data Fig. 3. As expected, genes with higher numbers of expected LOFs generally have greater concordance between their likelihoods and posterior distributions.

Using s_{het} to prioritize phenotypically important genes

To assess the accuracy of our s_{het} estimates and evaluate their ability to prioritize genes, we first used these estimates to classify genes essential for the survival of human cells in vitro. Genome-wide CRISPR growth screens have quantified the effects of gene knockouts on cell survival or proliferation[24,25]. We find that our estimates of s_{het} outperform other constraint metrics at classifying essential genes (Fig. 3a (left); bootstrap $P < 7 \times 10^{-7}$ for pairwise differences in AUPRC between our estimates and other metrics). The difference is largest for genes with few expected LOFs (Fig. 3a (right)). Our performance gains remain even when compared to LOEUF computed using gnomAD (v4), which contains roughly 6× as many individuals (Extended Data Fig. 4a), consistent with our companion work demonstrating the limited benefits of larger sample size for most genes[15]. In addition, our estimates of s_{het} outperform other metrics at classifying nonessential genes (Extended Data Fig. 4b).

DeepLOF[14], the only other method that combines information from both LOF data and gene features, outperforms methods that rely exclusively on LOF data. However, our method outperforms DeepLOF, likely because DeepLOF considers only the number of unique LOFs, discarding frequency information.

Next, we performed further comparisons of our estimates of s_{het} against LOEUF, as LOEUF and its predecessor pLI are extremely popular metrics of constraint.

In classifying curated developmental disorder genes[26], we find that s_{het} outperforms LOEUF (Fig. 3b; bootstrap $P = 5 \times 10^{-20}$ for the difference in AUPRC) and performs well compared to additional constraint metrics (Extended Data Fig. 4c). The performance of our s_{het} estimates is not strongly dependent on any individually important features (Extended Data Fig. 1b,c). In addition, s_{het} outperforms LOEUF even for genes with sufficient expected LOFs, although the measures become more concordant (Extended Data Fig. 5).

We further considered a broader range of phenotypic abnormalities annotated in the Human Phenotype Ontology (HPO)[27]. For each HPO term, we calculated the enrichment of the 10% most constrained genes and the depletion of the 10% least constrained genes, ranked using s_{het} or LOEUF. Genes considered constrained by s_{het} are more enriched in HPO terms than genes considered constrained by LOEUF (Fig. 3c, left). Additionally, genes considered unconstrained by s_{het} are more depleted in HPO terms than genes considered constrained by LOEUF (Fig. 3c, right).

X-linked inheritance is one of the terms with the largest enrichment of constrained genes (6.7-fold enrichment for s_{het} and 4.1-fold enrichment for LOEUF). The ability of s_{het} to prioritize X-linked genes may prove particularly useful, as the reduced number of X chromosomes in a cohort with males limits the power of population-scale sequencing alone to detect constraint on X chromosome genes[4].

We next assessed if de novo disease-associated variants are enriched in constrained genes, similar to the analyses in refs. 4,5. Using data from 31,058 trios, we calculated for each gene the enrichment of de novo missense and LOF mutations in offspring with developmental disorders (DDs) relative to unaffected parents[5]. We found that for both classes of variants, enrichment is higher for genes considered constrained by s_{het} (Fig. 3d). Consistent with previous findings, the excess burden of de novo variants is predominantly in highly constrained genes (Fig. 3d). Notably, this difference in enrichment remains

https://doi.org/10.1038/s41588-024-01820-9

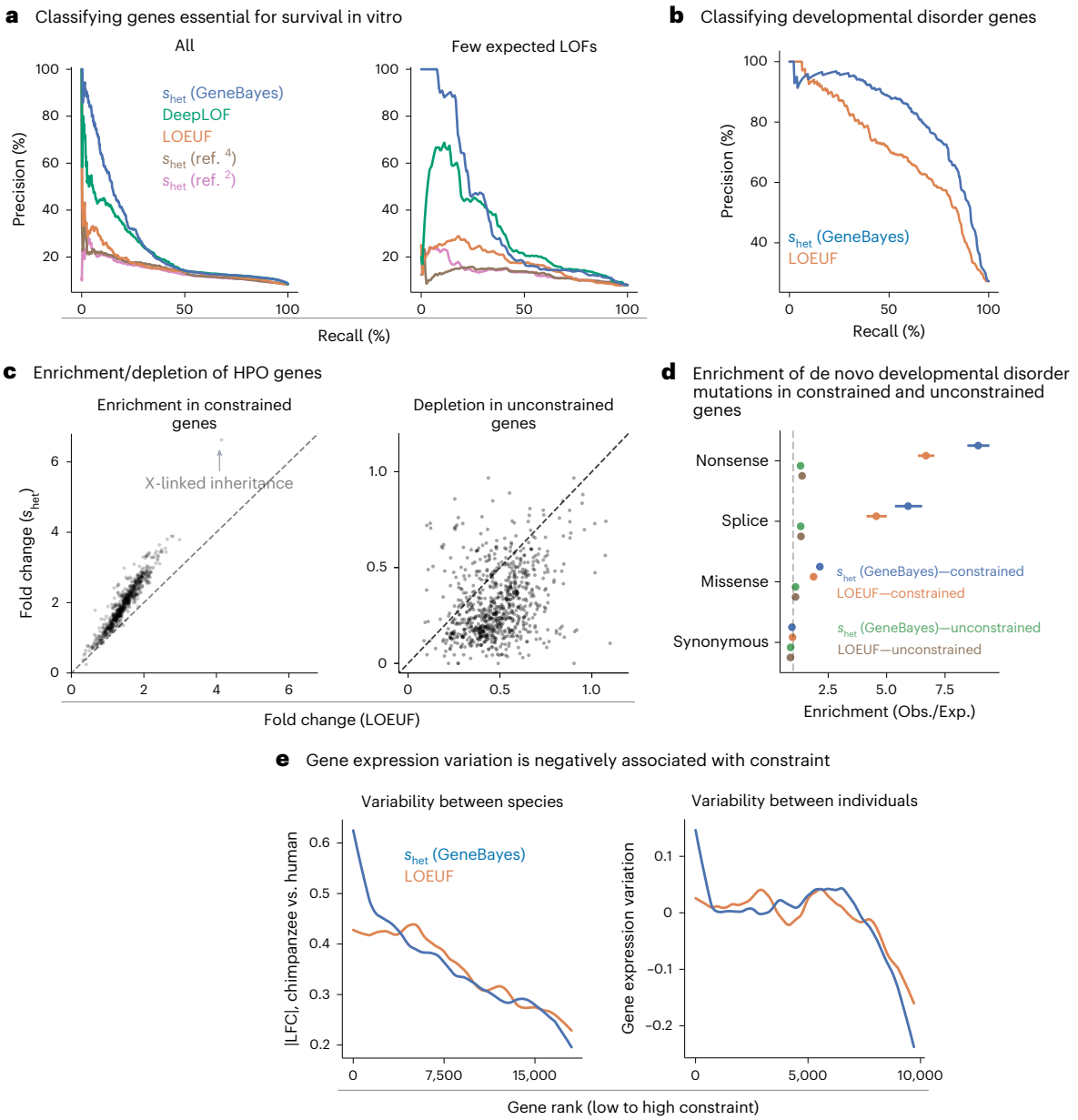


Fig. 3 | GeneBayes estimates of s_{het} perform well at identifying constrained and unconstrained genes. a, Precision–recall curves comparing the performance of s_{het} against other methods in classifying essential genes (left, all genes; right, quartile of genes with the fewest (<9.6) expected unique LOFs). Blue, green, orange, brown and pink lines correspond to s_{het} (GeneBayes), DeepLOF, LOEUF, s_{het} (ref. 4) and s_{het} (ref. 2), respectively. **b**, Precision–recall curves comparing the performance of s_{het} (blue) against LOEUF (orange) in classifying developmental disorder genes. **c**, Scatterplots showing the enrichment of the top 10% most constrained genes and the depletion of the top 10% least constrained genes in HPO terms, with genes ranked by s_{het} (*y* axis) or LOEUF (*x* axis). **d**, Enrichment of DNMs in patients with developmental disorders, calculated as the observed number of mutations over the expected number under a null mutational model ($n = 31{,}058$ parent–offspring trios). We plot the enrichment of synonymous, missense, splice and nonsense variants in the 10% most constrained genes, ranked by s_{het} (blue) or LOEUF (orange), or enrichment in the remaining genes, ranked by s_{het} (green) or LOEUF (brown). Bars represent 95% confidence intervals, centered around the mean. **e**, Left: LOESS curve showing the relationship between constraint (gene rank, *x* axis) and absolute LFC in expression between chimpanzee and human cortical cells (*y* axis). Genes are ranked by s_{het} (blue) or LOEUF (orange). Right: LOESS curve showing the relationship between constraint (gene rank, *x* axis) and gene expression variation in GTEx samples after controlling for mean expression levels (*y* axis). Genes are ranked by s_{het} (blue) or LOEUF (orange).

after removing known DD genes (Extended Data Fig. 4d). Together, these results indicate that s_{het} may facilitate the discovery of new DD genes[5].

In addition to rare de novo disease-associated variants, we find that common variant heritability computed using stratified linkage disequilibrium (LD) score regression is enriched in constrained genes (Extended Data Fig. 4e; Methods), consistent with the findings from ref. 12. For 380 of 438 highly heritable traits (87%), heritability is more highly enriched in the decile of genes most highly constrained by s_{het} than the decile most highly constrained by LOEUF.

Finally, constraint can also be related to longer-term evolutionary processes. For example, we expect constrained genes to maintain expression levels closer to their optimal values across evolutionary time scales. Consistent with this expectation, we find that less constrained genes have larger differences in expression between human and chimpanzee in cortical cells[28], with a stronger correlation for

