BMC Bioinformatics

RESEARCH

Open Access

An unsupervised deep learning framework for predicting human essential genes from population and functional genomic data

Troy M. LaPolice[1,2,3*] and Yi-Fei Huang[1,3*]

*Correspondence:
troy.lapolice@psu.edu;
yuh371@psu.edu

[1] Department of Biology, Pennsylvania State University, University Park, PA 16802, USA
[2] Bioinformatics and Genomics Graduate Program, Pennsylvania State University, University Park, PA 16802, USA
[3] Huck Institutes of the Life Sciences, Pennsylvania State University, University Park, PA 16802, USA

Abstract

Background: The ability to accurately predict essential genes intolerant to loss-of-function (LOF) mutations can dramatically improve the identification of disease-associated genes. Recently, there have been numerous computational methods developed to predict human essential genes from population genomic data. While the existing methods are highly predictive of essential genes of long length, they have limited power in pinpointing short essential genes due to the sparsity of polymorphisms in the human genome.

Results: Motivated by the premise that population and functional genomic data may provide complementary evidence for gene essentiality, here we present an evolution-based deep learning model, DeepLOF, to predict essential genes in an unsupervised manner. Unlike previous population genetic methods, DeepLOF utilizes a novel deep learning framework to integrate both population and functional genomic data, allowing us to pinpoint short essential genes that can hardly be predicted from population genomic data alone. Compared with previous methods, DeepLOF shows unmatched performance in predicting ClinGen haploinsufficient genes, mouse essential genes, and essential genes in human cell lines. Notably, at a false positive rate of 5%, DeepLOF detects 50% more ClinGen haploinsufficient genes than previous methods. Furthermore, DeepLOF discovers 109 novel essential genes that are too short to be identified by previous methods.

Conclusion: The predictive power of DeepLOF shows that it is a compelling computational method to aid in the discovery of essential genes.

Keywords: Deep Learning, Unsupervised, Essential Genes, Loss of Function Intolerance, Population Genomics, Functional Genomics

Introduction

Loss-of-function (LOF) mutations, including stop-gain, splice-site, and frameshift mutations, play a key role in the etiology of genetic disorders (Fig. 1a). While it is relatively straightforward to identify LOF mutations in protein-coding genes, it is challenging to infer their effects on evolutionary fitness and disease risk. Several computational methods [1–7] have recently been developed to predict human essential genes based on the

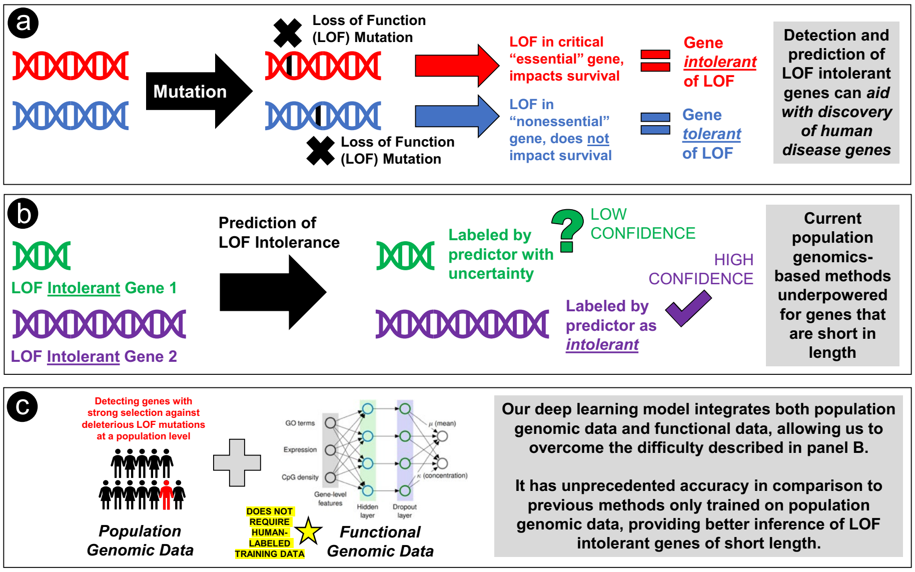
a
Loss of Function (LOF) Mutation
Mutation
LOF in critical "essential" gene, impacts survival
Gene intolerant of LOF
LOF in "nonessential" gene, does not impact survival
Gene tolerant of LOF
Loss of Function (LOF) Mutation
Detection and prediction of LOF intolerant genes can aid with discovery of human disease genes
b
LOF Intolerant Gene 1
LOF Intolerant Gene 2
Prediction of LOF Intolerance
Labeled by predictor with uncertainty
LOW CONFIDENCE
HIGH CONFIDENCE
Labeled by predictor as intolerant
Current population genomics-based methods underpowered for genes that are short in length
c
Detecting genes with strong selection against deleterious LOF mutations at a population level
Population Genomic Data
GO terms
Expression
CpG density
Gene-level features
Hidden layer
Dropout layer
μ (mean)
κ (concentration)
DOES NOT REQUIRE HUMAN-LABELED TRAINING DATA
Functional Genomic Data
Our deep learning model integrates both population genomic data and functional data, allowing us to overcome the difficulty described in panel B.
It has unprecedented accuracy in comparison to previous methods only trained on population genomic data, providing better inference of LOF intolerant genes of short length.

Fig. 1 Overview of the background information, motivation and methods behind DeepLOF. **a** *Background*: Introduction to loss of function (LOF) mutations, essential genes versus nonessential genes as well as LOF intolerance versus LOF tolerance. *Motivation*: Determining which genes are LOF intolerant can aid with discovery of human disease genes. **b** *Motivation*: The limitation of current population genomics-based methods for determining LOF intolerance is they are underpowered when predicting genes that are short in length. **c** Simple overview of the concept behind DeepLOF. *Methods*: Our method integrates a population genomics-based approach with a functional genomics approach, providing unparalleled ability to predict LOF intolerance, particularly in short genes. DeepLOF does not require human-labeled training data and thus, may not suffer from label leakage

premise that LOF mutations causing early-onset disorders may be subject to negative selection in human populations [8, 9]. Based on large-scale population genomic data, such as gnomAD [5], these methods seek to identify LOF-intolerant genes where the observed number of LOF variants is significantly smaller than the expected number under a neutral mutation model. The rationale behind these prediction methods is that essential genes, with minimal inter-individual variation, are subject to purifying selection in order to maintain their sequence in the population, as alterations would lead to reduced fitness. It has been shown that LOF-intolerant genes predicted by these methods are enriched with haploinsufficient genes associated with Mendelian disorders [1–7]. Furthermore, *de novo* LOF mutations in probands with autism [10, 11], schizophrenia [12, 13], and severe developmental disorders [14] are significantly overrepresented in LOF-intolerant genes. Therefore, population genetics-based prediction of LOF-intolerant genes is a powerful strategy to discover haploinsufficient genes associated with human disease.

However, despite the recent success of population genetics-based gene essentiality prediction, the statistical power of existing methods may heavily depend on the length of a gene [5, 9, 15]. Specifically, a long gene typically has a large expected number of LOF variants under a neutral mutation model. Thus, when we compare the observed number of LOF variants with the expected one, it is relatively easy to reject the null hypothesis of neutral evolution in a long gene. In contrast, a short gene is expected to have only a handful of LOF variants. Therefore, in a short gene it is difficult to distinguish the depletion of LOF mutations caused by negative selection from that by chance alone, which may hinder the discovery of many essential genes of short length in the human genome (Fig. 1b).

Complementary to population genomic data that manifest natural selection at the organism level, functional genomic assays, such as RNA-seq and ChIP-seq, provide rich information on the molecular functions of protein-coding genes. Thus, functional genomic data may also be utilized to predict gene essentiality. Based on this idea, several supervised methods have been developed to predict essential genes from functional genomic features [15–20]. One such method, DEEPLYESSENTIAL for example, utilizes functional features that include gene length, codon frequency and codon adaptation index to determine essentiality [19]. Other models use features such as CpG density [15], gene ontology terms [17], and epigenomic features [20] among other functional data to predict intolerance. Unlike population genetics-based methods, the predictive power of genomic feature-based methods may not heavily depend on the length of a gene. However, because functional genomic data are often from cell lines, gene scores solely derived from functional genomic features may not always be indicative of gene essentiality at the whole organism level.

We propose that integrating population and functional genomic data may improve gene essentiality prediction. To this end, we introduce DeepLOF, an evolution-based deep learning model for predicting human genes intolerant to LOF mutations. By combining a deep neural network and a population genetics-based likelihood function, DeepLOF can integrate genomic features and population genomic data to predict LOF-intolerant genes without human-labeled training data (Fig. 1c). Thus, DeepLOF may not suffer from label leakage and other pitfalls of supervised machine learning [21]. Compared to previous methods, DeepLOF shows unmatched performance in predicting ClinGen haploinsufficient genes [22], human orthologs of mouse essential genes [23], and genes essential to the survival of cell lines [24]. Furthermore, using DeepLOF we identify 109 LOF-intolerant genes of short length missed by previous methods. The 109 novel LOF-intolerant genes are enriched with essential genes and are depleted in benign genomic deletions. Taken together, DeepLOF is a powerful deep learning framework to predict essential genes in the human genome.

Methods

Details of the DeepLOF model

We denote η_i as the relative rate of observed LOF variants in gene i with respect to the expected number of LOF variants under a neutral mutation model. In DeepLOF, we seek to estimate the distribution of η_i from both genomic features and population genomic data. To this end, the DeepLOF model combines a feedforward neural network transforming genomic features and a likelihood function modeling the generation of LOF variants in human populations (Fig. 2). Denoting $\mathbf{x}_i$ as the column vector of genomic features associated with gene i, the feedforward neural network describes the relationship between $\mathbf{x}_i$ and the prior distribution of η_i. Denoting y_i and n_i as the observed and expected numbers of LOF variants in gene i, respectively, the likelihood function is defined as the probability of observing y_i given n_i and η_i.

