Distinct Epigenomic Patterns Are Associated with Haploinsufficiency and Predict Risk Genes of Developmental Disorders

Xinwei Han[1,2,*], Siying Chen[1,3,*], Elise Flynn[1,3], Dana Wintner[4], Yufeng Shen[1,5]

1. Department of Systems Biology, Columbia University, New York, NY
2. Department of Pediatrics, Columbia University, New York, NY
3. The Integrated Program in Cellular, Molecular and Biomedical Studies, Columbia University, New York, NY
4. Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY
5. Department of Biomedical Informatics, Columbia University, New York, NY

* These authors contributed to the work equally.

Correspondence: ys2411@cumc.columbia.edu (YS)

Abstract

Haploinsufficiency is a major mechanism of genetic risk in developmental disorders (DD). Accurate prediction of haploinsufficient genes is essential for prioritizing and interpreting deleterious variants. Current methods based on mutation intolerance in population data suffer from inadequate power for genes with short transcripts. Here we showed haploinsufficiency is strongly associated with epigenomic patterns, and then developed a new computational method (Episcore) to predict haploinsufficiency from epigenomic data using a Random Forest model. Based on data from recent exome sequencing studies of DD, we show that Episcore performs favorably to current methods in prioritizing loss of function *de novo* variants. Our method enables new applications of epigenomic data, and facilitates discovery and interpretation of novel candidate risk variants in genetic studies of DD.

Introduction

Haploinsufficiency (HIS) due to hemizygous deletions or heterozygous likely-gene-disrupting (LGD) variants plays a central role in the pathogenesis of various diseases. Recent large-scale exome and genome sequencing studies of developmental disorders, including autism, intellectual disability, developmental delay, and congenital heart disease [1-5], have estimated that *de novo* LGD mutations explain the cause of a significant portion of patients with these developmental disorders and the enrichment rate of *de novo* LGD variants indicates about half of these variants are associated with disease risk. However, relatively few genes have multiple LGD variants ("recurrence") in a cohort [1,2,6], lacking of which provides insufficient statistical evidence to distinguish individual risk genes from the ones with random mutations [7]. On the other hand, most of the enrichment of LGD variants can be explained by HIS genes [6]. Therefore, a comprehensive catalog of HIS genes can greatly help interpreting and prioritizing mutations in genetic studies.

Currently, there are two main approaches of predicting HIS genes based on high-throughput data. Huang et al. uses a combination of genetic, transcriptional and protein-protein interaction features from various sources to estimate haploinsufficient probabilities for 12,443 genes [8], on par with Steinberg et al., which generated the probabilities for more (over 19,700) human genes by a Support Vector Machine (SVM) model[9]. The other type is based on mutation intolerance [10-12] in populations that do not have early onset developmental disorders. Lek et al 2016 [11] estimated each gene's probability of haploinsufficiency (pLI: Probability of being Loss-of-function Intolerant) based on the depletion of observed rare LGD variants in over 60,000 exome sequencing samples. Although effective, ExAC pLI is biased towards genes with longer transcripts or higher background mutation rates, since the statistical power of assessing the significance depends on a relatively large expected number of rare LGD variants from background mutations.

We sought to predict HIS using epigenomic data that are orthogonal to genetic variants and generally independent of gene size. Our method is motivated by recent studies indicating that specific epigenetic patterns are associated with genes that are likely haploinsufficient. Specifically, genes with increased breadth of H3K4me3, typically associated with actively transcribing promoters, are enriched with tumor suppressor genes [13], which are predominantly haploinsufficient based on somatic mutation patterns [14]. Another study reported H3K4me3 breadth regulates transcriptional precision [15], which is critical for dosage sensitivity. These observations led us to hypothesize that haploinsufficient genes are tightly regulated by a combination of transcription factors and epigenetic modifications to achieve spatiotemporal precision of gene expression, and such regulation can be

detected by distinct patterns of epigenomic marks in relevant tissues and cell types. Based on this model, we developed a Random Forest–based method ("Episcore") using epigenetic data from the Epigenomic Roadmap [16] and ENCODE Projects [17] as input features and a few hundreds of curated HIS genes as positive training data. To assess the performance of prioritizing candidate risk variants in real-world genetic studies, we used large data sets of *de novo* mutations from recent studies of birth defects and neurodevelopmental disorders, and showed that Episcore had better performance than existing methods. Additionally, Episcore is less biased by gene length or background mutation rate and complementary to mutation-based metrics in HIS-based gene prioritization. Our analysis indicates that epigenetic features in stem cells, brain tissues, and fetal tissues have the highest contribution to Episcore.