https://doi.org/10.1038/s41588-024-01820-9

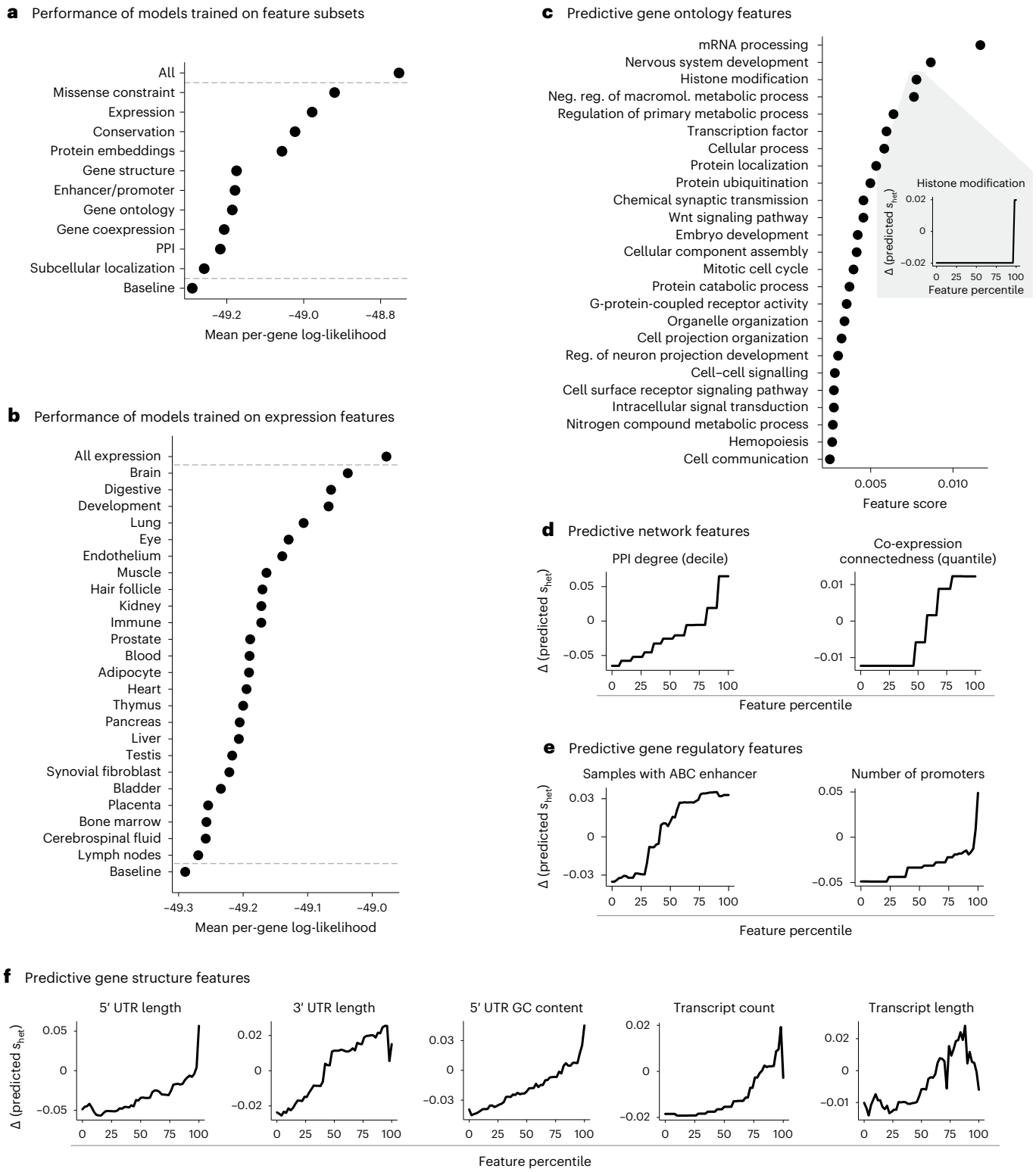


Fig. 4 | Breakdown of the gene features that are important for s_{het} prediction. **a**, Ordered from highest to lowest, a plot of the mean per-gene log-likelihood over the test genes for models separately trained on categories of features. 'All' and 'baseline' include all and no features, respectively. **b**, Plot of the mean per-gene log-likelihood, as in **a**, for models separately trained on expression features grouped by tissue, cell type or developmental stage. **c**, Ordered from highest to lowest, feature scores for individual GO terms. Inset: lineplot showing the change in predicted s_{het} for a feature as the feature value is varied. **d**–**f**, Lineplot as in **c** (inset) for PPI and co-expression features (**d**), enhancer and promoter features (**e**) and gene structure features (**f**).

s_{het} than for LOEUF (Fig. 3e). Similarly, we quantified gene expression variability in human populations[29] and found that variance decreases with increased constraint, again with a stronger correlation for s_{het} (Fig. 3e).

Interpreting the relationship between gene features and s_{het}

Our framework allows us to learn the relationship between gene features and s_{het} while accounting for dependencies between the features. To interrogate the relationship between features and s_{het}, we divided

https://doi.org/10.1038/s41588-024-01820-9

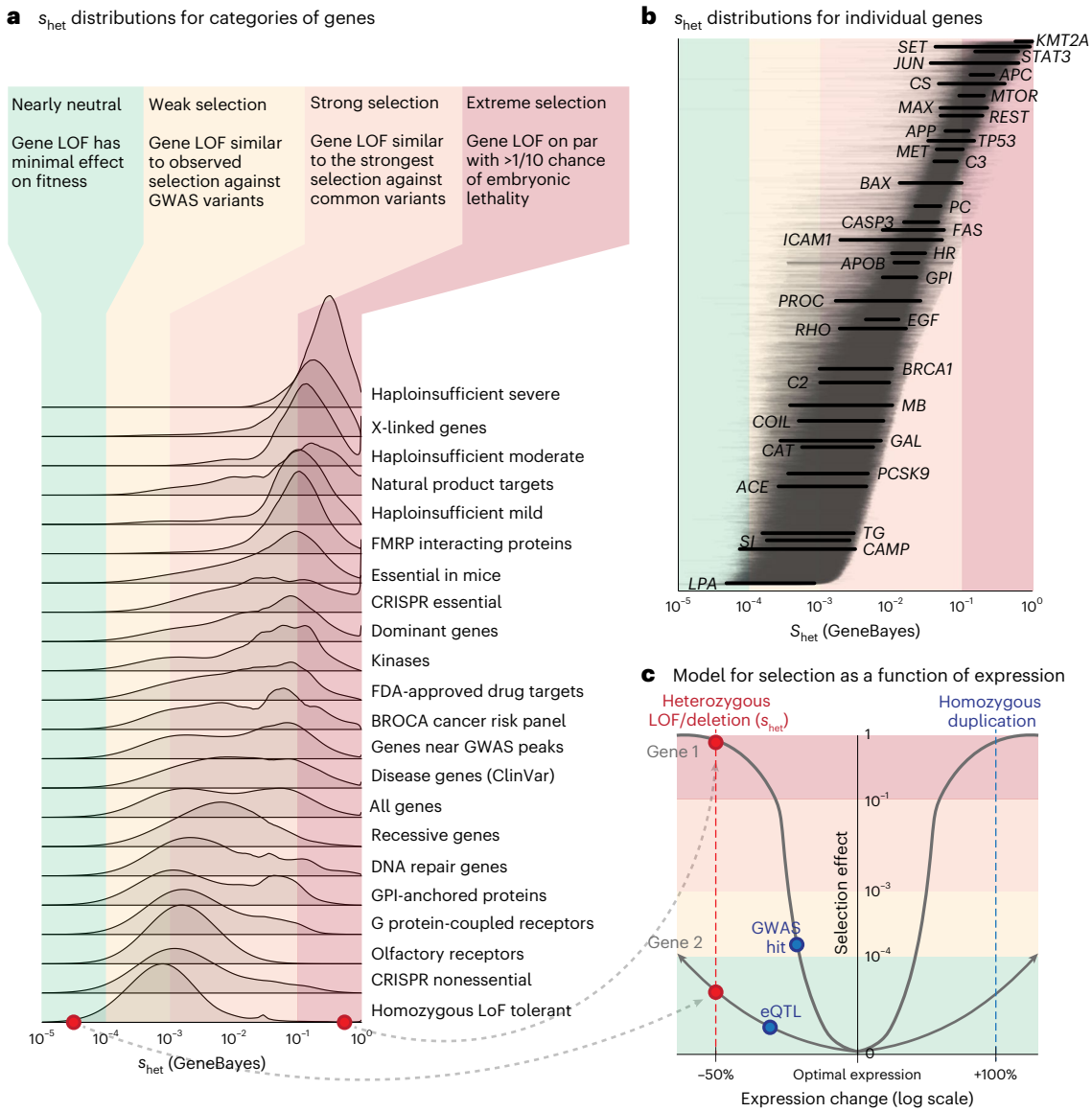


Fig. 5 | Comparing selection on LOFs (s_{het}) between genes and s_{het} to selection on other variant types. a, Distributions of s_{het} for gene sets, calculated by averaging the posterior distributions for the genes in each gene set. Gene sets are sorted by the mean of their distributions. Colors represent four general selection regimes—nearly neutral (light green), weak selection (light yellow), strong selection (light orange) and extreme selection (light red). See the text for a detailed description of the selection regimes. **b**, Posterior distributions of s_{het} for individual genes, ordered by mean. Lines represent 95% credible intervals, with labeled genes represented by thick black lines. Colors represent the selection regimes in **a**. **c**, Schematic representation demonstrating the hypothesized relationship between changes in expression (*x* axis, $\log_2$ scale) and selection (*y* axis) against these changes for two hypothetical genes, assuming stabilizing selection. The shapes of the curves are not estimated from real data. Background colors represent the selection regimes in **a**. The red points and line represent the effects of heterozygous LOFs and deletions on expression and selection, while the blue points and line represent the potential effects of other types of variants.

our gene features into ten distinct categories (Fig. 4a) and trained a separate model per category using only the features in that category. We found that missense constraint, gene expression patterns, evolutionary conservation and protein embeddings are the most informative categories.

Next, we further divided the expression features into 24 subgroups, representing tissues, cell types and developmental stages (Supplementary Table 3). Expression patterns in the brain, digestive system and during development are the most predictive of constraint (Fig. 4b). Notably, a study that matched Mendelian disorders to tissues through a literature review found that a sizable plurality affects the brain[30]. Meanwhile, most of the top digestive expression features are also related to development (for example, ref. 31). The importance of developmental features is consistent with the severity of many developmental disorders and the expectation that selection is stronger on early onset phenotypes[4,32].

To quantify the relationship between constraint and individual features, we changed the value of one feature at a time and used the variation in predicted s_{het} over the feature values as the score for each feature (Methods). First, consistent with the top expression features, the top Gene Ontology (GO) features highlight developmental and brain-specific processes as important for selection (Fig. 4c).

Next, we analyzed network (Fig. 4d), gene regulatory (Fig. 4e) and gene structure (Fig. 4f) features. Protein–protein interaction (PPI) and gene co-expression networks have highlighted 'hub' genes involved in numerous cellular processes[33,34], while genes linked to GWAS variants have more complex enhancer landscapes[35]. Consistent with these studies, we find that network connectedness and enhancer/promoter

https://doi.org/10.1038/s41588-024-01820-9

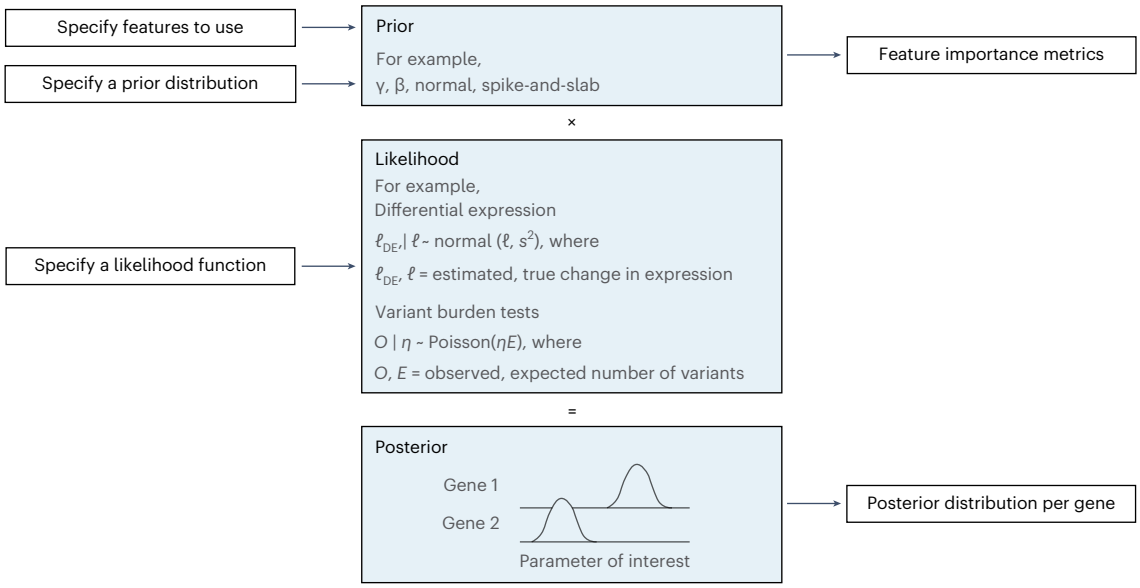


Fig. 6 | GeneBayes is a flexible framework for estimating gene-level properties. Schematic representation for how GeneBayes can be applied to estimate gene-level properties beyond s_{het}, showing the key inputs and outputs and two example applications. See Supplementary Note for more details.

count are positively associated with constraint (Fig. 4d,e). In addition, several gene structure features are predictive of constraint (Fig. 4f), consistent with recent work on UTRs[36]. Our results indicate that more complex genes—genes involved in more regulatory connections, more central to networks and with more complex gene structures—are generally more constrained.