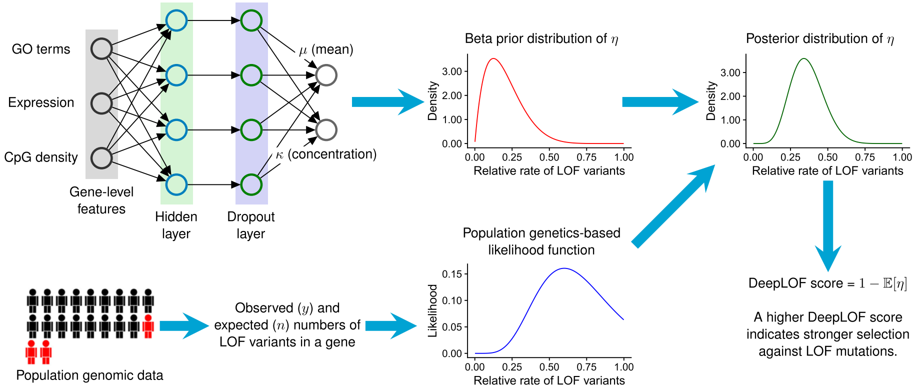
GO terms
Expression
CpG density
Gene-level features
Hidden layer
Dropout layer
μ (mean)
κ (concentration)
Beta prior distribution of η
Density
3.00
2.00
1.00
0.00
0.00 0.25 0.50 0.75 1.00
Relative rate of LOF variants
Posterior distribution of η
Density
3.00
2.00
1.00
0.00
0.00 0.25 0.50 0.75 1.00
Relative rate of LOF variants
Population genetics-based likelihood function
Likelihood
0.15
0.10
0.05
0.00
0.00 0.25 0.50 0.75 1.00
Relative rate of LOF variants
Population genomic data
Observed (y) and expected (n) numbers of LOF variants in a gene
DeepLOF score = $1 - \mathbb{E}[\eta]$
A higher DeepLOF score indicates stronger selection against LOF mutations.

Fig. 2 Overview of the DeepLOF model. DeepLOF combines a feedforward neural network and a population genetics-based likelihood function to infer the relative rate of LOF variants in a gene (η) with respect to the expected number under a neutral mutation model (n). The feedforward neural network transforms genomic features into a beta prior distribution of η, which represents our belief about η based on genomic features. The population genetics-based likelihood function describes the probability of observing y LOF variants in a gene conditional on η and n, which represents our belief about η based on population genomic data. Finally, DeepLOF combines the prior distribution and the likelihood function to compute the posterior distribution of η. The DeepLOF score is defined as $1 - \mathbb{E}[\eta]$, where $\mathbb{E}[\eta]$ is the mean of η under the posterior distribution

Specifically, we treat η_i as a random variable ranging from 0 to 1 and utilize a beta distribution to describe its prior distribution,

$$f(\eta_i|\mathbf{x}_i) = \frac{\eta_i^{\mu_i\kappa_i-1}(1-\eta_i)^{(1-\mu_i)\kappa_i-1}}{\mathrm{B}(\mu_\mathrm{i}\kappa_\mathrm{i}, (1-\mu_\mathrm{i})\kappa_\mathrm{i})} \tag{1}$$

where $f(\eta_i|\mathbf{x}_i)$ is the probability density function of η_i given feature vector $\mathbf{x}_i$; B is the beta function; μ_i and κ_i are the mean and concentration parameters of the beta distribution in gene i. It is worth noting that we employ an alternative parameterization of the beta distribution here [25]. The two shape parameters in the canonical parametrization of the beta distribution are equal to $\mu_i\kappa_i$ and $(1-\mu_i)\kappa_i$, respectively. Under the alternative parameterization, the mean of η_i is equal to μ_i, and the variance of η_i decreases with increasing κ_i. The alternative parametrization has been used in other Bayesian models due to the better interpretability of the mean and concentration parameters [25].

In the feedforward neural network, we seek to model the relationship between $\mathbf{x}_i$ and the parameters of the beta prior distribution (μ_i and κ_i). There are two versions of feedforward neural network in DeepLOF: a nonlinear version with hidden layer and a linear version without hidden layer. Specifically, in the nonlinear version of DeepLOF, the hidden layers can be represented by the following equation,

$$\mathbf{z}_i = \mathrm{Dropout}(\mathrm{ReLU}(\mathbf{W}_\mathrm{h}^\top \mathbf{x}_\mathrm{i} + \mathbf{b}_\mathrm{h})), \tag{2}$$

where $\mathbf{z}_i$ is the vector of hidden units; ReLU and Dropout are the the rectified linear layer [26] and the dropout layer [27]; $\mathbf{W}_h$ and $\mathbf{b}_h$ are the weight matrix and the bias vector of the rectified linear layer. After the hidden layers, we add an additional layer to transform $\mathbf{z}_i$ into μ_i and κ_i,

$$\begin{aligned}\mu_i &= \text{logistic}(\mathbf{w}_\text{m}^\top \mathbf{z}_\text{i} + \text{b}_\text{m}) \\ \kappa_i &= \exp(\mathbf{w}_k^\top \mathbf{z}_i + b_k),\end{aligned} \tag{3}$$

where $\mathbf{w}_m$ and b_m are the weight vector and the bias term associated with μ_i; $\mathbf{w}_k$ and b_k are the weight vector and the bias term associated with κ_i; the logistic function ensures that μ_i ranges from 0 to 1; the exponential function ensures that κ_i is positive.

In the alternative linear version of DeepLOF, the feedforward neural network does not include any hidden layer. Instead, we directly transform feature vector $\mathbf{x}_i$ into μ_i and κ_i,

$$\begin{aligned}\mu_i &= \text{logistic}(\mathbf{w}_\text{m}^\top \mathbf{x}_\text{i} + \text{b}_\text{m}) \\ \kappa_i &= \exp(\mathbf{w}_k^\top \mathbf{x}_i + b_k),\end{aligned} \tag{4}$$

which is similar to Eq. 3 expect that $\mathbf{z}_i$ is replaced by $\mathbf{x}_i$. The linear DeepLOF model allows us to directly infer the associations of genomic features with LOF intolerance based on the negative values of weights in $\mathbf{w}_m$.

In the likelihood function, we seek to model the generation of LOF variants in human populations. Specifically, we assume that the observed number of LOF variants in gene i follows a Poisson distribution,

$$P(y_i|\eta_i, n_i) = \frac{(\eta_i n_i)^{y_i} \exp(-\eta_i n_i)}{(\eta_i n_i)!}, \tag{5}$$

where y_i and n_i are the observed and expected numbers of LOF variants, respectively, and the mean of the Poisson distribution is equal to $\eta_i n_i$.

In the training step, the DeepLOF model combines the prior distribution (Eq. 1) and the likelihood function (Eq. 5) to obtain the marginal likelihood of the model,

$$P(y_i|\mathbf{x}_i, n_i) = \int_0^1 f(\eta_i|\mathbf{x}_i) P(y_i|\eta_i, n_i) d\eta_i, \tag{6}$$

which represents the probability of observing y_i LOF variants in gene i conditional on $\mathbf{x}_i$ and n_i. It is worth noting that we omit the parameters of the feedforward neural network in this equation for the sake of notation simplicity. Because there is no analytical solution for the integral in this equation, we use the midpoint Riemann sum to approximately compute $P(y_i|\mathbf{x}_i, n_i)$. To estimate the parameters of the feedforward neural network, we perform stochastic gradient descent on the following loss function,

$$-\frac{1}{|\Psi|} \sum_{i \in \Psi} \log(P(y_i|\mathbf{x}_i, n_i)), \tag{7}$$

where Ψ and $|\Psi|$ are the gene set and the number of genes in a mini-batch of data. We use the Adam algorithm [28] for the mini-batch gradient descent and utilize early stopping and L2 regularization to avoid overfitting.

In the prediction step, we fix the parameters of the feedforward neural network to the optimal values from the training step. Then, we obtain the density function of the posterior distribution of η_i using Bayes' rule,

$$f(\eta_i|y_i, \mathbf{x}_i, n_i) = \frac{f(\eta_i|\mathbf{x}_i)P(y_i|\eta_i, n_i)}{P(y_i|\mathbf{x}_i, n_i)}, \quad (8)$$

which represents our belief about η_i after integrating genomic features and population genomic data. The mean of η_i under the posterior distribution is equal to

$$\mathbb{E}[\eta_i] = \int_0^1 \eta_i f(\eta_i|y_i, \mathbf{x}_i, n_i) d\eta_i, \quad (9)$$

which we compute numerically using the midpoint Riemann sum. Finally, we define the DeepLOF score as $1 - \mathbb{E}[\eta_i]$. A higher DeepLOF score indicates that LOF mutations in the corresponding gene are subject to stronger negative selection.

Genomic features

The training data of DeepLOF included 18 genomic features. First, we obtained five sets of epigenomic data from various cell types [20]. These data included ChIP-seq peaks of H3K9ac, H3K27me3, H3K4me3, and H2A.Z in promoter regions and promoter-enhancer interactions predicted by EpiTensor [29]. We defined H3K9ac, H3K27me3, H3K4me3, and H2A.Z signals as the average length of the corresponding ChIP-seq peak in a gene's promoter across all cell types. We defined the enhancer number in a gene as the average number of promoter-enhancer interactions across all cell types. Second, we downloaded four development-related gene categories from MSigDB (version 7.1) [30]. These gene categories included 1029, 995, 508, and 1131 genes from two GO categories [31], i.e., embryo development and central nervous system development, and two Reactome pathways [32], i.e., nervous system development and developmental biology. We converted each development-related gene category into a binary feature indicating whether each gene was included in the category. Third, we obtained a list of 1254 transcription factor genes [33] and a list of 3431 genes encoding subunits of protein complexes [34]. We converted each gene list into a binary feature indicating whether each gene was included in the list. Fourth, we obtained promoter CpG density, promoter phastCons score, and exonic phastCons score from a previous study [15]. Fifth, we downloaded mean gene expression level, tissue specificity (tau) [35], PPI degree [36] from a recent study [37]. Finally, we obtained the UNEECON-G score from its original publication [38].