Results

Haploinsufficient (HIS) and Haplosufficient (HS) genes show distinct distributions of epigenomic features

To examine the correlation of gene haploinsufficiency and epigenomic patterns, we analyzed ChIP-seq data from Roadmap and ENCODE projects, including active (H3K4me3, H3K9ac, and H2A.Z) and repressive (H3K27me3) promoter modifications, and marks associated with enhancers (H3K4me1, H3K27ac, DNase I hypersensitivity sites). We used the width of called ChIP-seq peaks for promoter features and counted the number of peaks within 20kb upstream or downstream of transcription start sites (TSS) for enhancer features, unless otherwise stated. As each histone modification is characterized in multiple cell types, we refer to the combination of an epigenetic modification and a cell type as one epigenetic feature.

Figure 1A shows the correlation among epigenetic features, and the correlation of epigenetic features and ExAC pLI score. As expected, active promoter or enhancer marks are highly correlated with each other and with ExAC pLI score, and anti-correlated with repressor marks in general. The repressor marks from stem cells or fetal tissues have positive correlations with active marks and ExAC pLI scores, suggesting many genes with bivalent marks in stem cells are likely haploinsufficient.

To further investigate the association of haploinsufficiency and patterns of epigenetic modifications, we compiled a list of 287 known HIS genes (Supplementary Table 1) from a recent study [8,18] and human-curated ClinGen dosage sensitivity map. We also collected a list of 717 HS

genes, of which one copy of each gene had been deleted in two or more subjects based on a CNV study in 2,026 healthy individuals [19]. For promoter features, HIS and HS genes clearly have distinct distributions of peak length (Figure 1B-D). HIS genes on average have larger peak length for both the active marker H3K4me3 (Figure 1B) and the repressive marker H3K27me3 (Figure 1C), suggesting the difference between HIS and HS genes is not only on the level of expression but also on distinct mechanisms of regulation. Furthermore, other epigenetic modifications associated with active promoters, including H2A.Z and H3K9ac, also display wider peaks upstream of HIS genes (Supplementary Figure 1 A and B). In addition, HIS and HS genes also differ in the number of interacting enhancers. A naive approach was attempted first by assigning enhancers to genes by counting all the peaks of enhancer marks in 40kb windows around transcription start sites. As result, HIS genes have a slightly larger number of H3K4me peaks than HS genes ($p < 10^{-4}$, permutation test, Supplementary Figure 1C). To better infer enhancer-gene relationship, we adopted a recently published method named "EpiTensor" [20], which decomposes a 3D tensor representation of histone modifications, DNase-Seq, and RNA-Seq data to find associations between distant genomic regions. When restricted to pre-defined topologically-associated domains (TADs), associated regions identified by EpiTensor correspond well to enhancer-promoter interactions found by Hi-C. Epitenosr revealed much larger differences between HIS and HS genes than the naive approach. HIS genes have a median of 9 interacting enhancers, while HS genes have a median of 0 ($p < 10^{-4}$, permutation test, Supplementary Figure 1C). When averaged across tissues, HIS genes shifts towards a larger number of mean interacting enhancers, as compared to HS genes (Figure 1D), supporting the notion that HIS genes have more regulatory complexity.

Predicting haploinsufficiency with epigenomic features

To leverage the strong association between epigenomic patterns and gene haploinsufficiency, we developed a computational method to predict haploinsufficiency using Random Forest (Figure 2A) and other supervised learning models (Supplementary Figure 2 A and B). The input features included peak length of 4 promoter marks (H3K4me3, H3K9ac, H2A.Z and H3K27me3) and the number of EpiTensor-inferred interacting enhancers in various tissues. Performance evaluation by 10-fold cross validation and AUC (Area Under Curve) in ROC (Receiver Operating Characteristic) curves showed that all of these methods achieved high AUC values of 0.86~0.88 (Figure 2B and Supplementary Figure 2 A and B), supporting the utility of epigenetic features in predicting haploinsufficiency. As Random Forest performs best and requires minimal parameterization, results

from Random Forest are chosen as final metrics measuring the probability of being haploinsufficient, termed "Episcore" (Supplementary Table 2). Despite completely different data and methods used, Episcore and ExAC pLI score displayed overall concordance. The median Episcore of likely HIS genes defined by pLI is two times larger than the median score of likely HS genes ($p < 10^{-5}$) (Supplementary Figure 2C).