Gene length is predictive of constraint yet also correlates with the amount of information in the LOF data (Fig. 4f), such that measures like LOEUF depend strongly on gene length. Furthermore, it correlates with several other gene features (Extended Data Fig. 6a–c). However, gene length explains at most a modest amount of the correlation between most features and s_{het} (Extended Data Fig. 6d).

Contextualizing the strength of selection against gene LOF

A major benefit of s_{het} over LOEUF and pLI is that s_{het} has a precise, intrinsic meaning in terms of fitness[1–4]. This facilitates the comparison of s_{het} between genes, populations, species and studies. More broadly, consequences of noncoding, missense and copy number variants can be understood through the same framework, as we expect such variants to also be under negative selection[18] due to ubiquitous stabilizing selection on traits[37]. Quantifying differences in the selection on variants will deepen our understanding of the evolution and genetics of human traits ('Discussion').

To contextualize our s_{het} estimates, we compared the distributions of s_{het} for different gene sets (Fig. 5a) and genes (Fig. 5b) and analyzed them in terms of selection regimes. To define such regimes, we first conceptualized the selection on variants as a function of their effects on expression (Fig. 5c), where heterozygous LOFs usually reduce expression by ~50%. Under this framework, we can directly compare s_{het} to selection on other variant types—for the hypothetical genes in Fig. 5c, a GWAS hit affecting gene 1 has a stronger selective effect than an LOF affecting gene 2, despite having a smaller effect on expression.

Next, we divided the range of possible s_{het} values into four regimes determined by theoretical considerations[38] and comparisons to other types of variants[39,40]—nearly neutral, weak selection, strong selection and extreme selection. LOFs in nearly neutral genes ($s_{het} < 10^{-4}$) have minimal effects on fitness—the frequency of such variants is dominated by genetic drift rather than selection[38]. Under the weak selection regime (s_{het} from 10^{-4} to 10^{-3}), gene LOFs have similar effects on fitness as typical GWAS hits, which usually have small or context-specific effects on gene expression or function[39]. Under the strong selection regime (s_{het} from 10^{-3} to 10^{-1}), gene LOFs have fitness effects on par with the strongest selection coefficients measured for common variants, such as the selection estimated for adaptive mutations in *LCT*[40]. Finally, for genes in the extreme selection regime ($s_{het} > 10^{-1}$), LOFs have an effect on fitness equivalent to a >10% chance of embryonic lethality.

Gene sets vary widely in their constraint. For example, genes known to be haploinsufficient for severe diseases are almost all under extreme selection. In contrast, genes that can tolerate homozygous LOFs are generally under weak selection. One notable example of such a gene is *LPA*—while high expression levels are associated with cardiovascular disease, low levels have minimal phenotypic consequences[41,42], consistent with limited conservation in the sequence or gene expression of *LPA* across species and populations[43,44].

Other gene sets have much broader distributions of s_{het} values. For example, manually curated recessive genes are under weak to strong selection, indicating that many such genes are either not fully recessive or have pleiotropic effects on other traits under selection. For example, homozygous LOFs in *PROC* can cause life-threatening congenital blood clotting[45], yet s_{het} for *PROC* is nonnegligible (Fig. 5b), consistent with observations that heterozygous LOFs can also increase blood clotting and cause deep vein thrombosis[46].

Discussion

Here we developed an empirical Bayes approach to accurately infer s_{het}, an interpretable metric of gene constraint. Our approach uses powerful machine learning methods to leverage vast amounts of functional and evolutionary information about each gene while coupling them to a population genetics model.

There are two advantages of this approach. First, the additional data sources result in substantially better performance than LOEUF across tasks, from classifying essential genes to identifying pathogenic de novo mutations (DNMs). These improvements are especially pronounced for the large fraction of genes with few expected LOFs, where LOF data alone are underpowered for estimating constraint.

Second, by inferring s_{het}, our estimates of constraint are interpretable in terms of fitness, and we can directly compare the impact of an LOF across genes, populations, species and studies.

As a selection coefficient, s_{het} can also be directly compared to other selection coefficients, even for different types of variants[3,4]. Theory suggests that genes are generally close to their optimal levels of expression and are mainly subject to stabilizing selection[37], in which case expression-altering variants decrease fitness, with larger perturbations causing greater decreases (Fig. 5c). Estimating the fitness consequences of other types of expression-altering variants, such as duplications or expression quantitative trait loci (eQTLs), will allow us to map the relationship between genetic variation and fitness in detail, deepening our understanding of the interplay of expression, complex traits and fitness[10,39,47,48].

A recent method, DeepLOF[14], uses a similar empirical Bayes approach, but by estimating constraint from the number of observed and expected unique LOFs, it inherits the same difficulties regarding interpretation as pLI and LOEUF, and loses information by not considering variant frequencies. Another line of work[1,2], culminating in ref. 4, solved the issues with interpretability by directly estimating s_{het}. Yet, by relying exclusively on LOFs, these estimates are underpowered for ~25% of genes. Furthermore, by using the aggregate frequencies of all LOF variants, previous s_{het} estimates[1,2,4] are not robust to misannotated LOF variants. Our approach eliminates this tradeoff between power and interpretability present in existing metrics.

Similar insights that combine evolutionary modeling and genomic features have been used to estimate constraint on noncoding variation[49–52].

Our estimates of s_{het} will be useful for many applications. For example, by informing gene level priors, LOEUF, pLI and previous estimates of s_{het} have been used to increase the power of association studies based on rare mutations or DNMs[5,6,53]. In such contexts, our s_{het} estimates can be used as a drop-in replacement. Additionally, investigating highly constrained genes may give insights into the mechanisms by which cellular and organism-level phenotypes affect fitness[54].

While we primarily used the posterior means of s_{het} here, our approach provides the entire posterior distribution per gene, similar to ref. 4. In some applications, different aspects of the posterior may be more relevant than the mean. For example, when prioritizing rare variants for follow-up in a clinical setting, the posterior probability that s_{het} is high enough for the variant to severely reduce fitness may be more relevant.

As more exomes are sequenced, one might expect that we would be better able to more accurately estimate s_{het}. Indeed, for non-European genetic ancestry groups, larger samples may facilitate a more accurate estimation of ancestry-specific s_{het}, a challenging task given the sample sizes available in gnomAD (v2.1). Yet, we show in a companion paper[15] that increasing the sample size for estimating LOF frequencies beyond ~140,000 individuals (the approximate aggregate size of gnomAD (v2.1)) will only improve estimates slowly and provide essentially no additional information for the ~85% of genes with the lowest values of s_{het}. By sharing information across genes, we can overcome this fundamental limit on how accurately we can estimate constraint.

Here we focused on estimating s_{het}, but our empirical Bayes framework, GeneBayes, can be used in any setting where one has a model that ties a gene-level parameter to gene-level observable data (Supplementary Note). For example, GeneBayes can be used to find trait-associated genes using variants from case–control studies[55,56] or to improve the power to find differentially expressed genes in RNA sequencing (RNA-seq) experiments[57]. We provide a graphical overview of how GeneBayes can be applied more generally in Fig. 6. Briefly, GeneBayes requires users to specify a likelihood model and the form of a prior distribution for their parameter of interest. Then, using empirical Bayes and a set of gene features, it improves the power to estimate the parameter by flexibly sharing information across similar genes.

In summary, we developed a powerful framework for estimating a broadly applicable and readily interpretable metric of constraint, s_{het}. Our estimates provide a more informative ranking of gene importance than existing metrics, and our approach allows us to interrogate potential causes and consequences of natural selection.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-024-01820-9.

References

1. Cassa, C. A. et al. Estimating the selective effects of heterozygous protein-truncating variants from human exome data. *Nat. Genet.* **49**, 806–810 (2017).
2. Weghorn, D. et al. Applicability of the mutation–selection balance model to population genetics of heterozygous protein-truncating variants in humans. *Mol. Biol. Evol.* **36**, 1701–1710 (2019).
3. Fuller, Z. L., Berg, J. J., Mostafavi, H., Sella, G. & Przeworski, M. Measuring intolerance to mutation in human genetics. *Nat. Genet.* **51**, 772–776 (2019).
4. Agarwal, I., Fuller, Z. L., Myers, S. R. & Przeworski, M. Relating pathogenic loss-of-function mutations in humans to their evolutionary fitness costs. *eLife* **12**, e83172 (2023).
5. Kaplanis, J. et al. Evidence for 28 genetic disorders discovered by combining healthcare and research data. *Nature* **586**, 757–762 (2020).
6. Fu, J. M. et al. Rare coding variation provides insight into the genetic architecture and phenotypic context of autism. *Nat. Genet.* **54**, 1320–1331 (2022).
7. Whiffin, N. et al. The effect of LRRK2 loss-of-function variants in humans. *Nat. Med.* **26**, 869–877 (2020).
8. Gazal, S. et al. Combining SNP-to-gene linking strategies to identify disease genes and assess disease omnigenicity. *Nat. Genet.* **54**, 827–836 (2022).
9. Wang, X. & Goldstein, D. B. Enhancer domains predict gene pathogenicity and inform gene discovery in complex disease. *Am. J. Hum. Genet.* **106**, 215–233 (2020).
10. Mostafavi, H., Spence, J. P., Naqvi, S. & Pritchard, J. K. Systematic differences in discovery of genetic effects on gene expression and complex traits. *Nat. Genet.* **55**, 1866–1875 (2023).
11. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
12. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
13. Gillespie, J. H. *Population Genetics: A Concise Guide* (JHU Press, 2004).
14. LaPolice, T. M. & Huang, Y. F. An unsupervised deep learning framework for predicting human essential genes from population and functional genomic data. *BMC Bioinformatics* **24**, 347 (2023).
15. Spence, J. P., Zeng, T., Mostafavi, H. & Pritchard, J. K. Scaling the discrete-time Wright–Fisher model to biobank-scale datasets. *Genetics* **225**, iyad168 (2023).
16. Duan, T. et al. Ngboost: natural gradient boosting for probabilistic prediction. In *Proc. International Conference on Machine Learning* (eds Daumé, H. III & Singh, A.) 2690–2700 (PMLR, 2020).
17. Ewens, W. J. *Mathematical Population Genetics: Theoretical Introduction* Vol. 27 (Springer, 2004).
18. Agarwal, I. & Przeworski, M. Mutation saturation for fitness effects at human CpG sites. *eLife* **10**, e71513 (2021).
19. Huang, Y. F. Unified inference of missense variant effects and gene constraints in the human genome. *PLoS Genet.* **16**, e1008922 (2020).
20. Da Costa, L., Leblanc, T. & Mohandas, N. Diamond–Blackfan anemia. *Blood* **136**, 1262–1273 (2020).