We observed that several genomic features were nonnegative and had right-skewed distributions. Following a common practice in machine learning and statistics, we applied a log transformation to these features (Additional file 1: Table S1), $x'_{ij} = \log(x_{ij} + \delta_j)$, where x_{ij} is the raw value of feature j in gene i, x'_{ij} is the transformed feature, and δ_j is the minimum observed positive value of feature j. Then, we standardized each continuous feature by subtracting its mean and dividing by its standard deviation. We imputed missing values of each feature with the mean of the non-missing values.

Model training

We downloaded the observed number of LOF variants in each protein-coding gene and the expected number under a neutral mutation model from gnomAD (version 2.1.1) [5].

We combined the expected and observed numbers of LOF variants with the 18 genomic features to build a dataset of 19,197 genes for model training. We randomly split these genes into a training set (80% genes) and a validation set (20% genes). We trained the DeepLOF model on the training set and used a grid search to tune hyperparameters in the validation set. In the training of the linear DeepLOF model, these hyperparameters included the L2 penalty (0, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}) and the learning rate of the Adam algorithm (10^{-3}, 10^{-4}, 10^{-5}). In the training of the nonlinear DeepLOF model, we added an additional hyperparameter, i.e., the number of hidden units in the feedforward neural network (64, 128, 256, 512, 1024). We fixed the dropout rate to 0.5. We computed the contribution scores of genomic features using the optimal linear DeepLOF model with the lowest loss in the validation set. We computed the DeepLOF score using the nonlinear model with the lowest loss in the validation set. The optimal nonlinear model had a lower loss than the optimal linear model.

Comparison with other methods in predicting disease genes

For this evaluation, we evaluated the performance of DeepLOF and eight alternative methods, including LOEUF [5], pLI [2], mis-z [1], RVIS [39], GeVIR [6], CoNeS [7], VIRLOF [6], and UNEECON-G [38], in predicting essential genes and dominant-negative genes. We obtained LOEUF, pLI, and mis-z scores from the gnomAD database (version 2.1.1) [5]. We downloaded the RVIS score trained on the ExAC dataset [2] from dbNSFP (version 4.0) [40]. We obtained the other gene scores from the corresponding publications.

We downloaded 311 ClinGen haploinsufficient genes and 404 mouse genes where heterozygous knockouts resulted in lethality from the GitHub repository for gnomAD (https://github.com/macarthur-lab/gnomad_lof/). Then, we obtained 18,797 human-mouse orthologs from the Mouse Genome Database [23, 41] and used these data to map the mouse essential genes to the human genome, resulting in 397 human orthologs of mouse essential genes. We downloaded 683 human genes deemed essential in cell lines and 913 genes without significant fitness effects in cell lines from the GitHub repository for the MacArthur Lab (https://github.com/macarthur-lab/gene_lists). Finally, we obtained 364 OMIM dominant-negative genes from a previous study [39].

By using known LOF-intolerant genes from these sources, we sought to evaluate the performance of our model in comparison to previous methods. We used these data to provide matched gene sets of known essential genes with nonessential genes, allowing us to determine the true and false positive rates of the different predictive models. Essential genes in this data set are considered to be genes determined to be LOF-intolerant and conversely, nonessential genes are those which are considered LOF-tolerant.

To create receiver operating characteristic curves (ROCs) that would show this comparison, we first needed to match each essential gene with a nonessential gene containing a similar expected number of LOF variants. To this end, we first constructed a nonessential gene set for each of the essential gene sets that were of matching size. We matched each essential gene with a nonessential gene of similar expected number of LOF variants using MatchIt [42]. For the 311 ClinGen haploinsufficient genes, the 397 human orthologs of mouse essential genes, and the 364 dominant-negative genes, we considered all other human genes to be nonessential. For the 683 human genes deemed essential in cell lines, we considered the 913 human genes without significant fitness effects in cell lines to be nonessential. Finally, we used ROCR to plot the receiver operating characteristic curves and calculate the AUCs for all computational methods in the matched gene sets [43]. We evaluated the statistical significance of the difference in AUC using the DeLong test [44].

Evaluation of the 109 LOF-intolerant genes uniquely predicted by DeepLOF

We obtained comparable numbers of LOF-intolerant genes from DeepLOF, LOEUF, VIRLOP, and CoNeS. First, we obtained 2835 LOF-intolerant genes from LOEUF using an established cutoff of 0.35 [5]. To obtain similar numbers of LOF-intolerant genes from the other methods, we used cutoffs of 0.835, and −1.11, and 15 for DeepLOF, CoNeS, and VIRLOF percentile scores, respectively. Given these cutoffs, DeepLOF, CoNeS, and VIRLOF predicted 2817, 2847, and 2817 LOF-intolerant genes, respectively. To evaluate the power of these methods in predicting short essential genes, we retained LOF-intolerant genes with $\leq$ 10 expected LOF variants for downstream analysis.

We evaluated the enrichment of ClinGen haploinsufficient genes, human orthologs of mouse essential genes, and human genes essential for the survival of cell lines in the 109 LOF-intolerant genes uniquely predicted by DeepLOF. For each essential gene set, we defined the other genes as nonessential genes. Also, we defined LOF-tolerant genes as those genes with $\leq$ 10 expected LOF variants and not predicted to be LOF-intolerant by any method. We evaluated the enrichment of each essential gene set in the 109 LOF-intolerant genes using the log odds ratio, $\log(\mathrm{OR}) = \log(\frac{\mathrm{n_{11}/n_{12}}}{\mathrm{n_{21}/n_{22}}})$, where n_{11}, n_{12}, n_{21}, and n_{22} are the numbers of essential genes predicted to be LOF-intolerant, nonessential genes predicted to be LOF-intolerant, essential genes predicted to be LOF-tolerant, and nonessential genes predicted to be LOF-tolerant, respectively. We defined the confidence interval of the log odds ratio as $\log(\mathrm{OR}) \pm 1.96 \times \mathrm{SE}$, where SE is the standard error of the log odds ratio and is equal to $\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$.

We evaluated the depletion of the 109 LOF-intolerant genes uniquely predicted by DeepLOF in benign genomic deletions. To this end, we obtained clinical structural variants from the nstd102 study in dbVar [45] and retrained 5649 benign deletions overlapping coding regions of genes from GENCODE (version 19) [46]. Then, we computed the proportion of benign deletions overlapping at least one of the 109 LOF-intolerant genes. To examine whether the proportion of overlapping deletions was smaller than the expectation under a null model that the 109 LOF-intolerant genes are nonessential. We performed a permutation test with 10,000 permutations. In each permutation, we randomly selected 109 genes with $\leq$ 10 LOF variants and computed the proportion of deletions overlapping with the random genes. The one-tailed *P*-value of the permutation test was defined as the fraction of permutations where the proportion of deletions overlapping random genes was equal to or smaller than the observed proportion in empirical data.

Results

Overview of the DeepLOF model

DeepLOF is an evolution-based deep learning model for inferring protein-coding genes intolerant to LOF mutations. The key variable of interest in DeepLOF is η, i.e., the relative rate of LOF variants in a gene with respect to the expected number of LOF variants under a neutral mutation model. A smaller η indicates that a gene has a lower rate of LOF variants after adjusting for neutral evolutionary factors, such as mutation rate and genetic drift. Thus, a smaller η indicates stronger negative selection against LOF variants. To take into account the uncertainty of η, DeepLOF treats η as a random variable at the gene level. To integrate genomic features and population genomic data in a Bayesian manner, DeepLOF combines a feedforward neural network and a population genetics-based likelihood function (Fig. 2).

In this hybrid framework, the feedforward neural network consists of a sequence of neural network layers, which together transform genomic features into the beta prior distribution of η (Fig. 2). The genomic features include gene ontology (GO) terms [31], epigenomic data, gene expression patterns, and several other gene-level features potentially predictive of LOF intolerance. The outputs of the feedforward neural network are the mean and concentration parameters of the beta distribution, which represents our belief about η based on genomic features. In addition, the population genetics-based likelihood function describes the probability of observing y LOF variants in a gene given η and n, where n is the expected number of LOF variants in the same gene under a neutral mutation model (Fig. 2). Thus, the likelihood function represents evidence for LOF intolerance based on population genomic data.

Using Bayes' rule, DeepLOF combines the neural network-based beta prior distribution with the population genetics-based likelihood function to obtain a posterior distribution of η, which represents our belief about LOF intolerance after integrating genomic features and population genomic data. Denoting $\mathbb{E}[\eta]$ as the expectation of η under the posterior distribution, we define the DeepLOF score as $1 - \mathbb{E}[\eta]$, which can be interpreted as the proportion of LOF mutations purged by negative selection in a gene. Thus, a higher DeepLOF score indicates a higher level of LOF intolerance. We estimate model parameters, including the weights and biases of the feedforward neural network, using stochastic gradient descent on a loss function that integrates the feedforward neural network and the likelihood function.

DeepLOF elucidates genomic features predictive of LOF-intolerant genes

We trained the DeepLOF model on 18 genomic features (Additional file 2: Data 1) and the observed and expected numbers of LOF variants in 19,197 human genes. The observed number of LOF variants in each gene was from the exomes of 125,748 healthy individuals in the gnomAD database [5]. The expected number of LOF variants in each gene was from a neutral mutation model developed by gnomAD [5], which took into account the impact of trinucleotide sequence context, CpG methylation level, local mutation rate, and site-wise sequencing coverage on the occurrence of variants. The 18 genomic features included five epigenomic features [20], four gene categories associated with developmental processes [30], three protein annotations [33, 34, 36], two phastCons conservation scores [15, 47], two gene expression features [48, 49], the promoter CpG density [15], and the UNEECON-G score [38]. A detailed description of these genomic features is available in Additional file 1: Table S1. We used 80% randomly selected genes as a training set and used the remaining 20% genes as a validation set for hyperparameter tuning.