Episcore provides better prioritization of *de novo* LGD variants in developmental disorders

A major goal of predicting haploinsufficiency is to facilitate prioritization of variants identified in genetic studies of developmental disorders. We compared Episcore with pLI scores from ExAC [11], S_{het} values (denoting selective effects of heterozygous LGD variants) [12], and ranks of mouse heart expression level [21], using *de novo* LGD variants identified in a recently published whole exome sequencing study of 1,365 trio families with congenital heart disease (CHD) [22]. LGD variants include frameshift, nonsense and canonical splice site mutations. Genes with all 4 metrics were included for comparison, although we note Episcore (19,430 genes) made predictions for more genes than pLI (18,225 genes), S_{het} (17,200 genes) and ranks of mouse heart expression level (17,624 genes, due to loss in orthologue matching). Different predictions are compared by the enrichment rate of variants. For the same number of top-ranked genes from each metric, we calculated the number of LGD variants located in these genes and estimated the number of LGD variants due to background mutation [23]. Across a wide range of top-ranked genes, Episcore always showed larger enrichment than ExAC pLI, S_{het}, or heart expression level (Figure 3A). We also applied the same approach to *de novo* synonymous variants identified in the CHD dataset and observed no enrichment (Supplementary Figure 3A). Additionally, we compared these predictions by precision-recall-like curve (PR-like) based on enrichment. Since the total number of positive variants (true disease-causing variants) is unknown, we used estimated number of "true positives" instead of "true positive rate (recall)" in this comparison. For top-ranked genes from each method, the number of true positives were estimated by subtracting expected number of LGD variants based on background mutation rate from the observed in these genes. We measured precision by dividing the estimated number of true positives by the total number of observed LGD variants in these genes. Across a wide range of precision, Episcore consistently showed superior recall compared to pLI, S_{het} and heart expression level (Figure 3B) and other methods [8,9] (Supplementary Figure 3 B and C).

We further assessed Episcore based on a second CHD WES cohort ("PCGC") of 2,645 parent-offspring trios from a recent publication [24]. Genes with multiple LGD variants in this data

set, or the ones with one LGD variant and at least one LGD variant in the first cohort ("DDD CHD") have much higher Episcore compared to genes with LGD variants in controls (unaffected siblings in Simons Simplex Collection autism study [25])(Supplementary Figure 4).

Episcore provides complementary information to mutation intolerance metrics

Haploinsufficiency predicted by mutation intolerance in a general population (such as ExAC pLI metric) is intrinsically biased towards genes with longer CDS (coding sequence) lengths or higher background mutation rates. The distribution of genes with top 20% pLI scores shifts towards longer CDS length or higher background mutation rate, as compared to the distribution of all genes or top 20% highly-expressed genes in developing heart [21] (Figure 3 C and D). Top 20% genes ranked by Episcore have similar distribution to all genes in the genome for either CDS length or background mutation rate (Figure 3 C and D).

Since Episcore and pLI use distinct types of input data, a combination of these two scores might achieve better performance. We used a logistic regression method to integrate Episcore and pLI based on a collection of 4,293 trio families affected by various developmental disorders (DD) [6]. Specifically, we used a total of 45 genes with *de novo* LGD variants in 3 or more probands as positives, and randomly sampled 45 genes from genes with no observed *de novo* LGD variant as negatives to estimate coefficients in the logistic model. Both Episcore and pLI have significant coefficients ($P < 10^{-5}$), supporting these two methods convey complementary information. We found that the resulting meta-score outperformed Episcore or pLI alone (Figure 3 E and F). The meta-score obtained the same sensitivity as pLI, while maintaining the precision equal to Episcore (Figure 3F).

Brain tissues, fetal tissues, and stem cells have highest contribution to the predicted haploinsufficiency.