21. Berger, W. et al. Mutations in the candidate gene for Norrie disease. *Hum. Mol. Genet.* **1**, 461–465 (1992).
22. Howard, T. D. et al. Mutations in TWIST, a basic helix–loop–helix transcription factor, in Saethre–Chotzen syndrome. *Nat. Genet.* **15**, 36–41 (1997).
23. Ghouzzi, V. E. et al. Mutations of the TWIST gene in the Saethre–Chotzene syndrome. *Nat. Genet.* **15**, 42–46 (1997).
24. Meyers, R. M. et al. Computational correction of copy number effect improves specificity of CRISPR–Cas9 essentiality screens in cancer cells. *Nat. Genet.* **49**, 1779–1784 (2017).
25. Ghandi, M. et al. Next-generation characterization of the cancer cell line encyclopedia. *Nature* **569**, 503–508 (2019).
26. Wright, C. F. et al. Genomic diagnosis of rare pediatric disease in the United Kingdom and Ireland. *N. Engl. J. Med.* **388**, 1559–1571 (2023).
27. Köhler, S. et al. The Human Phenotype Ontology in 2021. *Nucleic Acids Res.* **49**, D1207–D1217 (2021).
28. Agoglia, R. M. et al. Primate cell fusion disentangles gene regulatory divergence in neurodevelopment. *Nature* **592**, 421–427 (2021).
29. GTEx Consortium The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
30. Basha, O. et al. Differential network analysis of multiple human tissue interactomes highlights tissue-selective processes and genetic disorder genes. *Bioinformatics* **36**, 2821–2828 (2020).
31. Gao, S. et al. Tracing the temporal-spatial transcriptome landscapes of the human fetal digestive tract using single-cell RNA-sequencing. *Nat. Cell Biol.* **20**, 721–734 (2018).
32. Charlesworth, B. et al. *Evolution in Age-Structured Populations* Vol. 2 (Cambridge University Press, 1994).
33. Barrio-Hernandez, I. et al. Network expansion of genetic associations defines a pleiotropy map of human cell biology. *Nat. Genet.* **55**, 389–398 (2023).
34. Van Dam, S., Vosa, U., van der Graaf, A., Franke, L. & de Magalhaes, J. P. Gene co-expression analysis for functional classification and gene–disease predictions. *Brief. Bioinform.* **19**, 575–592 (2018).
35. Nasser, J. et al. Genome-wide enhancer maps link risk variants to disease genes. *Nature* **593**, 238–243 (2021).
36. Wieder, N. et al. Differences in 5′ untranslated regions highlight the importance of translational regulation of dosage sensitive genes. *Genome Biol.* **25**, 111 (2024).
37. Sella, G. & Barton, N. H. Thinking about the evolution of complex traits in the era of genome-wide association studies. *Annu. Rev. Genomics Hum. Genet.* **20**, 461–493 (2019).
38. Charlesworth, B. Effective population size and patterns of molecular evolution and variation. *Nat. Rev. Genet.* **10**, 195–205 (2009).
39. Simons, Y. B., Mostafavi, H., Smith, C. J., Pritchard, J. K. & Sella, G. Simple scaling laws control the genetic architectures of human complex traits. Preprint at *bioRxiv* https://doi.org/10.1101/2022.10.04.509926 (2022).
40. Mathieson, I. & Terhorst, J. Direct detection of natural selection in Bronze Age Britain. *Genome Res.* **32**, 2057–2067 (2022).
41. Emdin, C. A. et al. Phenotypic characterization of genetically lowered human lipoprotein(a) levels. *J. Am. Coll. Cardiol.* **68**, 2761–2772 (2016).
42. Langsted, A., Nordestgaard, B. G. & Kamstrup, P. R. Low lipoprotein(a) levels and risk of disease in a large, contemporary, general population study. *Eur. Heart J.* **42**, 1147–1156 (2021).
43. Rausell, A. et al. Common homozygosity for predicted loss-of-function variants reveals both redundant and advantageous effects of dispensable human genes. *Proc. Natl Acad. Sci. USA* **117**, 13626–13636 (2020).
44. Reyes-Soffer, G. et al. Lipoprotein(a): a genetically determined, causal, and prevalent risk factor for atherosclerotic cardiovascular disease: a scientific statement from the American Heart Association. *Arterioscler. Thromb. Vasc. Biol.* **42**, e48–e60 (2022).
45. Millar, D. S. et al. Molecular genetic analysis of severe protein C deficiency. *Hum. Genet.* **106**, 646–653 (2000).
46. Romeo, G. et al. Hereditary thrombophilia: identification of nonsense and missense mutations in the protein C gene. *Proc. Natl Acad. Sci. USA* **84**, 2829–2832 (1987).
47. O'Connor, L. J. et al. Extreme polygenicity of complex traits is explained by negative selection. *Am. J. Hum. Genet.* **105**, 456–476 (2019).
48. Benton, M. L. et al. The influence of evolutionary history on human health and disease. *Nat. Rev. Genet.* **22**, 269–283 (2021).
49. Gulko, B., Hubisz, M. J., Gronau, I. & Siepel, A. A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nat. Genet.* **47**, 276–283 (2015).
50. Huang, Y. F., Gulko, B. & Siepel, A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat. Genet.* **49**, 618–624 (2017).
51. Huang, Y. F. & Siepel, A. Estimation of allele-specific fitness effects across human protein-coding sequences and implications for disease. *Genome Res.* **29**, 1310–1321 (2019).
52. Chen, S. et al. A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* **625**, 92–100 (2024).
53. Satterstrom, F. K. et al. Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. *Cell* **180**, 568–584 (2020).
54. Gardner, E. J. et al. Reduced reproductive success is associated with selective constraint on human genes. *Nature* **603**, 858–863 (2022).
55. He, X. et al. Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. *PLoS Genet.* **9**, e1003671 (2013).
56. Zhu, X. & Stephens, M. Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. *Ann. Appl. Stat.* **11**, 1561–1592 (2017).
57. Boyeau, P. et al. An empirical Bayes method for differential expression analysis of single cells with deep generative models. *Proc. Natl Acad. Sci. USA* **120**, e2209124120 (2023).
58. Des Portes, V. et al. A novel CNS gene required for neuronal migration and involved in X-linked subcortical laminar heterotopia and lissencephaly syndrome. *Cell* **92**, 51–61 (1998).
59. Nascimento, R. M., Otto, P. A., de Brouwer, A. P. & Vianna-Morgante, A. M. UBE2A, which encodes a ubiquitin-conjugating enzyme, is mutated in a novel X-linked mental retardation syndrome. *Am. J. Hum. Genet.* **79**, 549–555 (2006).
60. Stevenson, R. E. et al. Renpenning syndrome comes into focus. *Am. J. Med. Genet. A* **134**, 415–421 (2005).
61. Esmailpour, T. et al. A splice donor mutation in NAA10 results in the dysregulation of the retinoic acid signalling pathway and causes Lenz microphthalmia syndrome. *J. Med. Genet.* **51**, 185–196 (2014).
62. Laumonnier, F. et al. Transcription factor SOX3 is involved in X-linked mental retardation with growth hormone deficiency. *Am. J. Hum. Genet.* **71**, 1450–1455 (2002).
63. Faundes, V. et al. Impaired eIF5A function causes a Mendelian disorder that is partially rescued in model systems by spermidine. *Nat. Commun.* **12**, 833 (2021).
64. Hatada, I. et al. An imprinted gene *p57 KIP2* is mutated in Beckwith–Wiedemann syndrome. *Nat. Genet.* **14**, 171–173 (1996).
65. Cacciagli, P. et al. Mutations in BCAP31 cause a severe X-linked phenotype with deafness, dystonia, and central hypomyelination and disorganize the Golgi apparatus. *Am. J. Hum. Genet.* **93**, 579–586 (2013).
66. Fantes, J. et al. Mutations in SOX2 cause anophthalmia. *Nat. Genet.* **33**, 462–463 (2003).

67. Nichols, K. E. et al. Inactivating mutations in an SH2 domain-encoding gene in X-linked lymphoproliferative syndrome. *Proc. Natl Acad. Sci. USA* **95**, 13765–13770 (1998).
68. Garg, V. et al. GATA4 mutations cause human congenital heart defects and reveal an interaction with TBX5. *Nature* **424**, 443–447 (2003).
69. Bione, S. et al. A novel X-linked gene, G4. 5. is responsible for Barth syndrome. *Nat. Genet.* **12**, 385–389 (1996).
70. Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh, A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* **43**, D789–D798 (2015).

Methods

Inclusion and ethics

The study did not require any specific ethics approval. It relies on summary data from aggregated exome sequencing data made publicly available by gnomAD but does not use any individual-level data. Specifically, we use allele count and frequency data for predicted LOF variants for ancestry groups assigned by gnomAD, such as 'NFE'. All other data used are also publicly available and contain no individual-level data (see Methods for descriptions and data sources).

Statistics and reproducibility

No preliminary statistical analyses were conducted to determine sample sizes. The inclusion of individuals in the summary data made available by gnomAD was based on criteria described in ref. 12. Feature selection, model training and model evaluation are described in Methods, and the code for model training is publicly available (Code availability).

Empirical Bayes overview

Many genes have few observed LOF variants, making it challenging to infer constraint without additional information. Bayesian approaches that specify a prior distribution for each gene can provide such information to improve constraint estimates, but specifying prior distributions is challenging as we have limited prior knowledge about the selection coefficients, s_{het}. Empirical Bayes procedures allow us to learn a prior distribution for each gene by combining information across genes.

To use the information contained in the gene features, we learn a mapping from a gene's features to a prior specific for that gene. We parameterize this mapping using gradient-boosted trees, as implemented in NGBoost[16]. Intuitively, this approach learns a notion of 'similarity' between genes based on their features and then shares information across similar genes to learn how s_{het} relates to the gene features. This approach has two major benefits. First, by sharing information between similar genes, it can dramatically improve the accuracy of the predicted s_{het} values, particularly for genes with few expected LOFs. Second, by leveraging the LOF data, this approach allows us to learn about how the various gene features relate to fitness, which cannot be modeled from first principles.

For a more in-depth description of our approach along with mathematical and implementation details, see Supplementary Note.

Population genetic likelihood

To model how s_{het} relates to the frequency of individual LOF variants, we used the discrete-time Wright–Fisher model, with an approximation of diploid selection with additive fitness effects. We used a composite likelihood approach, assuming independence across individual LOF variants, to obtain gene-level likelihoods. Within this composite likelihood, we model each individual variant as either having a selection coefficient of s_{het} with probability $1 - P_{mis}$ or having a selection coefficient of 0 with probability P_{mis}. That is, P_{mis} acts as the prior probability that a given variant is misannotated, and we assume that misannotated variants evolve neutrally regardless of the strength of selection on the gene. All likelihoods were computed using machinery developed in a companion paper[15].

Our model depends on a number of parameters—a demographic model of past population sizes, mutation rates for each site and the probability of misannotation. The demographic model is taken from the literature[71] with modifications as described in ref. 4. The mutation rates account for trinucleotide context as well as methylation status at CpGs[12]. Finally, we estimated the probability of misannotation from the data.

For additional technical details and intuition, see Supplementary Note.

Curation of LOF variants

We obtained annotations for the consequences of all possible single-nucleotide changes to the hg19 reference genome from ref. 72. In ref. 72, the effects of variants on protein function were predicted using Variant Effect Predictor (VEP; v85)[73] using GENCODE (v19) gene annotations[74] as a reference. We defined a variant as an LOF if it was predicted by VEP to be a splice acceptor, splice donor or stop-gain variant. In ref. 72, predicted LOFs were further annotated using LOFTEE[12], which implements a series of filters to identify variants that may be misannotated (for example, LOFTEE considers predicted LOFs near the ends of transcripts as likely misannotations). For our analyses, we only kept predicted LOFs labeled as high confidence by LOFTEE, which are LOFs that passed all of LOFTEE's filters.

Next, we considered potential criteria for further filtering LOFs—cutoffs for the median exome sequencing read depth, cutoffs for the mean pext (proportion expressed across transcripts) score[72], whether to exclude variants that fall in segmental duplications or regions with low mappability[75] and whether to exclude variants flagged by LOFTEE as potentially problematic but that passed LOFTEE's primary filters.