To obtain insights into which genomic features may be predictive of gene-level intolerance to LOF mutations, we trained a linear DeepLOF model without hidden layer in the feedforward neural network. While the linear DeepLOF model may not provide most accurate predictions of LOF intolerance, it allows us to estimate the association of each genomic feature with LOF intolerance. Specifically, in the linear DeepLOF model, we defined the contribution score of a genomic feature as the negative value of its weight with respective to the mean of the beta prior distribution of η. The absolute value of a contribution score indicates the strength of association between a feature and LOF intolerance, whereas the sign of the contribution score indicates the direction of association.

Among the 18 genomic features, the UNEECON-G score had the strongest positive association with LOF intolerance (Fig. 3a). Because the UNEECON-G score is a measure of a gene's intolerance to missense mutations, it corroborates a previous finding that missense intolerance is strongly correlated with LOF intolerance at the gene level [38]. Two GO categories [31], i.e., central nervous system development and embryo development, and the Reactome category of nervous system development [32] had strong positive associations with LOF intolerance, suggesting that developmental genes are highly intolerant to LOF mutations. Two protein annotations, i.e., transcription factor [33] and protein complex [34], also had strong positive associations with LOF intolerance, suggesting that genes encoding transcription factors or subunits of protein complexes may be more intolerant to LOF mutations than other protein-coding genes. In agreement with previous studies [15, 20], epigenomic features in a gene's promoter, including the signals of H3K9ac, H3K27me3, and H3K4me3 histone modifications [20], and the promoter CpG density [15], had positive associations with LOF intolerance. Furthermore, the phastCons score [47] in a gene's promoter had a positive association with LOF intolerance, suggesting that genes with conserved promoter sequences may be intolerant to LOF mutations.

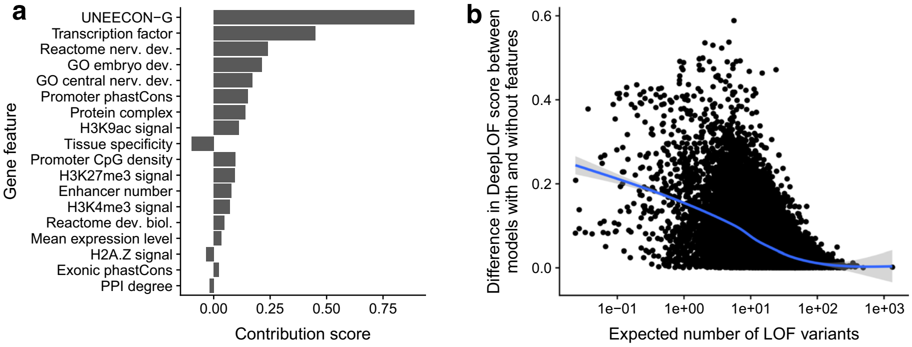

Fig. 3 Impact of genomic features on the inference of LOF intolerance. **a** Association of genomic features with LOF intolerance. We define the contribution score of a genomic feature as the negative weight of the feature in the linear DeepLOF model. The absolute value of a contribution score indicates the strength of association between a feature and LOF intolerance, whereas the sign of a contribution score indicates the direction of association. **b** DeepLOF automatically adjusts the relative importance of genomic features in a gene length-dependent manner. The *x* axis represents the expected number of LOF variants. The *y* axis represents the absolute difference in DeepLOF score between the linear DeepLOF model with genomic features and the counterpart model without genomic features. A higher absolute difference in DeepLOF score indicates a stronger impact of genomic features on the inference of LOF intolerance. Each dot represents a gene. The blue and grey curves represent the fit of the generalized additive model with integrated smoothness and its 95% confidence interval [66]

H2A.Z signal had a negative association with LOF intolerance. The H2A.Z histone variant plays a vital role in gene regulation during mammalian development, specifically at promoter regions. In embryonic stem cells, bivalent domains, characterized by both activating and repressive histone modifications, are common and frequently feature H2A.Z. As lineage specification progresses, bivalent domains of crucial developmental genes often resolve. Genes unexpressed within the relevant lineage maintain repressive H3K27me3 domains and lose H2A.Z [50]. Consequently, our model may detect these genes and their post-developmental H2A.Z depletion, leading to a negative correlation between LOF intolerance and H2A.Z signal. Finally, tissue specificity [35] also had a negative association with LOF intolerance, suggesting that housekeeping genes may be more intolerant to LOF mutations than tissue-specific genes.

DeepLOF automatically adjusts the relative importance of genomic features and population genomic data in a gene length-dependent manner

Because DeepLOF uses Bayes' rule to infer the distribution of η, we hypothesized that, similar to other Bayesian models, DeepLOF may automatically adjust the relative importance of the beta prior distribution in a data-dependent manner. Specifically, in a long gene where a large number of LOF variants is expected under a neutral mutation model, the posterior distribution of η may be dominated by the the population genetics-based likelihood function. Thus, DeepLOF may primarily leverage population genomic data to predict long LOF-intolerant genes. Because population genomic data are indicative of negative selection at the organism level, this may allow DeepLOF to unbiasedly infer LOF intolerance at the organism level in long genes. Conversely, in a short gene where a small number of LOF variants is expected under a neutral mutation model, the posterior distribution of η may be dominated by the beta prior distribution of η. Thus, DeepLOF may automatically upweight genomic features to improve the inference of LOF intolerance in short genes.

To test this hypothesis, we retrained the linear DeepLOF model without using any genomic features. This model effectively assumed an identical prior distribution of η across genes and solely used population genomic data to infer LOF intolerance. We computed the absolute difference in DeepLOF score between the linear DeepLOF model with genomic features and the model without genomic features, which indicates the relative importance of genomic features in the inference of LOF intolerance. We observed that the absolute difference in DeepLOF score was negatively correlated with the expected number of LOF variants in a gene (Fig. 3b), supporting our hypothesis that DeepLOF automatically upweights genomic features in short genes to improve the prediction of LOF intolerance.

DeepLOF shows unmatched performance in predicting essential genes intolerant to LOF mutations

We hypothesized that, by integrating a large number of genomic features and population genomic data, DeepLOF may show improved performance in predicting essential genes. To test this hypothesis, we obtained three sets of essential genes, including 311 ClinGen haploinsufficient genes [22], 397 human orthologs of mouse essential genes where heterozygous knockouts resulted in lethality [23], and 683 human genes essential to the survival of cell lines [24]. For each essential gene set, we constructed a nonessential gene set of matching size. To this end, we used MatchIt [42] to match each essential gene with a putatively nonessential gene of similar number of LOF variants.

We trained a nonlinear DeepLOF model with hidden layer and observed that the nonlinear DeepLOF model had a lower loss than the linear DeepLOF model in the validation set. Thus, we used scores from the nonlinear DeepLOF model in downstream analysis (Additional file 3: Data 2). We compared the performance of DeepLOF with eight alternative methods in distinguishing essential genes from matched nonessential genes. The eight alternative methods included two measures of gene-level intolerance to LOF mutations (LOEUF [5] and pLI [2]), three measures of gene-level intolerance to missense mutations (mis-z [1], GeVIR [6], and UNEECON-G [38]), and three metrics that considered both LOF intolerance and missense intolerance (RVIS [39], VIRLOF [6], and CoNeS [7]). In the prediction of ClinGen haploinsufficient genes, DeepLOF showed substantially better performance than the other methods as evidenced by its significantly higher area under the receiver operating characteristic curve (AUC) (Fig. 4a; Additional file 1: Table S2). In particular, DeepLOF showed unmatched performance when the false positive rate was low. For instance, the true positive rate of DeepLOF was approximately

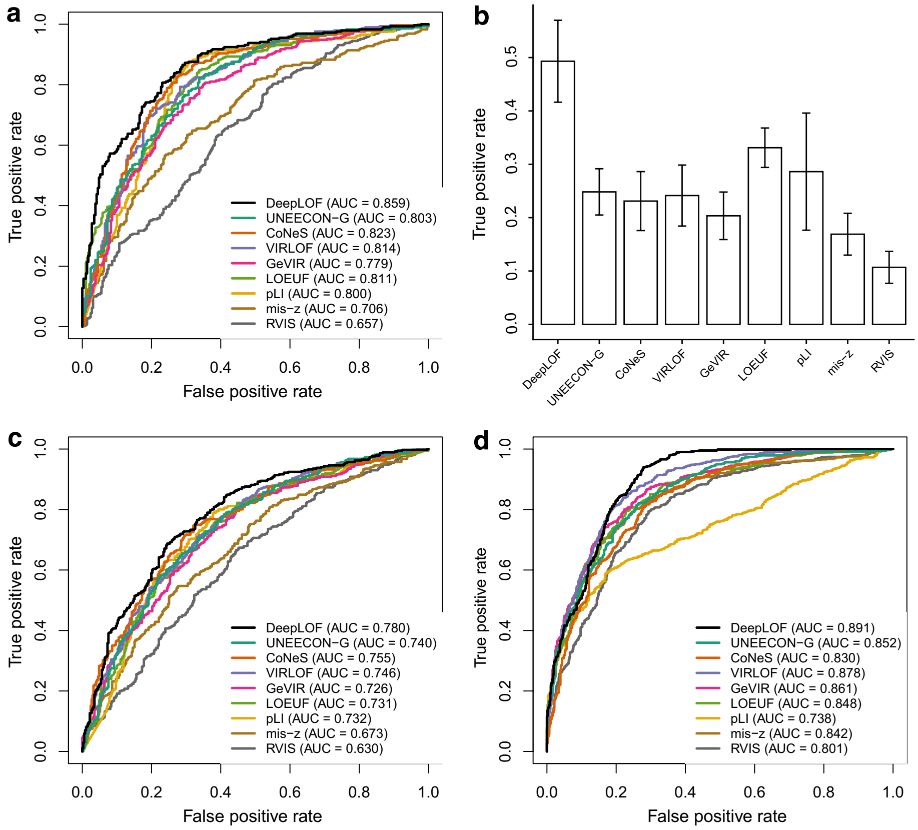

Fig. 4 Predictive power of DeepLOF and alternative methods in distinguishing essential genes from putatively nonessential genes. **a** Performance in predicting ClinGen haploinsufficient genes. **b** True positive rates in predicting ClinGen haploinsufficient genes at a fixed false negative rate of 5%. Error bars represent bootstrap standard errors of true positive rates. **c** Performance in predicting human orthologs of mouse essential genes. **d** Performance in predicting human genes essential for the survival of cell lines

0.5 at a false positive rate of 0.05 (Fig. 4b), which was about 50% higher than that of the second best method (LOEUF; true positive rate = 0.33). DeepLOF also outperformed the other methods in predicting human orthologs of mouse essential genes and human genes essential for the survival of cell lines (Fig. 4c, d; Additional file 1: Table S2). In sum, DeepLOF had superior performance in predicting essential genes.