To evaluate contribution of each epigenetic feature to HIS prediction, we calculated Spearman correlation coefficients between each feature and Episcore. These correlation coefficients were analyzed in two ways. First, we grouped them based on the molecular entities they represent, such that the same epigenetic modification from different tissues would be in one group. Each of the 5 resulting categories has distinct distributions of Spearman correlation coefficients, suggesting different contributions to Episcore (Figure 4A). Except for the repressive mark H3K27me3, most of them have larger correlation coefficients than gene expression values, suggesting these features and

our model not merely reflected expression abundance but also epigenetic regulation specific to HIS genes.

Second, we grouped the correlation coefficients based on tissues, thus each tissue had several correlation coefficients for different epigenetic modifications. As contributions of different epigenetic modifications varied considerably, we converted every correlation coefficient to a Z-score using the mean and standard deviation of each epigenetic modification and further averaged them for each tissue. The averaged Z-score represents the importance of this tissue to haploinsufficiency prediction. In general, stem cells and neural tissues have large average Z-scores (Figure 4B). Interestingly, for tissues in the same category, fetal tissues usually have larger average Z-scores than postnatal tissues.

Finally, to illustrate the contribution of different tissues to HIS, we examined in detail the histone modifications around TSS of several known HIS genes. A CHD risk gene recently discovered through *de novo* LGD variants, *RBFOX2* [5], was shown in Figure 4C. *RBFOX2* had expanded H3K4me3 and H3K9ac in stem/fetal cells, and heart and brain tissues, but not in blood cells. Correspondingly, H3K27me3 displayed the reverse pattern, extensive in blood cells but limited in other tissues. On the contrary, a known house-keeping gene, *CWC22*, showed unanimous amount of histone modification across tissues. Together, these showcased that HIS genes had wide deposition of epigenetic modification specifically in certain tissues.

Discussion

In this study we showed there is a strong correlation between epigenomics patterns across tissues and gene haploinsufficiency, and then developed a computational method (Episcore) to predict HIS using epigenomic features and Random Forest method. Episcore had superior performance in prioritization of *de novo* LGD variants in congenital heart disease and neurodevelopmental disorders, compared to mutation intolerance metrics such as ExAC pLI [11]. Additionally, we showed that Episcore and pLI are complement to each other and can be combined to achieve better performance.

Existing HIS prediction methods based on intolerance of mutations or gene network properties have inadequate statistical power and accuracy in genes with small transcript size and difficulty in interpretability in specific diseases or biases towards well-studied genes. Epigenomic data have several advantages to address these issues: (a) orthogonal to genetic mutations, and therefore provide additional information that could improve power; (b) much less biased by

transcript size, and will be most helpful to predict HIS of genes with shorter transcripts; (c) intrinsically tissue specific, lending a direct link to interpretation in specific diseases; (d) integrating large amount of data that are not biased towards well-studied genes. These advantages contribute to the superior performance of Episcore in prioritizing *de novo* LGD variants from exome sequencing studies.

There are likely a variety of mechanisms underlining the correlation of epigenomics patterns and haploinsufficiency. First, broad H3K4me3 peaks contributed most to Episcore prediction of HIS. Broad H3K4me3 peaks are associated with reduced transcriptional noise at cell population and single cell levels [15], which is probably required to maintain precision expression level of HIS genes in specific cell types and developmental stages. Second, associated with transcriptional programs in hematopoietic differentiation, "regulatory complexity" is required to achieve cell-type specific expression patterns of the lineage defining genes [26]. Consistently, we found the number of enhancers interacting with the promotor of a gene is highly correlated with predicted HIS score. Third, many HIS genes are regulators that define cell lineages during differentiation but have low expression in stem cells. Bivalent chromatin domains in embryonic stem cells, in which both active marker H3K4me3 and repressor marker H3K27me3 are present, are generally associated with lineage control genes [27]. We observed that H3K27me3 are positively correlated with H3K4me3 in stem cells, mutation intolerance, known HIS genes (Figure 1A and 1C) and Episcore predicted HIS scores (Figure 4A), supporting the association of bivalent marks with HIS. Finally, we found epigenetic features from stem cells and fetal tissues contribute most to prediction, highlighting the importance of early developmental stages in the arise of HIS.