We trained models with these filters one at a time and in combination and chose the model that had the best AUPRC in classifying essential from nonessential genes in mice. The filters we evaluated and chose for the final model are reported in Supplementary Table 4. Because we used mouse gene essentiality data to choose the filters, we do not further evaluate s_{het} on these data.

We considered genes to be essential in mice if they are heterozygous lethal, as determined by Karczewski et al.[12] using data from heterozygous knockouts reported in Mouse Genome Informatics[76]. We classify genes as nonessential if they are reported as Homozygous-Viable or Hemizygous-Viable by the International Mouse Phenotyping Consortium[77] (annotations downloaded on 8 December 2022 from https://www.ebi.ac.uk/mi/impc/essential-genes-search/).

Finally, we annotated each variant with its frequency in the gnomAD (v2.1.1) exomes[12], a dataset of 125,748 uniformly analyzed exomes that were largely curated from case–control studies of common adult-onset diseases. gnomAD provides precomputed allele frequencies for all variants that they call.

For potential LOFs that are not segregating, gnomAD does not release the number of individuals that were genotyped at those positions. For these sites, we used the median number of genotyped individuals at the positions for which gnomAD provides this information. We performed this separately on the autosomes and X chromosomes.

Data sources for the variant annotations, filters and frequencies, as well as additional information used to compute likelihoods, are listed in Supplementary Table 5.

Feature processing and selection

We compiled the following ten types of gene features from several sources: gene structure (for example, number of transcripts, number of exons and GC content), gene expression across tissues and cell lines, biological pathways and GO terms, PPI networks, co-expression networks, gene regulatory landscape (for example, number and properties of enhancers and promoters), conservation across species, protein embeddings, subcellular localization and missense constraint.

Additionally, we included an indicator variable that is 1 if the gene is in the nonpseudoautosomal region of the X chromosome and 0 otherwise.

For a description of the features within each category and where we acquired them, see Supplementary Note.

Training and validation

We fine-tuned a set of hyperparameters for our full empirical Bayes approach, using the best hyperparameters from an initial feature selection step (see Supplementary Note for description) as a starting point. To minimize overfitting, we split the genes into the following

three sets: a training set (chromosomes 7–22, X), a validation set for hyperparameter tuning (chromosomes 2, 4 and 6) and a test set to evaluate overfitting (chromosomes 1, 3 and 5). During each training iteration, one or more trees were added to the model to fit the gradient of the loss on the training set. We stopped model training once the loss on the validation set did not improve for ten iterations in a row (or the maximum number of iterations, 1,000, was reached). Using this approach, we performed a grid search over the hyperparameters listed in Supplementary Table 6 and used the combination with the lowest validation loss and best performance at classifying mouse essential genes (mean of the ranks on the two metrics).

Choosing OMIM genes

To identify genes that are considered constrained by s_{het} but not by LOEUF (Table 1), we filtered for genes with $s_{het} > 0.1$ (top ~15% most constrained genes, analogous to the recommended LOEUF cutoff of 0.35 (ref. 78), which corresponds to the top ~16% of genes) and LOEUF > 0.47 (least constrained ~75% of genes). Of these, we identified genes where heterozygous or hemizygous mutations that decrease the amount of functional protein (for example, LOF mutations) are associated with Mendelian disorders in the Online Mendelian Inheritance in Man (OMIM) database[70]. We chose genes for Table 1 primarily based on their prominence in the existing literature.

We define a gene as having a pathogenic variant in ClinVar if it contains a variant annotated with CLNSIG = Pathogenic. We downloaded ClinVar variants from https://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh38/ on 3 December 2023.

Evaluation of additional datasets

Definition of human essential and nonessential genes. We obtained data from 1,085 CRISPR knockout screens quantifying the effects of genes on cell survival or proliferation from the DepMap portal (22Q2 release)[24,25]. Scores from each screen are normalized such that nonessential genes identified by Hart et al.[79] have a median score of 0 and that common essential genes identified by Hart et al.[79] and Blomen et al.[80] have a median score of −1.

In classifying essential genes (Fig. 3a), we define a gene as essential if its score is <−1 in at least 25% of screens and as not essential if its score is >−1 in all screens. In classifying nonessential genes, we define a gene as nonessential if it has a minimal effect on growth in most cell lines (absolute effect <0.25 in at least 99% of screens) and as not nonessential if its score is <0 in all screens.

Definition of developmental disorder genes. Through the Deciphering Developmental Disorders (DDD) study[26], clinicians have annotated a subset of genes with the strength and nature of their association with developmental disorders. We classify genes as developmental disorder genes if they are annotated by the DDD study with confidence_category = definitive and allelic_requirement = monoallelic_autosomal, monoallelic_X_hem (hemizygous), or monoallelic_X_het (heterozygous).

We classify genes as not associated with developmental disorders if they are annotated by the DDD study, do not meet the abovementioned criteria for association with a disorder and are not annotated with confidence_category = strong, moderate or limited and allelic_requirement = monoallelic_autosomal, monoallelic_X_hem or monoallelic_X_het.

We downloaded genes with DDD annotations from https://www.deciphergenomics.org/ddd/ddgenes on 19 November 2023.

Enrichment/depletion of HPO genes. The HPO provides a structured organization of phenotypic abnormalities and the genes associated with them, with each HPO term corresponding to a phenotypic abnormality. We calculated the enrichment of constrained genes in each HPO term with at least 200 genes as the ratio (fraction of HPO genes under constraint)/(fraction of background genes under constraint). We defined genes under constraint to be the decile of genes considered most constrained by s_{het} or LOEUF. To choose background genes, we sampled from the set of all genes to match each HPO term's distribution of expected unique LOFs. Similarly, we calculated the depletion of unconstrained genes in each HPO term as the ratio (fraction of HPO genes not under constraint)/(fraction of background genes not under constraint), where we define genes not under constraint to be the decile of genes considered least constrained by s_{het} or LOEUF.

We downloaded HPO phenotype-to-gene annotations from http://purl.obolibrary.org/obo/hp/hpoa/phenotype_to_genes.txt on 27 January 2023.

Enrichment of DNMs in patients with developmental disorders. We used the enrichment metric developed by Kaplanis et al.[5] in their analysis of DNMs identified from the exome sequencing of 31,058 patients with developmental disorders and their unaffected parents. Enrichment of DNMs in patients with developmental disorders was calculated as the ratio of observed DNMs in patients over the expected number under a null mutational model that accounts for the study sample size and triplet mutation rate at the mutation sites[81].

For Fig. 3d, we calculated the enrichment of DNMs in constrained genes, defined as the decile of genes considered most constrained by s_{het} or LOEUF. For Extended Data Fig. 4d, we calculated the enrichment of DNMs in constrained genes with and without known associations with development disorders. We defined a gene as having a known association if it is annotated by the DDD study ('Definition of developmental disorder genes') with confidence_category = definitive or strong and allelic_requirement = monoallelic_autosomal, monoallelic_X_hem (hemizygous) or monoallelic_X_het (heterozygous).

For each set of genes, we computed the mean enrichment over sites and 95% Poisson confidence intervals for the mean using the code provided in ref. 5.

Heritability enrichment in constrained genes. We computed the heritability enrichment in the top 10% of genes constrained by s_{het} or LOEUF using stratified LD score regression (S-LDSC)[82]. To do this, we divided the heritability enrichment in constrained genes as reported by S-LDSC by the heritability enrichment in all genes. We linked variants to genes if they were in or within 100 kb of the gene body, and ran S-LDSC using 1000G EUR Phase3 genotype data to estimate LD scores, baseline v2.2 annotations and HapMap 3 SNPs excluding the major histocompatibility complex region as regression SNPs. We performed this analysis using summary statistics from 438 traits in the UK Biobank (downloaded from https://nealelab.github.io/UKBB_ldsc) with highly statistically significant SNP heritability (LDSC z score > 7, the threshold recommended in ref. 82).

Expression variability across species. To understand the variability in expression between humans and other species, we focused on gene expression differences between humans and chimpanzees as estimated from RNA-seq of an in vitro model of the developing cerebral cortex for each species[28]. As a metric of variability between the two species, we used the absolute log fold change (LFC) in gene expression between human and chimpanzee cortical spheroids, which was calculated from samples collected at several time points throughout the differentiation of the spheroids. LFC estimates were obtained from Supplementary Table 9 of ref. 28.

To visualize the relationship between constraint and absolute LFC, we plotted a LOESS curve between the constraint on a gene (gene rank from least to most constrained using either s_{het} or LOEUF as the constraint metric) and the absolute LFC for the gene. Curves were calculated using the LOWESS function from the statsmodels package with parameters frac = 0.15 and $\delta = 10$.

Expression variability across individuals. To calculate a measure of expression variance across Genotype-Tissue Expression (GTEx) samples, we log-transformed the per-gene mean and variance of gene expression levels (where expression is in units of transcripts per million) and used the residuals from LOESS regression of the transformed expression variance on the transformed mean expression. LOESS regression was computed using the LOWESS function from the statsmodels package with parameters frac = 0.1 and $\delta = 0$. This procedure reduces the correlation between mean expression and expression variance (Spearman $\rho = 0.02$ between mean expression and residual variance, compared to Spearman $\rho = 0.90$ between mean expression and variance before regression). We calculated expression variance using 17,398 RNA-seq samples in the GTEx (v8) release[29] (838 donors and 52 tissues/cell lines) for all genes with a median TPM of ≥5. LOESS curves for visualization were computed as in 'Expression variability across species'.

Feature interpretation

Training models on feature subsets. We grouped features into categories (see Supplementary Table 8 for the features in each category) and trained a model for each category to predict s_{het} from the corresponding features. For each model, we tuned hyperparameters over a subset of the values we considered for the full model (Supplementary Table 7) and chose the combination of hyperparameters that minimized the loss over genes in the validation set. As a baseline, we trained a model with no features, such that all genes have a shared prior distribution that is learned from the LOF data—this model is analogous to a standard empirical Bayes model.

Definition of expression feature subsets. We grouped gene expression features into 24 categories representing tissues, cell types and developmental stages using terms present in the feature names (Supplementary Table 3).

Scoring individual features. To score individual gene features, we varied the value of one feature at a time and calculated the variance in predicted s_{het} as a feature score. In more detail, we fixed each feature to values spanning the range of observed values for that feature (0th, 2nd, ..., 98th and 100th percentiles), such that all genes shared the same feature value. Then, for each of these 51 feature values, we averaged the s_{het} values predicted by the learned priors over all genes, where the predicted s_{het} for each gene is the mean of its prior. We denote this averaged prediction by $s_{het}^{(f)}(p)$ for some feature f and percentile p. Finally, we define the score for feature f as $\text{score}_f = \text{s.d.}(s_{het}^{(f)}(0), s_{het}^{(f)}(2), \ldots, s_{het}^{(f)}(98), s_{het}^{(f)}(100))$, where s.d. is a function computing the sample standard deviation. In other words, a feature with a high score is one for which varying its value causes high variance in the predicted s_{het}.

For the lineplots in Fig. 4c–f, we scale the predictions $s_{het}^{(f)}(p)$ for each feature f by subtracting $(s_{het}^{(f)}(0) + s_{het}^{(f)}(100))/2$ from each prediction.

Pruning features before computing feature scores. While investigating the effects of features on predicted s_{het}, we found that including highly correlated features in the model could produce unintuitive results, such as opposite correlations with s_{het} for highly similar features. Therefore, for Fig. 4c–f, we first pruned the set of features to minimize pairwise correlations between the remaining features. To do this, we randomly kept one feature in each group of correlated features, where such a group is defined as a set of features where each feature in the set has an absolute Spearman $\rho > 0.7$ to some other feature in the set.