Because the DeepLOF score is a measure of LOF intolerance, we hypothesized that it might not be the best method in predicting disease genes via a mechanism different from haploinsufficiency. To test this hypothesis, we obtained 364 OMIM dominant-negative genes where a heterozygous mutation may adversely affect the function of the wild-type allele in the same individual through interlocus or intralocus interactions [39, 51, 52]. In agreement with our hypothesis, UNEECON-G instead of DeepLOF showed the best performance in predicting dominant-negative genes (Additional file 1: Fig. S1), suggesting that missense intolerance scores, such as UNEECON-G, might be better predictors of dominant-negative genes than LOF intolerance scores.

DeepLOF predicts 109 novel LOF-intolerant genes of short length

Previous predictions of LOF intolerance are often biased towards longer genes because it is easier to reject neutral evolution when the expected number of LOF variants is high [6, 15]. In contrast, by leveraging a genomic feature-based prior distribution, DeepLOF may have higher power to predict LOF-intolerant genes of short length. To test this hypothesis, we examined LOF-intolerant genes predicted by four methods, including DeepLOF, CoNeS, LOEUF, and VIRLOP, which showed better performance than other methods in predicting ClinGen haploinsufficient genes (Fig. 4a).

To ensure that the number of predicted LOF-intolerant genes is comparable between the four methods, we obtained ~ 2800 LOF-intolerant genes from each method. Specifically, using a previously established cutoff of 0.35 [5], we obtained 2835 LOF-intolerant genes from LOEUF. Using comparable cutoffs, we obtained 2817, 2847, and 2817 LOF-intolerant genes from DeepLOF, CoNeS, and VIRLOP, respectively. Because previous studies suggested that the difficulty of LOF intolerance prediction mainly occurred in genes with ≤ 10 expected LOF variants [5, 15], we focused on predicted LOF-intolerant genes with ≤ 10 expected LOF variants in downstream analysis.

In total, 452 genes with ≤ 10 expected LOF variants were predicted to be LOF-intolerant by at least one method (Additional file 4: Data 3). DeepLOF predicted that 364 genes with ≤ 10 expected LOF variants were LOF-intolerant, which was the largest number among all the four methods (Fig. 5a). Also, 109 LOF-intolerant genes, or 24.1% of the total number (109/452), were uniquely predicted by DeepLOF (Fig. 5a) (Additional file 5: Data 4). Because we used comparable cutoffs for the four methods, these results suggest that DeepLOF may have much higher power to pinpoint LOF-intolerant genes of short length.

We examined the enrichment of ClinGen haploinsufficient genes, human orthologs of mouse essential genes, and human genes essential for the survival of cell lines in the 109 novel LOF-intolerant genes predicted by DeepLOF, using genes with ≤ 10 expected LOF variants but not predicted to be LOF-intolerant by any method as a background set. We observed that the 109 novel LOF-intolerant genes were significantly enriched with

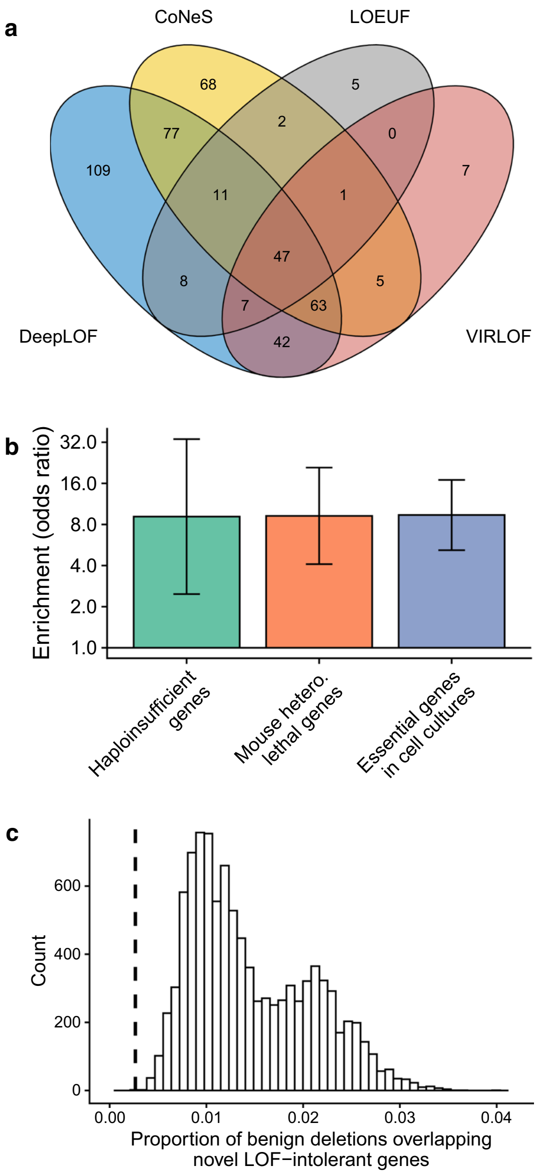

Fig. 5 Predicted LOF-intolerant genes with ≤ 10 expected LOF variants. **a** Venn diagram of predicted LOF-intolerant genes. **b** Enrichment of essential genes in the 109 LOF-intolerant genes uniquely predicted by DeepLOF. Error bars represent 95% confidence intervals. **c** Proportion of benign genomic deletions overlapping the 109 LOF-intolerant genes uniquely predicted by DeepLOF. The histogram represents the null distribution of the proportion from a permutation test, and the dashed vertical line represents the observed proportion in empirical data

essential genes (Fig. 5b), highlighting that the novel LOF-intolerant genes predicted by DeepLOF may play key roles in important biological processes.

We hypothesized that the 109 novel LOF-intolerant genes predicted by DeepLOF might be depleted in benign genomic deletions due to the detrimental effects of deletions overlapping LOF-intolerant genes. To test this hypothesis, we obtained 5,649 benign genomic deletions overlapping protein-coding genes from dbVar [45]. We observed that 0.27% of benign deletions overlapped the 109 novel LOF-intolerant genes (Fig. 5c). To evaluate whether the proportion of overlapping was smaller than the expectation under a null model that postulated the 109 genes to be nonessential, we performed a permutation test with 10,000 permutations. In each permutation, we randomly selected 109 genes with ≤ 10 LOF variants and computed the proportion of benign genomic deletions overlapping the random genes. The mean proportion of benign deletions overlapping random genes was 1.46% (Fig. 5c), which was 5.4 fold higher than the observed proportion in empirical data (1.46% vs. 0.27%; $P = 0$; one-tailed permutation test). Thus, benign genomic deletions were depleted with the 109 novel LOF-intolerant genes predicted by DeepLOF.

Finally, we further investigated the 109 short genes uniquely predicted by DeepLOF by conducting a GO enrichment analyses on the 109 genes. We used The Database for Annotation, Visualization and Integrated Discovery (DAVID) [53, 54] and the PANTHER Classification System (Protein Analysis Through Evolutionary Relationships) [55] for our analysis. We found many overrepresented essential biological functions in our dataset, including many found in the development of organs and morphogenesis.

An example of a critical developmental gene determined by DeepLOF to be LOF-intolerant was HAND2. HAND2 is involved during cardiac development. HAND proteins are involved in the development of ventricular chambers and aortic arch arteries [56]. As such, dysfunction of these important proteins has been associated with congenital heart defects. Specifically, a heterozygous deletion in HAND1 and HAND2 has been shown to be associated with heart defects. Therefore, HAND2 (predicted by DeepLOF) is thought to be haploinsufficient [57] and the cause of certain congenital heart issues [56, 57].

In the 109 DeepLOF uniquely predicted genes, we also found significant enrichment for several genes associated with various ribosomal proteins. Ribosomal proteins are responsible for many important functions and dysfunction can lead to serious complications. Diamond-Blackfan anemia (DBA) is a rare disease in which patients' bone marrow fails to produce enough red blood cells. The disease also is responsible for malformations in the hands, face, or heart in approximately 50% of DBA patients. Haploinsufficient mutations in various ribosomal proteins have been implicated in causing this serious disease [58]. TSR2 was a gene uniquely deemed by DeepLOF to be LOF-intolerant. Dysfunction of this gene has been associated with DBA [56, 58, 59]. DeepLOF also uniquely predicted several other ribosomal proteins where mutation has been associated with DBA. This list of ribosomal proteins novelly predicted by our model and validated in association with DBA [58] included proteins RPL27, RPL35, RPS27, RPS28.