Currently, Episcore is limited by availability and resolution of epigenomic data, especially cell-type specific data from complex tissues or organs such as the brain, and data from tissues at various developmental stages. Complex developmental disorders, such as autism, involve a large number of cell types during a broad range of developmental stages. It is critical to generate and integrate much more fine-grained epigenomic data from cells of specific types at specific time points to improve genetic discoveries by Episcore approach in studies of such diseases. We expect such data sets will become available in near future from ongoing projects [28-30], and will enable us to improve prediction of HIS and facilitate novel discoveries in genetic studies.

Methods

Collection and Preprocessing of Training Genes

In this study, we used Ensembl release 75 for gene annotation and TSS (transcription start site) locations. All genomic coordinates are based on hg19 human genome assembly. Any non-hg19 coordinates were lifted over to hg19 using UCSC LiftOver tool (https://genome.ucsc.edu/cgi-bin/hgLiftOver). Conversion of gene symbols to Ensembl IDs were based on annotation tables downloaded from Ensembl BioMart.

Positive training set data (curated haploinsufficient genes) were collected from these two sources: (1) haploinsufficent training genes used in previous studies [8,18] and (2) genes with haploinsufficient score of 3 in ClinGen Dosage Sensitivity Map (http://www.ncbi.nlm.nih.gov/projects/dbvar/clingen/). For the negative training set (curated haplosufficient genes), we used genes deleted in two or more healthy people, based on CNVs detected in 2,026 normal individuals [19]. Only genes with half or more of its length covered by any deletion were considered "deleted" in an individual.

The raw training set may have some false positives and false negatives, as it contained results from automated literature mining that is known to give noisy output. To optimize the performance, we did the following pruning of the raw training set: (1) we only kept protein-coding genes in autosomes, as non-protein-coding genes or genes on sex chromosomes may be under different epigenetic regulation; (2) from the positive training set, we removed genes with both ExAC pLI ≤ 0.1 and expected loss-of-function variants > 10 [11] and (3) from the negative training set, we removed genes with both pLI ≥ 0.9 and expected loss-of-function variants > 10. After filtering, the positive training set has 287 genes and the negative training set has 717 genes. The full list of training genes is available in Supplementary Table 1.

Preprocessing of Epigenetic Feature Data

The uniformly processed peak calling results of Roadmap and ENCODE projects were downloaded from http://egg2.wustl.edu/roadmap/web_portal/processed_data.html. For promoter features (H2A.Z, H3K27me3, H3K4me3 and H3K9ac), "GappedPeaks" were used to allow for broad domains of ChIP-seq signal. The assignment of a GapppedPeak to a gene follows these steps in order: (1) for each gene, only TSSs of Ensembl canonical transcripts were used. (2) assigned a GappedPeak to a TSS if the GappedPeak overlaps with the upstream 5kb to downstream 1kb region

around the TSS. This definition of basal cis-regulatory region around promoter follows GREAT tool [31]. Assigning one GappedPeak to multiple TSSs was allowed. (3) For TSSs having more than 1 GappedPeak assigned, kept the closest one. (4) For genes with multiple TSSs and hence mutilple assigned GappedPeaks, kept the longest GappedPeak. After these four steps, if one gene had been associated with a GappedPeak, then we used the width of the peak as an epigenetic feature in the following machine learning models. If a gene had no associated GappedPeak, then the peak width is 0.

To calculate the number of interacting enhancers of a gene, we used two approaches. In a simple approach, we counted peaks of ChIP-seq signals that are associated with enhancers. The ChIP-seq signals we used include H3K4me1, H3K27ac and DNase I hypersensitivity site, and each ChIP signal was counted and recorded separately. We used "NarrowPeak" instead of "GappedPeak" in the counting to better estimate the number of interacting enhancers, as enhancer regions are not long and GappedPeak has the risk of merging nearby ChIP-seq signals. For each gene, we counted peaks in (1) the surrounding TAD (Topologically Associated Domain), based on TADs reported in [32]; or (2) +/- 20kb of each TSS (Only TSSs of Ensembl canonical transcripts were used. For genes with multiple TSSs and thus several numbers of interacting enhancers, we kept the largest one). In a more advanced approach, we adapted EpiTensor [20] to infer gene-enhancer relationship. We made a few changes when using EpiTensor: (1) we used normalized coverage of ChIP-seq signal instead of raw coverage in Zhu et al. 2016 [20]; (2) we used the coverage of H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me3, H3K9me3, DNase I and RNA-seq as input for EpiTensor to balance between more input data types and more cell types included, as not every cell type has all these histone modifications characterized. The number of data types included are fewer than the ones used in Zhu et al. 2016 [20], but it could still achieve desirable performance (personal communications); (3) we used enhancer annotation from 15-state chromHMM (http://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html#core_15state), while the original EpiTensor paper [20] used results of an early study. Based on the output of EpiTensor, which predicts enhancer-promoter pairs, we counted the number of interacting enhancers each gene has in various tissues.