For Fig. 4c–f, we trained models on the relevant features in this pruned set (GO, network, gene regulatory and gene structure features, respectively).

Reporting summary

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

Data availability

Posterior means and 95% credible intervals for s_{het} are available in Supplementary Table 1. Data sources for pLOF annotations, CpG methylation levels, exome sequencing coverage, variant frequencies and mappability/segmental duplication annotations are available in Supplementary Table 5. A description of the gene features is available in Supplementary Table 8. Posterior densities for s_{het}, likelihoods for s_{het}, LOF variants with misannotation probabilities and gene feature tables are available in ref. 83. Additional publicly available datasets used in this study are described in Methods and Supplementary Information and are accessible at IMPC essential genes (https://www.ebi.ac.uk/mi/impc/essential-genes-search/); pLOF annotations (gs://gnomad-public/papers/2019-tx-annotation/pre_computed/all.possible.snvs.tx_annotated.GTEx.v7.021520.tsv); mean methylation for CpG sites (gs://gcp-public-data–gnomad/resources/methylation); exome sequencing coverage (gs://gcp-public-data–gnomad/release/2.1/coverage/exomes/gnomad.exomes.coverage.summary.tsv.bgz); variant frequencies (gs://gcp-public-data–gnomad/release/2.1.1/vcf/exomes/gnomad.exomes.r2.1.1.sites.vcf.bgz); low mappability and segmental duplications (https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/genome-stratifications/v3.1/GRCh37/Union/GRCh37_alllowmapandsegdupregions.bed.gz); ClinVar variants (https://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh38/); DepMap 22Q2 release (https://depmap.org/portal/download/all/); DDD annotations (https://www.deciphergenomics.org/ddd/ddgenes); HPO phenotype-to-gene annotations (http://purl.obolibrary.org/obo/hp/hpoa/phenotype_to_genes.txt); DNMs from developmental disorder patients[5]; UK Biobank summary statistics (https://nealelab.github.io/UKBB_ldsc); RNA-seq from chimpanzee/human cortical models[28]; GTEx v8 release[29].

Code availability

GeneBayes and code for estimating s_{het} are available at https://github.com/tkzeng/GeneBayes and in ref. 84. Analysis code is available in ref. 85. All analyses were performed using Python v3.8, Python v3.9 or R v4.2. To train models, we used a modified version of NGBoost (v0.3.12)[16,86] (https://github.com/tkzeng/ngboost), XGBoost (v2.0.2)[87] and PyTorch (v1.12.1)[88]. Likelihoods were computed with fastDTWF (v.0.0.3)[15] (https://github.com/jeffspence/fastDTWF). For hyperparameter tuning, we used shap-hypetune v0.2 (https://github.com/cerlymarco/shap-hypetune). For heritability enrichment analyses, we used ldsc (v1.0.1)[89]. For additional analyses, we used NumPy (v1.26.0)[90], SciPy (v1.8.1)[91], Pandas (v2.1.3)[92], Scikit-learn (1.3.0)[93] and Statsmodels (v0.14.0)[94].

References

71. Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.* **46**, 919–925 (2014).
72. Cummings, B. B. et al. Transcript expression-aware annotation improves rare variant interpretation. *Nature* **581**, 452–458 (2020).
73. McLaren, W. et al. The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
74. Frankish, A. et al. GENCODE: reference annotation for the human and mouse genomes in 2023. *Nucleic Acids Res.* **51**, D942–D949 (2023).
75. Olson, N. D. et al. PrecisionFDA Truth Challenge V2: calling variants from short and long reads in difficult-to-map regions. *Cell Genom.* **2**, 100129 (2022).
76. Blake, J. A. et al. Mouse Genome Database (MGD): knowledgebase for mouse–human comparative biology. *Nucleic Acids Res.* **49**, D981–D987 (2021).

77. Groza, T. et al. The International Mouse Phenotyping Consortium: comprehensive knockout phenotyping underpinning the study of human disease. *Nucleic Acids Res.* **51**, D1038–D1045 (2023).
78. Gudmundsson, S. et al. Variant interpretation using population databases: lessons from gnomAD. *Hum. Mutat.* **43**, 1012–1030 (2022).
79. Hart, T., Brown, K. R., Sircoulomb, F., Rottapel, R. & Moffat, J. Measuring error rates in genomic perturbation screens: gold standards for human functional genomics. *Mol. Syst. Biol.* **10**, 733 (2014).
80. Blomen, V. A. et al. Gene essentiality and synthetic lethality in haploid human cells. *Science* **350**, 1092–1096 (2015).
81. Samocha, K. E. et al. A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* **46**, 944–950 (2014).
82. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
83. Zeng, T., Spence, J. P., Mostafavi, H. & Pritchard, J. K. s_het estimates from GeneBayes and other supplementary datasets. *Zenodo* https://doi.org/10.5281/zenodo.10403680 (2023).
84. Zeng, T. tkzeng/GeneBayes: GeneBayes v1.0. *Zenodo* https://doi.org/10.5281/zenodo.10939506 (2024).
85. Zeng, T. Code and data to reproduce GeneBayes figures. *Zenodo* https://doi.org/10.5281/zenodo.11141460 (2024).
86. Schuler, A. et al. tkzeng/ngboost: NGBoost for GeneBayes v1.0. *Zenodo* https://doi.org/10.5281/zenodo.10944711 (2024).
87. Chen, T. & Guestrin, C. Xgboost: a scalable tree boosting system. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (Association for Computing Machinery, 2016).
88. Paszke, A. et al. Pytorch: an imperative style, high-performance deep learning library. In *Proc. Advances in Neural Information Processing Systems* (eds Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F. & Fox, E. B.) 32 (Curran Associates Inc., 2019).
89. Finucane, H. K. et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621–629 (2018).
90. Harris, C. R. et al. Array programming with NumPy. *Nature* **585**, 357–362 (2020).
91. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
92. Van der Walt, S. & Millman, J. (eds). Data structures for statistical computing in Python. In *Proc. 9th Python in Science Conference* 56–61 (SciPy, 2010).
93. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
94. Van der Walt, S. & Millman, J. (eds). Statsmodels: econometric and statistical modeling with Python. In *Proc. 9th Python in Science Conference* 92–96 (SciPy, 2010).

Acknowledgements

We would like to thank I. Agarwal, M. Przeworski, J. Engreitz and members of the Pritchard Lab for valuable feedback and discussions. This work was supported by the National Institutes of Health (NIH; grants R01HG011432, R01HG008140 and U01HG009431 to J.K.P. and R01AG066490 to S. Montgomery). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the paper.

Author contributions

J.P.S., H.M. and J.K.P. conceived and designed the study. T.Z. and J.P.S. performed all data analyses and developed the model. H.M. provided intellectual contributions to all aspects of the study. T.Z., J.P.S., H.M. and J.K.P. wrote the paper. J.K.P. supervised the study and acquired funding.

Competing interests

The authors declare no competing interests.

Additional information

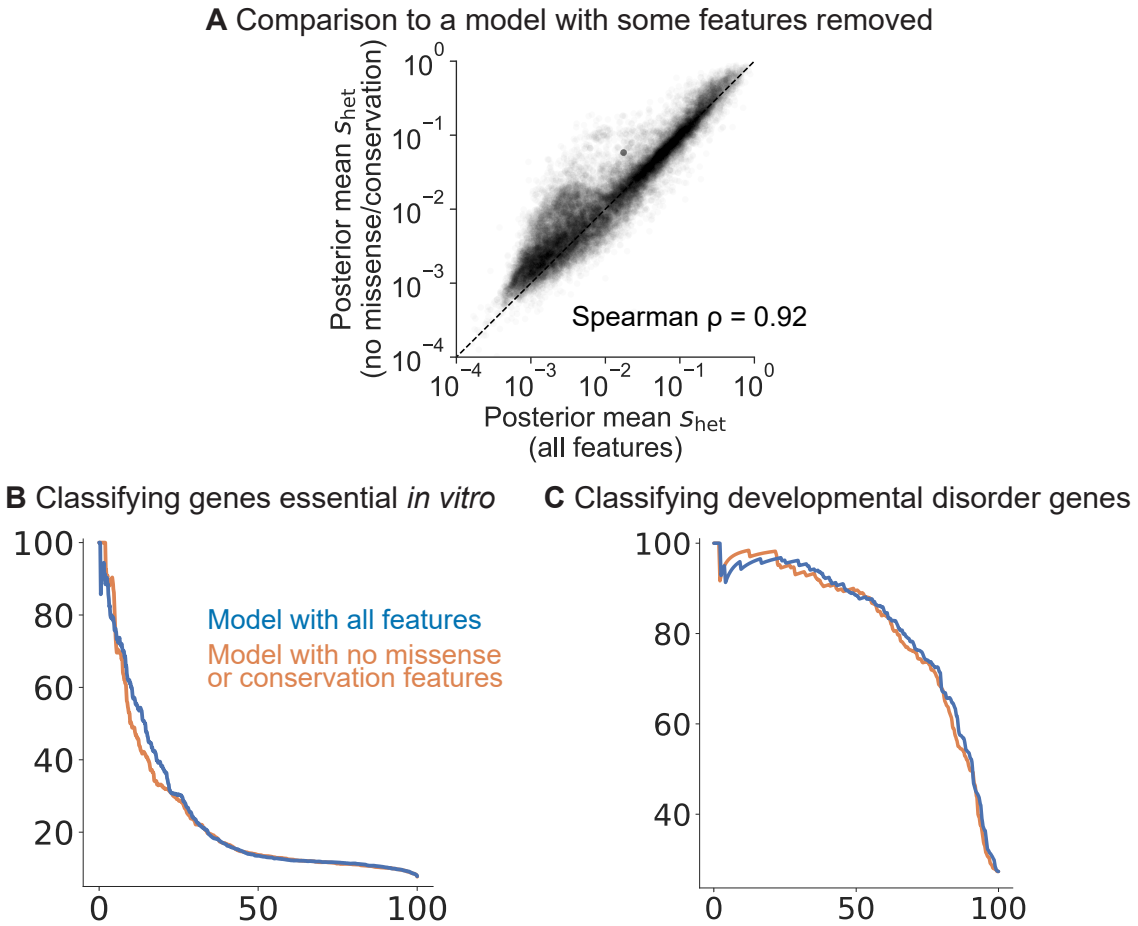
Extended data is available for this paper at https://doi.org/10.1038/s41588-024-01820-9.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-024-01820-9.

Correspondence and requests for materials should be addressed to Tony Zeng, Jeffrey P. Spence or Jonathan K. Pritchard.