Full names, symbols and Ensembl ID of all other genes uniquely predicted by DeepLOF can be found in Additional file 5: Data 4.

Discussion

We present an evolution-based machine learning framework, DeepLOF, for predicting human genes intolerant to LOF mutations. Unlike previous LOF intolerance scores, such as pLI and LOEUF, the DeepLOF model leverages both population and functional genomic data to predict LOF intolerance. Therefore, DeepLOF may be particularly powerful in predicting short essential genes without sufficient polymorphisms for selection inference. Furthermore, unlike supervised methods, DeepLOF does not use known essential genes as training data. Thus, it may not suffer from the pitfalls of supervised machine learning, such as the potential leakage of information from training data to test data and the ascertainment bias in human-annotated essential genes.

The linear DeepLOF model without hidden layer allows us to directly estimate the association of a genomic feature with LOF intolerance after adjusting for other genomic features (Fig. 3a). Using this approach, we show that the UNEECON-G score has the strongest positive association with LOF intolerance, which suggests that missense intolerance scores may also be informative of gene-level intolerance to LOF mutations. Because there are typically more missense variants than LOF variants in a gene under a neutral mutation model, the sample size for missense intolerance inference is larger than that for LOF intolerance inference. Therefore, it may be easier to reliably estimate missense intolerance than LOF intolerance, and in turn it may be beneficial to incorporate missense intolerance scores, such as UNEECON-G, into computational pipelines for LOF variant interpretation.

We also show that genes encoding transcription factors or protein complex subunits and genes associated with developmental processes may be highly intolerant to LOF mutations (Fig. 3a). Previous studies have shown that many transcription factors are haploinsufficient and are associated with dominant genetic disorders [60]. Transcription factors often cooperatively bind to regulatory sequences, which may result in a sigmoid-shaped dose-response curve [61, 62]. Therefore, transcription factors may be particularly susceptible to heterozygous knockouts. Also, in agreement with our observation, it has been shown that many protein complex subunits are haploinsufficient [63] because the reduced expression of a subunit may lead to a stoichiometric imbalance between different subunits of the same protein complex [62]. Finally, in agreement with our observation, it has been found that many developmental genes are haploinsufficient [64], highlighting that developmental processes may be particularly sensitive to reduced gene dosage.

While both genomic features and population genomic data are predictive of LOF intolerance, their relative importance may depend on the length of a gene. In short genes where population genomic data provide limited information on negative selection, it is critical to incorporate genomic features to improve the inference of LOF intolerance. In contrast, in long genes where polymorphisms are abundant, population genomic data may be more informative than genomic features because they directly reflect LOF intolerance at the organism level. Because DeepLOF infers the relative rate of LOF variants, η, in a Bayesian manner, it can automatically adjust the relative importance of genomic features and population genomic data to optimize LOF intolerance inference in a data-dependent manner (Fig. 3b).

By integrating genomic features and population genomic data, DeepLOF outperforms alternative methods in predicting essential genes (Fig. 4). Because most variants in the gnomAD database are of low allele frequency [5], the DeepLOF score may be indicative of negative selection against LOF variants in their heterozygous state. Thus, it shows unmatched performance in predicting ClinGen haploinsufficient genes (Fig. 4a). In particular, at a false positive rate of 0.05, the true positive rate of DeepLOF is more than 50% higher than other methods (Fig. 4b). Because the true positive rate is the proportion of haploinsufficient genes correctly discovered by a method, our result suggests that DeepLOF can detect 50% more ClinGen haploinsufficient genes than other methods at a false positive rate of 5%.

We observe that the predictive power of DeepLOF and other methods in disease gene prediction depends on the genetic mechanism of a disease. Notably, DeepLOF is outperformed by a missense intolerance score, UEECON-G, in the prioritization of dominant-negative disease genes (Additional file 1: Fig. S1), possibly because many dominant-negative mutations are missense mutations. Therefore, it is critical to take into account the genetic mechanism of a disease in gene prioritization [65].

Because DeepLOF leverages genomic features to improve the inference of LOF intolerance in short genes, DeepLOF has predicted the largest number of short LOF-intolerant genes compared to other methods (Fig. 5a). Furthermore, DeepLOF has predicted 109 novel LOF-intolerant genes of short length. These novel LOF-intolerant genes are enriched with essential genes and are depleted in benign genomic deletions (Fig. 5b, c), implicating that they may play an underappreciated role in human disease.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-023-05481-z.

Additional file 1. **Supplemental file containing: Supplementary Fig. 1** depicting the performance of DeepLOF and alternative methods in predicting dominant negative genes, **Supplementary Table 1** which provides a detailed description of the genomic features used for model training, and **Supplementary Table 2** which shows the statistical significance of the differences in AUC between DeepLOF and alternative methods in predicting essential genes.

Additional file 2. **Data File 1:** File containing DeepLOF training data.

Additional file 3. **Data File 2:** Scores from the nonlinear DeepLOF model used in downstream analysis.

Additional file 4. **Data File 3:** File containing the 452 genes with ≤ 10 expected LOF variants predicted to be LOF-intolerant by at least one method. The file shows which method(s) predicted the gene to be LOF-intolerant.

Additional file 5. Data File 4: File containing the 109 LOF-intolerant genes uniquely predicted by DeepLOF. The file shows gene IDs and brief description of each gene.

Acknowledgements

We thank Zhihan Liu for helpful discussions.

Author contributions

Y-FH developed the DeepLOF method. TML and Y-FH analyzed data. TML and Y-FH prepared and reviewed the manuscript.

Funding

Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R35GM142560, by the Pennsylvania State University and the National Institutes of Health T32 GM102057 Computation, Bioinformatics, and Statistics (CBIOS) Training Program Grant. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Availability of data and materials

The DeepLOF program is available at https://github.com/yifei-lab/DeepLOF. The training data, precomputed DeepLOF scores, and predicted LOF-intolerant genes with ≤ 10 LOF variants from this study are available as Supplementary Data files.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Received: 13 September 2022 Accepted: 13 September 2023
Published online: 18 September 2023