Finally, the results of peak width and interacting enhancers were consolidated into a matrix, with each row being a different gene and each column representing a combination of a tissue and a data type, e.g. "H3K4me3 peak width in foreskin fibroblast". One combination of a tissue and a data

type was referred to as one epigenetic feature in the main text. This matrix was used as input for machine learning models described in the following section.

Machine learning approaches to predict haploinsufficiency

Several machine learning approaches have been tried, including Random Forest, SVM and SVM with LASSO feature selection. Random Forest was implemented using R package "randomForest". SVM was implemented using R package "e1071". LASSO was implemented using R package "glmnet", with alpha value equal to 1. For each machine learning approach, method performance was assessed based on 100 runs of 10-fold cross-validation. In each run, 10% of the training genes were randomly selected and left out to form a test set for validation. The remaining data were used to train the model, after which the test set was used to calculate model sensitivity and specificity. We used R package "ROCR" to make an ROC curve based on the 100 runs and calculated AUC values.

For the prediction step, all training genes were used to train the model. For each machine learning approach, we made it to output the probabilities of being positive (i.e. probabilities of being HIS in our study). The whole process was repeated 30 times and we took the arithmetic mean of the 30 sets of probabilities as the final results.

Comparing Episcore and other metrics in variant prioritization

We used two approaches to compare Episcore and other metrics in variant prioritization, based upon customized "enrichment of *de novo* LGD variants", "the number of true-positives" and "precision". The formula to calculate these three statistics are as follows.

For any gene $\boldsymbol{i}$, the number of expected *de novo* LGD variants in each gene, $\boldsymbol{E_i}$, was calculated as:

$$\boldsymbol{E_i = 2 \times N \times R_i}$$

where $\boldsymbol{N}$ is the number of probands in the sequencing cohort and $\boldsymbol{R_i}$ is gene-specific LGD mutation rate. LGD variants include nonsense, frameshift and canonical splice site mutations. The background mutation rate per gene of each mutation type was obtained from Samocha et al. 2014 [23]. For each gene, $\boldsymbol{R_i}$ is the sum of background mutation rate of nonsense, frameshift and canonical splice site mutations.

For any gene set, the enrichment of *de novo* LGD variants, $\boldsymbol{D}$, was calculated as:

$$\boldsymbol{D} = \frac{\boldsymbol{M}}{\sum_i \boldsymbol{E_i}}$$

where $\boldsymbol{M}$ is the total number of observed *de novo* LGD variants in this gene set and $\boldsymbol{i}$ is any gene in the gene set. $\boldsymbol{M}$ was calculated using an empirical dataset. In this study, we used results from two whole exome sequencing studies on congenital heart disease [5,22] and another whole exome sequencing study on various developmental disorders [6].

For any gene set, the number of true positives, $\boldsymbol{TP}$, was calculated as:

$$\boldsymbol{TP} = \boldsymbol{M} - \sum_i \boldsymbol{E_i}$$

For any gene set, the precision (positive predictive value), $\boldsymbol{PPV}$, was calculated as:

$$\boldsymbol{PPV} = \frac{\boldsymbol{M} - \sum_i \boldsymbol{E_i}}{\boldsymbol{M}}$$

For each metric (Episcore, pLI, etc.), a series of top-ranked genes were selected, such as top 500 genes, top 2000 genes, etc. In the first approach, enrichment of de novo LGD variants, $\boldsymbol{D}$, was calculated for any set of top-ranked genes, and then enrichment values were plotted and compared, as shown in Figure 3A. In the second approach, the number of true positives, $\boldsymbol{TP}$, and the precision (true discovery rate), $\boldsymbol{PPV}$, were calculated for any set of top-ranked genes. $\boldsymbol{TP}$ and $\boldsymbol{PPV}$ were plotted and compared, as shown in Figure 3B. Ideally, recall (true positive rate) should be calculated and plotted along with precision (true discovery rate). However, for any empirical dataset, it is unknown to us whether or not a variant is disease-causing. In other words, when comparing variant prioritization results with the real roles of these variants, we cannot determine what part is false negative. Thus, we are unable to calculate true positive rate and have to use the number of true positives instead.