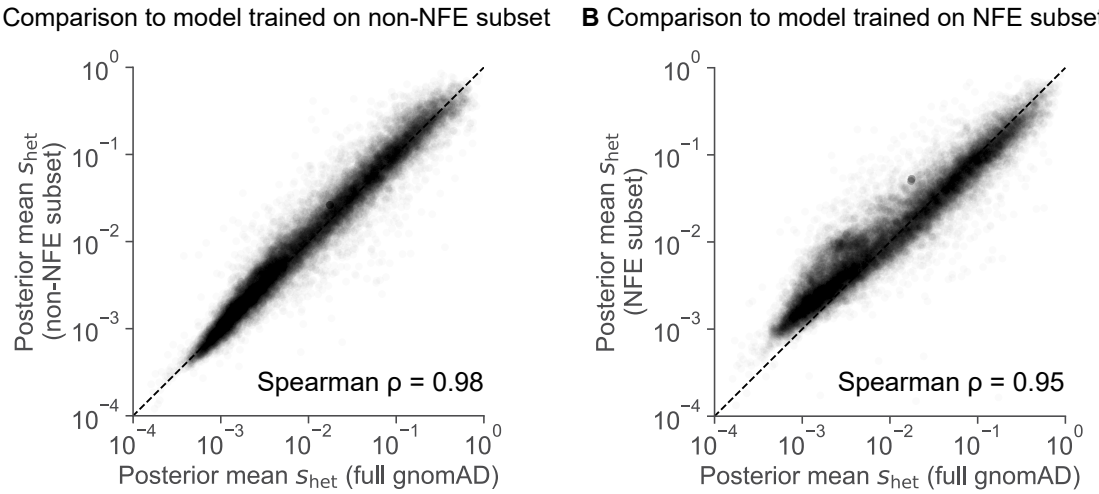
Peer review information *Nature Genetics* thanks Zilin Li and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

https://doi.org/10.1038/s41588-024-01820-9



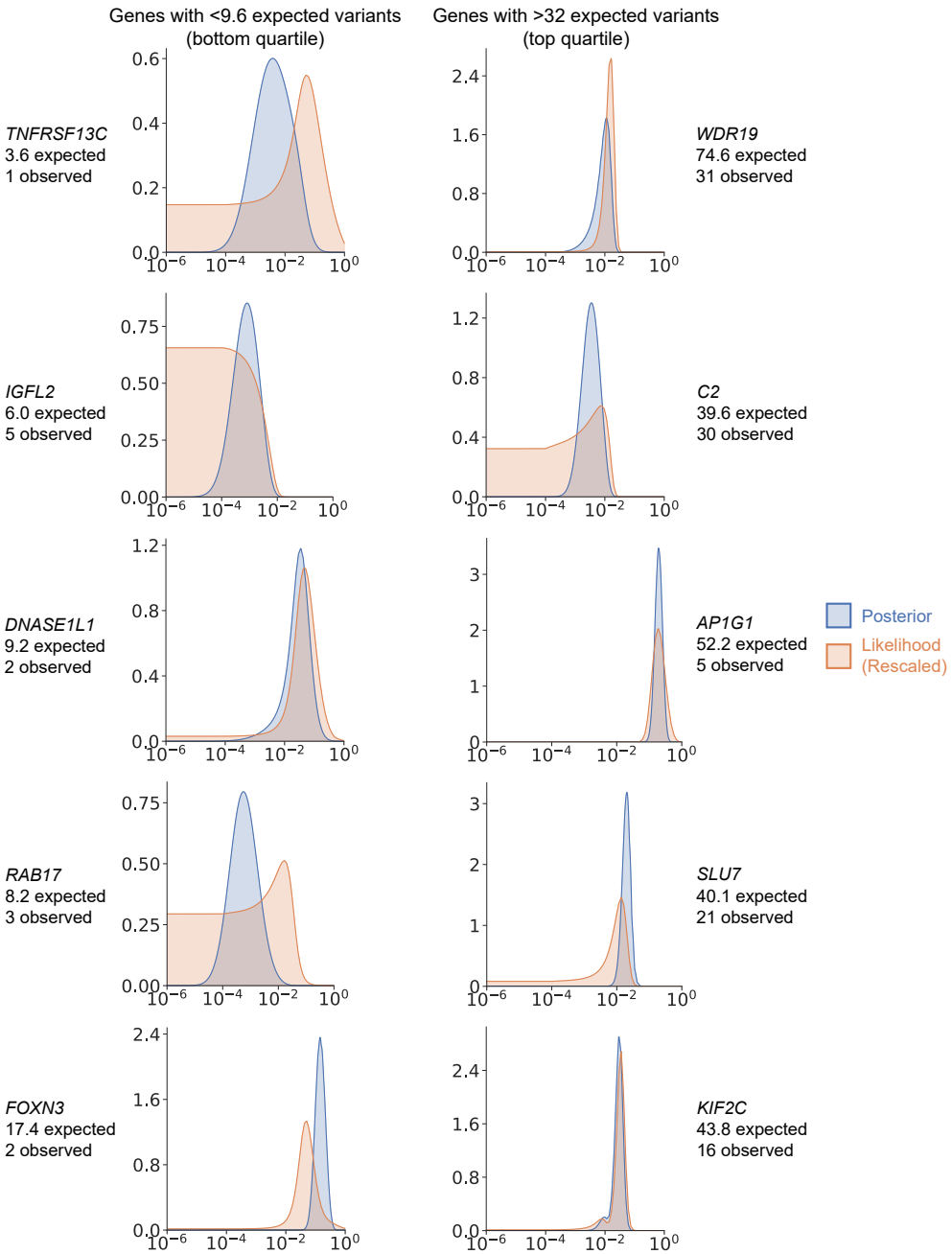
Extended Data Fig. 1 | Performance of s_{het} estimates from a model with some features removed. a, Scatterplot of posterior mean s_{het} estimated from a model trained without missense constraint or cross-species conservation features (y axis) against s_{het} estimated from the full model (x axis). **b**, Precision–recall curves comparing the performance of s_{het} estimated from the full model (blue) and from the model without missense/conservation features (orange) in classifying essential genes. **c**, Precision–recall curves comparing the performance of s_{het} estimated from the two models in classifying developmental disorder genes.

https://doi.org/10.1038/s41588-024-01820-9



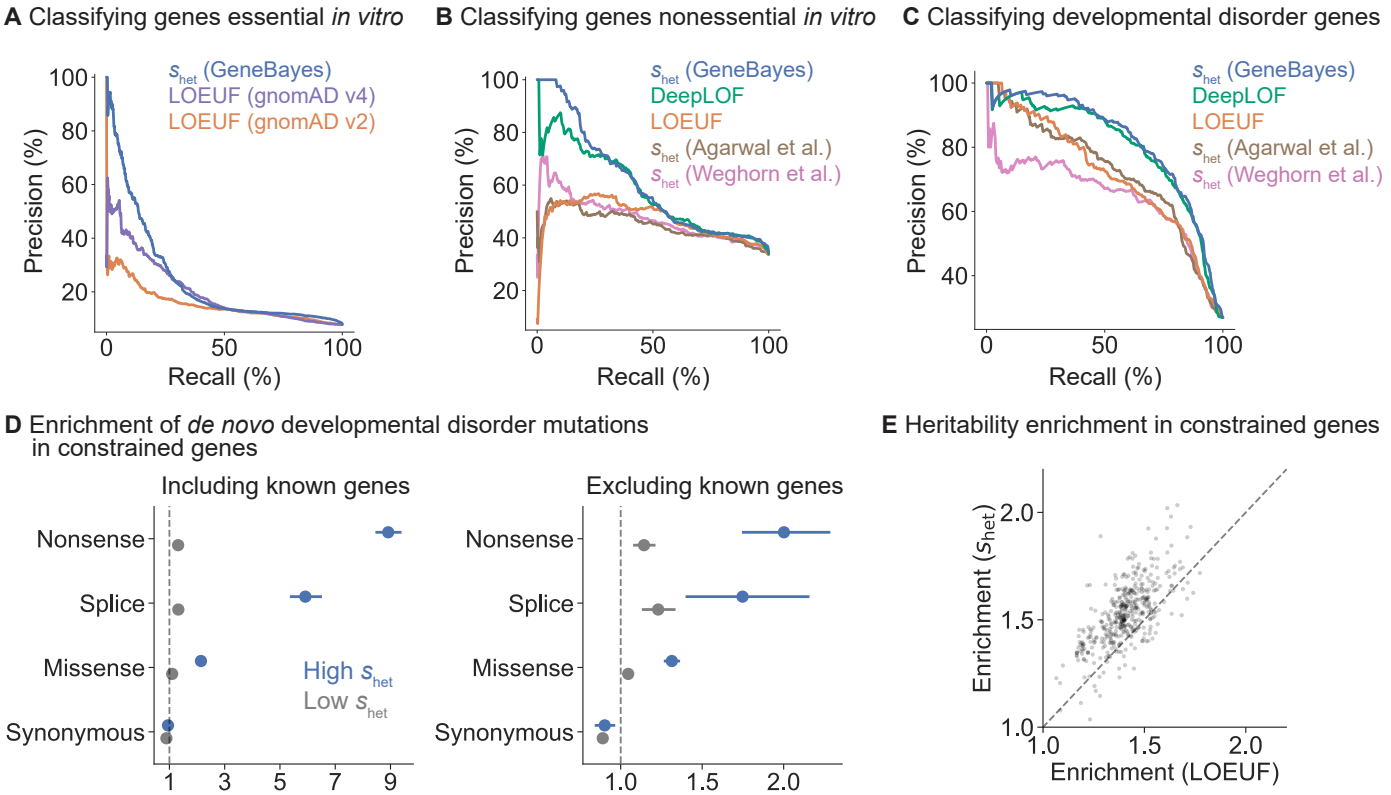
Extended Data Fig. 2 | Comparison of s_{het} estimates from models trained on subsets of gnomAD. a, Scatterplot of posterior mean s_{het} estimated from a model trained with non-NFE individuals (y axis) against s_{het} estimated from the full model (x axis). NFE, Non-Finnish European. This subset consists of 56,000 individuals or 45% of the total dataset. **b**, Scatterplot of posterior mean s_{het} estimated from a model trained with NFE individuals (y axis) against s_{het} estimated from the full model (x axis). This subset consists of 67,000 individuals or 55% of the total dataset.

https://doi.org/10.1038/s41588-024-01820-9



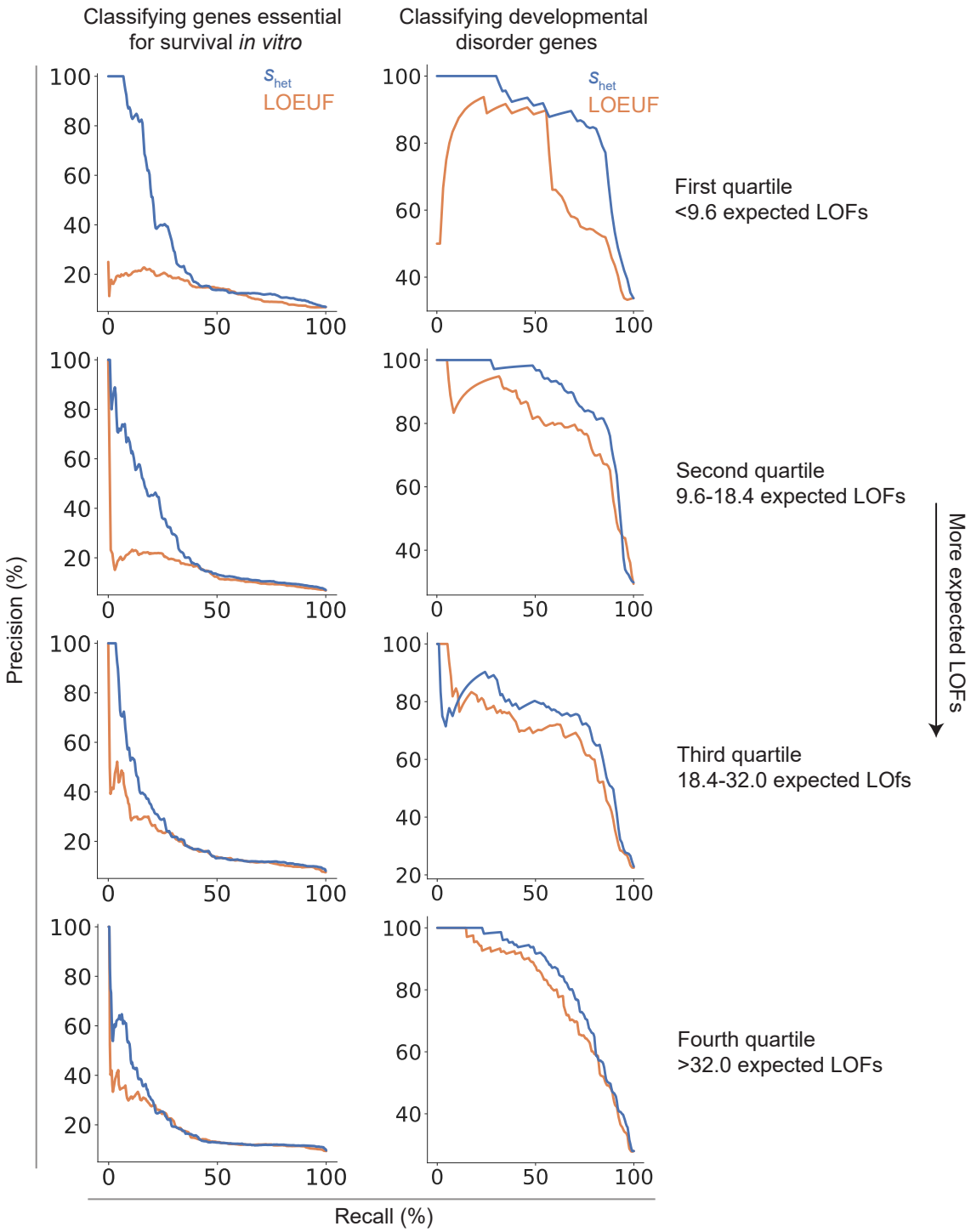
Extended Data Fig. 3 | s_{het} distributions for additional example genes. Left: posterior distributions and rescaled likelihoods for genes with few expected LOFs (genes in the bottom quartile). Right: posterior distributions and rescaled likelihoods for genes with many expected LOFs (genes in the top quartile).