References

1. Samocha KE, Robinson EB, Sanders SJ, Stevens C, Sabo A, McGrath LM, Kosmicki JA, Rehnstrom K, Mallick S, Kirby A, Wall DP, MacArthur DG, Gabriel SB, DePristo M, Purcell SM, Palotie A, Boerwinkle E, Buxbaum JD, Cook EH Jr, Gibbs RA, Schellenberg GD, Sutcliffe JS, Devlin B, Roeder K, Neale BM, Daly MJ. A framework for the interpretation of de novo mutation in human disease. Nat Genet. 2014;46(9):944–50.
2. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, ODonnell-Luria AH, Ware JS, Hill AJ, Cummings BB, Tukiainen T, Birnbaum DP, Kosmicki JA, Duncan LE, Estrada K, Zhao F, Zou J, Pierce-Hoffman E, Berghout J, Cooper DN, Deflaux N, DePristo M, Do R, Flannick J, Fromer M, Gauthier L, Goldstein J, Gupta N, Howrigan D, Kiezun A, Kurki MI, Moonshine AL, Natarajan P, Orozco L, Peloso GM, Poplin R, Rivas MA, Ruano-Rubio V, Rose SA, Ruderfer DM, Shakir K, Stenson PD, Stevens C, Thomas BP, Tiao G, Tusie-Luna MT, Weisburd B, Won H-H, Yu D, Altshuler DM, Ardissino D, Boehnke M, Danesh J, Donnelly S, Elosua R, Florez JC, Gabriel SB, Getz G, Glatt SJ, Hultman CM, Kathiresan S, Laakso M, McCarroll S, McCarthy MI, McGovern D, McPherson R, Neale BM, Palotie A, Purcell SM, Saleheen D, Scharf JM, Sklar P, Sullivan PF, Tuomilehto J, Tsuang MT, Watkins HC, Wilson JG, Daly MJ, MacArthur DG. Exome aggregation consortium: analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536(7616):285–91.
3. Cassa CA, Weghorn D, Balick DJ, Jordan DM, Nusinow D, Samocha KE, O'Donnell-Luria A, MacArthur DG, Daly MJ, Beier DR, Sunyaev SR. Estimating the selective effects of heterozygous protein-truncating variants from human exome data. Nat Genet. 2017;49:806–10.
4. Fadista J, Oskolkov N, Hansson O, Groop L. Loftool: a gene intolerance score based on loss-of-function variants in 60,706 individuals. Bioinformatics. 2017;33(4):471–4.
5. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, Gauthier LD, Brand H, Solomonson M, Watts NA, Rhodes D, Singer-Berk M, England EM, Seaby EG, Kosmicki JA, Walters RK, Tashman K, Farjoun Y, Banks E, Poterba T, Wang A, Seed C, Whiffin N, Chong JX, Samocha KE, Pierce-Hoffman E, Zappala Z, O'Donnell-Luria AH, Minikel EV, Weisburd B, Lek M, Ware JS, Vittal C, Armean IM, Bergelson L, Cibulskis K, Connolly KM, Covarrubias M, Donnelly S, Ferriera S, Gabriel S, Gentry J, Gupta N, Jeandet T, Kaplan D, Llanwarne C, Munshi R, Novod S, Petrillo N, Roazen D, Ruano-Rubio V, Saltzman A, Schleicher M, Soto J, Tibbetts K, Tolonen C, Wade G, Talkowski ME, Aguilar Salinas CA, Ahmad T, Albert CM, Ardissino D, Atzmon G, Barnard J, Beaugerie L, Benjamin EJ, Boehnke M, Bonnycastle LL, Bottinger EP, Bowden DW, Bown MJ, Chambers JC, Chan JC, Chasman D, Cho J, Chung MK, Cohen B, Correa A, Dabelea D, Daly MJ, Darbar D, Duggirala R, Dupuis J, Ellinor PT, Elosua R, Erdmann J, Esko T, Färkkilä M, Florez J, Franke A, Getz G, Glaser B, Glatt SJ, Goldstein D, Gonzalez C, Groop L, Haiman C, Hanis C, Harms M, Hiltunen M, Holi MM, Hultman CM, Kallela M, Kaprio J, Kathiresan S, Kim B-J, Kim YJ, Kirov G, Kooner J, Koskinen S, Krumholz HM, Kugathasan S, Kwak SH, Laakso M, Lehtimäki T, Loos RJF, Lubitz SA, Ma RCW, MacArthur DG, Marrugat J, Mattila KM, McCarroll S, McCarthy MI, McGovern D, McPherson R, Meigs JB, Melander O, Metspalu A, Neale BM, Nilsson PM, O'Donovan MC, Ongur D, Orozco L, Owen MJ, Palmer CNA, Palotie A, Park KS, Pato C, Pulver AE, Rahman N, Remes AM, Rioux JD, Ripatti S, Roden DM, Saleheen D, Salomaa V, Samani NJ, Scharf J, Schunkert H, Shoemaker MB, Sklar P, Soininen H, Sokol H, Spector T, Sullivan PF, Suvisaari J, Tai ES, Teo YY, Tiinamaija T, Tsuang M, Turner D, Tusie-Luna T, Vartiainen E, Ware JS, Watkins H, Weersma RK, Wessman M, Wilson JG, Xavier RJ, Neale BM, Daly MJ, MacArthur DG. Genome aggregation database consortium: the mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434–43.
6. Abramovs N, Brass A, Tassabehji M. Gevir is a continuous gene-level metric that uses variant distribution patterns to prioritize disease candidate genes. Nat Genet. 2020;52(1):35–9.
7. Rapaport F, Boisson B, Gregor A, Béziat V, Boisson-Dupuis S, Bustamante J, Jouanguy E, Puel A, Rosain J, Zhang Q, Zhang S-Y, Gleeson JG, Quintana-Murci L, Casanova J-L, Abel L, Patin E. Negative selection on human genes underlying inborn errors depends on disease outcome and both the mode and mechanism of inheritance. Proc Natl Acad Sci. 2021;118(3):2001248118.
8. Bartha I, di Iulio J, Venter JC, Telenti A. Human gene essentiality. Nat Rev Genet. 2017;19(1):51–62.
9. Fuller ZL, Berg JJ, Mostafavi H, Sella G, Przeworski M. Measuring intolerance to mutation in human genetics. Nat Genet. 2019;51(5):772–6.
10. Coe BP, Stessman HAF, Sulovari A, Geisheker MR, Bakken TE, Lake AM, Dougherty JD, Lein ES, Hormozdiari F, Bernier RA, Eichler EE. Neurodevelopmental disease genes implicated by de novo mutation and copy number variation morbidity. Nat Genet. 2019;51(1):106–16.
11. Satterstrom FK, Kosmicki JA, Wang J, Breen MS, De Rubeis S, An J-Y, Peng M, Collins R, Grove J, Klei L, Stevens C, Reichert J, Mulhern MS, Artomov M, Gerges S, Sheppard B, Xu X, Bhaduri A, Norman U, Brand H, Schwartz G, Nguyen R, Guerrero EE, Dias C, Aleksic B, Anney R, Barbosa M, Bishop S, Brusco A, Bybjerg-Grauholm J, Carracedo A, Chan MCY, Chiocchetti AG, Chung BHY, Coon H, Cuccaro ML, Curró A, Dalla Bernardina B, Doan R, Domenici E, Dong S, Fallerini C, Fernández-Prieto M, Ferrero GB, Freitag CM, Fromer M, Gargus JJ, Geschwind D, Giorgio E, Giorgio E, González-Peñas J, Guter S, Halpern D, Hansen-Kiss E, He X, Herman GE, Hertz-Picciotto I, Hougaard DM, Hultman CM, Ionita-Laza I, Jacob S, Jamison J, Jugessur A, Kaartinen M, Knudsen GP, Kolevzon A, Kushima I, Lee SL, Lehtimäki T, Lim ET, Lintas C, Lipkin WI, Lopergolo D, Lopes F, Ludena Y, Maciel P, Magnus P, Mahjani B, Maltman N, Manoach DS, Meiri G, Menashe I, Miller J, Minshew N, Montenegro EMS, Moreira D, Morrow EM, Mors O, Mortensen PB, Mosconi M, Muglia P, Neale BM, Nordentoft M, Ozaki N, Palotie A, Parellada M, Passos-Bueno MR, Pericak-Vance M, Persico AM, Pessah I, Puura K, Reichenberg A, Renieri A, Riberi E, Robinson EB, Samocha KE, Sandin S, Santangelo SL, Schellenberg G, Scherer SW, Schlitt S, Schmidt R, Schmitt L, Silva IMW, Singh T, Siper PM, Smith M, Soares G, Stoltenberg C, Suren P, Susser E, Sweeney J, Szatmari P, Tang L, Tassone F, Teufel K, Trabetti E, Trelles MdP, Walsh CA, Weiss LA, Werge T, Werling DM, Wigdor EM, Wilkinson E, Willsey AJ, Yu TW, Yu MHC, Yuen R, Zachi E, Agerbo E, Als TD, Appadurai V, Bækvad-Hansen M, Belliveau R, Buil A, Carey CE, Cerrato F, Chambert K, Churchhouse C, Dalsgaard S, Demontis D, Dumont A, Goldstein J, Hansen CS, Hauberg ME, Hollegaard MV,

Howrigan DP, Huang H, Maller J, Martin AR, Martin J, Mattheisen M, Moran J, Pallesen J, Palmer DS, Pedersen CB, Pedersen MG, Poterba T, Poulsen JB, Ripke S, Schork AJ, Thompson WK, Turley P, Walters RK, Betancur C, Cook EH, Gallagher L, Gill M, Sutcliffe JS, Thurm A, Zwick ME, Børglum AD, State MW, Cicek AE, Talkowski ME, Cutler DJ, Devlin B, Sanders SJ, Roeder K, Daly MJ, Buxbaum JD. Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. Cell. 2020;180(3):568–58423.

12. Singh T, Walters JTR, Johnstone M, Curtis D, Suvisaari J, Torniainen M, Rees E, Iyegbe C, Blackwood D, McIntosh AM, Kirov G, Geschwind D, Murray RM, Di Forti M, Bramon E, Gandal M, Hultman CM, Sklar P, Palotie A, Sullivan PF, O'Donovan MC, Owen MJ, Barrett JC, Study INTERVAL, Consortium U. The contribution of rare variants to risk of schizophrenia in individuals with and without intellectual disability. Nat Genet. 2017;49(8):1167–73.
13. Howrigan DP, Rose SA, Samocha KE, Fromer M, Cerrato F, Chen WJ, Churchhouse C, Chambert K, Chandler SD, Daly MJ, Dumont A, Genovese G, Hwu H-G, Laird N, Kosmicki JA, Moran JL, Roe C, Singh T, Wang S-H, Faraone SV, Glatt SJ, McCarroll SA, Tsuang M, Neale BM. Exome sequencing in schizophrenia-affected parent-offspring trios reveals risk conferred by protein-coding de novo mutations. Nat Neurosci. 2020;23(2):185–93.
14. Deciphering developmental disorders study: prevalence and architecture of de novo mutations in developmental disorders. Nature. 2017;542:433–8.
15. Boukas L, Bjornsson HT, Hansen KD. Promoter CpG density predicts downstream gene loss-of-function intolerance. Am J Hum Genet. 2020;107(3):487–98.
16. Chen Y, Xu D. Understanding protein dispensability through machine-learning analysis of high-throughput data. Bioinformatics. 2005;21(5):575–81.
17. Acencio ML, Lemke N. Towards the prediction of essential genes by integration of network topology, cellular localization and biological process information. BMC Bioinform. 2009;10(1):290.
18. Deng J, Deng L, Su S, Zhang M, Lin X, Wei L, Minai AA, Hassett DJ, Lu LJ. Investigating the predictability of essential genes across distantly related organisms using an integrative approach. Nucl Acids Res. 2011;39(3):795–807.
19. Hasan MA, Lonardi S. DeeplyEssential: a deep neural network for predicting essential genes in microbes. BMC Bioinform. 2020;21(14):367.
20. Han X, Chen S, Flynn E, Wu S, Wintner D, Shen Y. Distinct epigenomic patterns are associated with haploinsufficiency and predict risk genes of developmental disorders. Nat Commun. 2018;9(1):2138.
21. Whalen S, Schreiber J, Noble WS, Pollard KS. Navigating the pitfalls of applying machine learning in genomics. Nat Rev Genet. 2022;23:169–81.
22. Rehm HL, Berg JS, Brooks LD, Bustamante CD, Evans JP, Landrum MJ, Ledbetter DH, Maglott DR, Martin CL, Nussbaum RL, Plon SE, Ramos EM, Sherry ST, Watson MS. ClinGen: he clinical genome resource. N Engl J Med. 2015;372(23):2235–42.
23. Bult CJ, Blake JA, Smith CL, Kadin JA, Richardson JE. The mouse genome database group: mouse genome database (MGD) 2019. Nucl Acids Res. 2019;47(D1):801–6.
24. Hart T, Tong AHY, Chan K, Van Leeuwen J, Seetharaman A, Aregger M, Chandrashekhar M, Hustedt N, Seth S, Noonan A, Habsid A, Sizova O, Nedyalkova L, Climie R, Tworzyanski L, Lawson K, Sartori MA, Alibeh S, Tieu D, Masud S, Mero P, Weiss A, Brown KR, Usaj M, Billmann M, Rahman M, Costanzo M, Myers CL, Andrews BJ, Boone C, Durocher D, Moffat J. Evaluation and design of genome-wide CRISPR/SpCas9 knockout screens. G3: Genes Genomes Genet. 2017;7(8):2719–27.
25. Kruschke JK. Doing bayesian data analysis: a tutorial with R and BUGS. Burlington: Academic Press; 2011.
26. Nair V, Hinton GE. Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th international conference on international conference on machine learning. ICML'10, Omnipress, USA 2010, pp. 807–814.
27. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Lear Res. 2014;15:1929–58.
28. Kingma DP, Adam BJ. A method for stochastic optimization 2014.
29. Zhu Y, Chen Z, Zhang K, Wang M, Medovoy D, Whitaker JW, Ding B, Li N, Zheng L, Wang W. Constructing 3D interaction maps from 1D epigenomes. Nat Commun. 2016;7(1):10812.
30. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics. 2011;27(12):1739–40.
31. The Gene Ontology Consortium. The gene ontology resource: 20 years and still going strong. Nucl Acids Res. 2019;47(D1):330–8.
32. Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, Sidiropoulos K, Cook J, Gillespie M, Haw R, Loney F, May B, Milacic M, Rothfels K, Sevilla C, Shamovsky V, Shorser S, Varusai T, Weiser J, Wu G, Stein L, Hermjakob H, D'Eustachio P. The reactome pathway knowledgebase. Nucl Acids Res. 2020;48(D1):498–503.
33. Barrera LA, Vedenko A, Kurland JV, Rogers JM, Gisselbrecht SS, Rossin EJ, Woodard J, Mariani L, Kock KH, Inukai S, Siggers T, Shokri L, Gordân R, Sahni N, Cotsapas C, Hao T, Yi S, Kellis M, Daly MJ, Vidal M, Hill DE, Bulyk ML. Survey of variation in human transcription factors reveals prevalent DNA binding changes. Science. 2016;351(6280):1450–4.
34. Giurgiu M, Reinhard J, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, Montrone C, Ruepp A. Corum: the comprehensive resource of mammalian protein complexes-2019. Nucl Acids Res. 2019;47(D1):559–63.
35. Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, Ophir R, Bar-Even A, Horn-Saban S, Safran M, Domany E, Lancet D, Shmueli O. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics. 2005;21(5):650–9.
36. Luck K, Kim D-K, Lambourne L, Spirohn K, Begg BE, Bian W, Brignall R, Cafarelli T, Campos-Laborie FJ, Charloteaux B, Choi D, Coté AG, Daley M, Deimling S, Desbuleux A, Dricot A, Gebbia M, Hardy MF, Kishore N, Knapp JJ, Kovács IA, Lemmens I, Mee MW, Mellor JC, Pollis C, Pons C, Richardson AD, Schlabach S, Teeking B, Yadav A, Babor M, Balcha D, Basha O, Bowman-Colin C, Chin S-F, Choi SG, Colabella C, Coppin G, D'Amata C, De Ridder D, De Rouck S, Duran-Frigola M, Ennajdaoui H, Goebels F, Goehring L, Gopal A, Haddad G, Hatchi E, Helmy M, Jacob Y, Kassa Y, Landini S, Li R, van Lieshout N, MacWilliams A, Markey D, Paulson JN, Rangarajan S, Rasla J, Rayhan A, Rolland T, San-Miguel A, Shen Y, Sheykhkarimli D, Sheynkman GM, Simonovsky E, Taşan M, Tejeda A, Tropepe V, Twizere J-C, Wang Y, Weatheritt RJ, Weile J, Xia Y, Yang X, Yeger-Lotem E, Zhong Q, Aloy P, Bader GD, De Las Rivas J, Gaudet S, Hao T, Rak J, Tavernier J, Hill DE, Vidal M, Roth FP, Calderwood MA. A reference map of the human binary protein interactome. Nature. 2020;580(7803):402–8.