https://doi.org/10.1038/s41588-024-01820-9



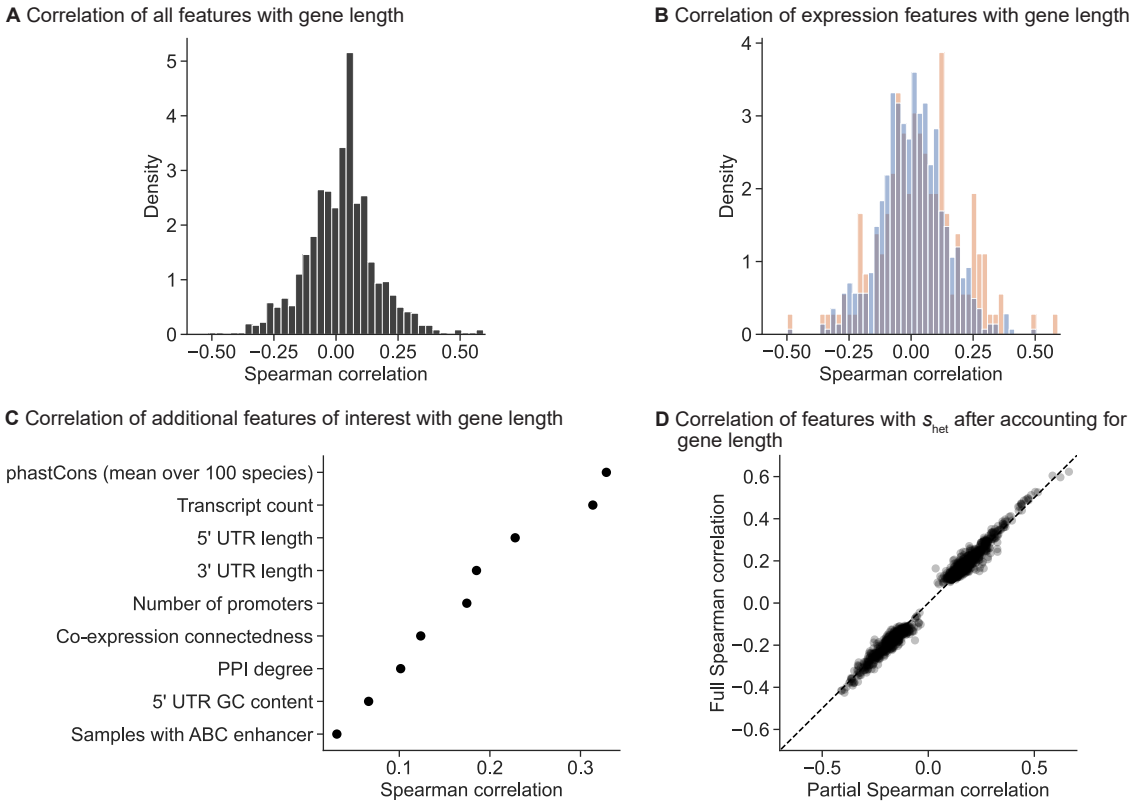
Extended Data Fig. 4 | Additional validation analyses. a, Precision–recall curves comparing the performance of s_{het} estimates from GeneBayes against LOEUF from gnomAD v4.0.0 (731k exomes) or LOEUF from gnomAD v2.1.1 (125k exomes) in classifying essential genes. **b**, Precision–recall curves comparing the performance of s_{het} estimates from GeneBayes against other constraint metrics in classifying nonessential genes. **c**, Precision–recall curves comparing the performance of s_{het} against other constraint metrics in classifying developmental disorder genes. **d**, Enrichment of de novo mutations in patients with developmental disorders, calculated as the observed number of mutations over the expected number under a null mutational model (n = 31,058 parent–offspring trios). We plot the enrichment of synonymous, missense, splice and nonsense variants in the 10% of genes considered most constrained by s_{het} (blue) and the enrichment of these variants in all other genes (gray), including (left) and excluding (right) known developmental disorder genes. Bars represent 95% confidence intervals, centered around the mean. **e**, Scatterplot of the enrichment of common variant heritability in the 10% of genes considered most constrained by s_{het} (y axis) or LOEUF (x axis), normalized by the enrichment of heritability in all genes. Each point represents one trait.

https://doi.org/10.1038/s41588-024-01820-9



Extended Data Fig. 5 | Performance of s_{het} and LOEUF for genes with differing numbers of expected LOFs. Left: precision–recall curves comparing the performance of s_{het} against LOEUF in classifying essential genes for groups of genes binned by their expected number of LOFs. Right: precision–recall curves comparing the performance of s_{het} against LOEUF in classifying developmental disorder genes for binned genes.

https://doi.org/10.1038/s41588-024-01820-9



Extended Data Fig. 6 | Correlation of gene features with gene length. **a**, Histogram of the Spearman ρ between gene features and coding sequence (CDS) length. **b**, Histogram of the Spearman ρ between gene features and CDS length for gene expression features, colored by category. **c**, Spearman ρ between gene features and CDS length for additional features of interest. **d**, Scatterplot of the Spearman ρ between gene features and posterior mean s_{het} (y axis) against the partial Spearman ρ (x axis) after controlling for the effect of gene (CDS) length.

nature portfolio

Corresponding author(s): Tony Zeng, Jeffrey P. Spence, Jonathan K. Pritchard

Last updated by author(s): Apr 8, 2024

Reporting Summary

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

Statistics

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

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

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

Software and code

Policy information about availability of computer code

Data collection	No software was used in the collection of data.
Data analysis	All analyses were performed using Python v3.8, Python v3.9, or R v4.2. To train models, we used a modified version of NGBoost v0.3.12 (https://github.com/tkzeng/ngboost), XGBoost v2.0.2, and PyTorch v1.12.1. Likelihoods were computed with fastDTWF v.0.0.3 (https://github.com/jeffspence/fastDTWF). For hyperparameter tuning, we used shap-hypetune v0.2. Model training code is released as GeneBayes v1.0 (https://github.com/tkzeng/GeneBayes). For heritability enrichment analyses, we used ldsc v1.0.1. For additional analyses, we used NumPy v1.26.0, SciPy v1.8.1, pandas v2.1.3, scikit-learn 1.3.0, and statsmodels v0.14.0.

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

Data

Policy information about availability of data

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

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Posterior means and 95% credible intervals for s_het are available in Supplementary Table 1. Data sources for pLOF annotations, CpG methylation levels, exome sequencing coverage, variant frequencies, and mappability/segmental duplication annotations are available in Supplementary Table 5. A description of the gene features is available in Supplementary Table 8. Posterior densities for s_het, likelihoods for s_het, LOF variants with misannotation probabilities, and gene feature tables are available at: https://zenodo.org/records/10403680. Additional publicly available datasets used in this study are described in Methods and Supplementary Information, and are accessible at: IMPC essential genes: https://www.ebi.ac.uk/mi/impc/essential-genes-search; pLOF annotations: gs://gnomad-public/papers/2019-tx-annotation/pre_computed/all.possible.snvs.tx_annotated.GTEx.v7.021520.tsv; Mean methylation for CpG sites: gs://gcp-public-data--gnomad/resources/methylation; Exome sequencing coverage: gs://gcp-public-data--gnomad/release/2.1/coverage/exomes/gnomad.exomes.coverage.summary.tsv.bgz; Variant frequencies: gs://gcp-public-data--gnomad/release/2.1.1/vcf/exomes/gnomad.exomes.r2.1.1.sites.vcf.bgz; Low mappability and segmental duplications: https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/genome-stratifications/v3.1/GRCh37/Union/GRCh37_alllowmapandsegdupregions.bed.gz; ClinVar variants: https://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh38/; DepMap 22Q2 release: https://depmap.org/portal/download/all/; DDD annotations: https://www.deciphergenomics.org/ddd/ddgenes; HPO phenotype-to-gene annotations: http://purl.obolibrary.org/obo/hp/hpoa/phenotype_to_genes.txt; De novo mutations from developmental disorder patients: Kaplanis et al. 2020 Nature; UK Biobank summary statistics: https://nealelab.github.io/UKBB_ldsc; RNA-seq from chimp/human cortical models: Agoglia et al. 2021 Nature; GTEx v8 release: The Gtex Consortium 2020 Science.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	N/A
Reporting on race, ethnicity, or other socially relevant groupings	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

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

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size	We used data from the gnomAD consortium (v2.1.1 exomes), which consists of 125,748 uniformly-analyzed exomes. No methods were used to predetermine sample size; the sample size was determined by the availability of aggregate exome sequencing data prior to publication of Karczewski et al. 2020 Nature. This sample size has previously facilitated useful estimates of constraint for many genes (Karczewski et al. 2020 Nature, Lek et al. 2016 Nature).
Data exclusions	Samples were excluded from gnomAD if they had lower sequencing quality, were from related individuals, had inadequate consent for release, or were from individuals with severe childhood-onset disease or their close relatives (see Karczewski et al. 2020 Nature for details). We additionally filtered out pLOFs that may be misannotated: pLOFs not annotated as HC by LOFTEE, expressed in too few tissues, that fall in a segmental duplication or low mappability region, or that are flagged as potentially problematic (see Methods for details).
Replication	We trained models on subsets of individuals from gnomAD (non-NFE and NFE individuals) and on subsets of features (Extended Data Fig. 1) to show that our method is robust to the included data/features. We performed bootstrapping over genes to demonstrate the robustness of our method's improved performance over other methods on independent benchmarking tasks (classifying essential and developmental disorder genes). We further evaluated our estimates of s_het on additional benchmarking metrics, including the enrichment of constrained genes in HPO terms and the enrichment of common variant heritability in constrained genes.

Randomization	Any publicly available data for which randomization may have been relevant were collected prior to this study. None of our analyses relied on causal interpretations of exogenous treatments, so randomization is not relevant for the purposes of this study.
Blinding	Any publicly available data for which blinding may have been relevant were collected prior to this study. As we did not collect any data, blinding is not relevant for the purposes of this study.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

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

Plants

Seed stocks	N/A
Novel plant genotypes	N/A
Authentication	N/A