37. Huang Y-F. Dissecting genomic determinants of positive selection with an evolution-guided regression model. Mol Biol Evolut. 2022;39(1):291.
38. Huang Y-F. Unified inference of missense variant effects and gene constraints in the human genome. PLOS Genet. 2020;16(7):1008922.
39. Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. Genic intolerance to functional variation and the interpretation of personal genomes. PLOS Genet. 2013;9(8):1003709.
40. Liu X, Li C, Mou C, Dong Y, Tu Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site snvs. Genome Med. 2020;12(1):103.
41. Dolan ME, Baldarelli RM, Bello SM, Ni L, McAndrews MS, Bult CJ, Kadin JA, Richardson JE, Ringwald M, Eppig JT, Blake JA. Orthology for comparative genomics in the mouse genome database. Mamm Genome. 2015;26(7):305–13.
42. Ho DE, Imai K, King G, Stuart EA. MatchIt: nonparametric preprocessing for parametric causal inference. J Stat Softw. 2011;42(8):1–28.
43. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. Bioinformatics. 2005;21(20):3940–1.
44. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44:837–45.
45. Lappalainen I, Lopez J, Skipper L, Hefferon T, Spalding JD, Garner J, Chen C, Maguire M, Corbett M, Zhou G, Paschall J, Ananiev V, Flicek P, Church DM. DbVar and DGVa: public archives for genomic structural variation. Nucl Acids Res. 2013;41(D1):936–41.
46. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, Boychenko V, Hunt T, Kay M, Mukherjee G, Rajan J, Despacio-Reyes G, Saunders G, Steward C, Harte R, Lin M, Howald C, Tanzer A, Derrien T, Chrast J, Walters N, Balasubramanian S, Pei B, Tress M, Rodriguez JM, Ezkurdia I, van Baren J, Brent M, Haussler D, Kellis M, Valencia A, Reymond A, Gerstein M, Guigo R, Hubbard TJ. GENCODE: the reference human genome annotation for the encode project. Genome Res. 2012;22(9):1760–74.
47. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005;15(8):1034–50.
48. Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson A, Kampf C, Sjostedt E, Asplund A, Olsson I, Edlund K, Lundberg E, Navani S, Szigyarto CA-K, Odeberg J, Djureinovic D, Takanen JO, Hober S, Alm T, Edqvist P-H, Berling H, Tegel H, Mulder J, Rockberg J, Nilsson P, Schwenk JM, Hamsten M, von Feilitzen K, Forsberg M, Persson L, Johansson F, Zwahlen M, von Heijne G, Nielsen J, Ponten F. Tissue-based map of the human proteome. Science. 2015;347(6220):1260419.
49. The GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 2020;369(6509):1318–30.
50. Ku M, Jaffe JD, Koche RP, Rheinbay E, Endoh M, Koseki H, Carr SA, Bernstein BE. H2a.z landscapes and dual modifications in pluripotent and multipotent stem cells underlie complex genome regulatory functions. Genome Biol. 2012;13(10):85.
51. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucl Acids Res. 2005;33(suppl 1):514–7.
52. Veitia RA. Exploring the molecular etiology of dominant-negative mutations. Plant Cell. 2007;19(12):3843–51.
53. Sherman BT, Hao M, Qiu J, Jiao X, Baseler MW, Lane HC, Imamichi T, Chang W. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). Nucl Acids Res. 2022;50(W1):216–21.
54. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using david bioinformatics resources. Nat Protoc. 2009;4(1):44–57.
55. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A. PANTHER: a library of protein families and subfamilies indexed by function. Genome Res. 2003;13(9):2129–41.
56. Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, Connor R, Funk K, Kelly C, Kim S, Madej T, Marchler-Bauer A, Lanczycki C, Lathrop S, Lu Z, Thibaud-Nissen F, Murphy T, Phan L, Skripchenko Y, Tse T, Wang J, Williams R, Trawick BW, Pruitt KD, Sherry ST. Database resources of the national center for biotechnology information. Nucl Acids Res. 2022;50(D1):20–6.
57. Cohen ASA, Simotas C, Webb BD, Shi H, Khan WA, Edelmann L, Scott SA, Singh R. Haploinsufficiency of the basic helix-loop-helix transcription factor hand2 causes congenital heart defects. Am J Med Genet Part A. 2020;182(5):1263–7.
58. Da Costa L, O'Donohue M-F, van Dooijeweert B, Albrecht K, Unal S, Ramenghi U, Leblanc T, Dianzani I, Tamary H, Bartels M, Gleizes P-E, Wlodarski M, MacInnes AW. Molecular approaches to diagnose diamond-blackfan anemia: the eurodba experience. Eur J Med Genet. 2018;61(11):664–73.
59. Gripp KW, Curry C, Olney AH, Sandoval C, Fisher J, Chong JX-L, for Mendelian Genomics UC, Pilchman L, Sahraoui R, Stabley DL, Sol-Church K. Diamond-blackfan anemia with mandibulofacial dystostosis is heterogeneous, including the novel dba genes tsr2 and rps28. Am J Med Genet Part A. 2014;164(9):2240–9.
60. Seidman JG, Seidman C. Transcription factor haploinsufficiency: when half a loaf is not enough. J Clin Investig. 2002;109(4):451–5.
61. Veitia RA. Exploring the etiology of haploinsufficiency. BioEssays. 2002;24(2):175–84.
62. Johnson AF, Nguyen HT, Veitia RA. Causes and effects of haploinsufficiency. Biol Rev. 2019;94(5):1774–85.
63. Bergendahl LT, Gerasimavicius L, Miles J, Macdonald L, Wells JN, Welburn JPI, Marsh JA. The role of protein complexes in human genetic disease. Prot Sci. 2019;28(8):1400–11.
64. Dang VT, Kassahn KS, Marcos AE, Ragan MA. Identification of human haploinsufficient genes and their genomic proximity to segmental duplications. Eur J Hum Genet. 2008;16(11):1350–7.
65. Ziegler A, Colin E, Goudenège D, Bonneau D. A snapshot of some pli score pitfalls. Hum Mutat. 2019;40(7):839–41.
66. Wood SN, Pya N, Säfken B. Smoothing parameter and model selection for general smooth models. J Am Stat Assoc. 2016;111(516):1548–63.

Publisher's Note