To examine the utility of Episcore in prioritizing genes with only one LGD mutation, we utilized two independent Congenital Heart Disease (CHD) cohorts: DDD (Deciphering Developmental Disorders consortium) CHD [22] and PCGC (Pediatric Cardiac Genomics Consortium) CHD [24]. Both these two study included trios from an earlier CHD study [21] to increase detection power. To avoid duplication, we removed these earlier trios from DDD CHD cohort and then selected genes based on the following two criteria: (1) only have 1 LGD variant reported in PCGC CHD WES cohort and (2) among top 3000 genes ranked by Episcore. For each of the selected genes, we enumerated the number of LGD variants it had in DDD CHD WES cohort. Finally, we found many of these genes selected based on Episcore and PCGC cohort also had LGD variant in DDD cohort.

Epigenetic features critical in the prediction

A Spearman correlation coefficient was calculated between each epigenetic feature and Episcore. One epigenetic feature here corresponds to a data type (like H3K4me3 peak width) in certain tissue/cell type (like foreskin fibroblast). To examine which data types are more important, these Spearman correlation coefficients were plotted by data type, e.g. correlation coefficients from H3K4me3 peak width were plotted in one section. To examine what tissue/cell types are more important, we calculated averaged z-score for each tissue/cell type. The average z-score is calculated following these two steps: (1) we converted every Spearman correlation coefficient to a Z-score using mean and standard deviation specific to each data type and (2) for each tissue/cell type, we averaged the Z-scores it has. To select example genes, epigenetic profiles were visualized in Integrative Genomic Viewer.

Acknowledgements

We thank Shuang Wu for helpful discussions. The work was partly supported by R01GM120609 (S.C. and Y.S.).

Author Contributions

All authors contributed to data analysis and interpretation and manuscript writing. X.H. curated and processed epigenomics data sets. Y.S. conceived and designed the study.

Competing Financial Interests statement.

None declared

Figure Legends

Figure 1. The potential of using epigenetic features to classify HIS genes. (A) Heatmap showing Spearman correlation between epigenetic features contains three groups of features: active promoter, repressive promoter and enhancer features. Epigenetic features inside one group strongly correlate with each other. Different feature types, including various histone modifications, histone variant and DNase I hypersensitivity sites, are color-coded. Above the heatmap, a bar denoting Spearman correction between epigenetic features and pLI shows many epigenetic features relate to HIS with varying degree. Data from stem cells or fetal tissues are also marked by color lines. (B-C) HIS and HS genes have different distributions of peak length from promoter features (B,

H3K4me3; C, H3K27me3). For each gene, peak length was averaged across tissues. (D) HIS and HS genes have different distributions of interacting enhancers inferred by Epitensor. For each gene, the number of interacting enhancers was averaged across tissues.

Figure 2. The Random Forest model to predict haploinsufficiency. (A) A flowchart of the method. (B) ROC curve of 10-fold cross-validation. The red curve is the average of 100 randomized cross-validation runs, with error bar showing standard deviation. The mean and median AUC of the 100 runs are 0.88 and 0.89, respectively.

Figure 3. Benchmark the performance of Episcore in variant prioritization using the empirical data. (A-B) Comparison of Episcore, pLI, S_{het} and heart expression level (HE) in variant prioritization using CHD exome sequencing data [22]. In (A), burden refers to the ratio between the number of *de novo* LGD variants observed in top genes ranked by each metric and the number of expected *de novo* LGD variants due to background mutation. Episcore has higher enrichment in top 1000-2500 genes and similar enrichment afterwards. In (B), true positive is the difference between the observed and expected *de novo* LGD variants. Precision is calculated by dividing the number of true positives by the number of observed *de novo* LGD variants. The blue curve for Episcore shifts upright, showing Episcore has better precision at the same number of true positives and vice versa. (C-D) Episcore has much less bias towards genes with longer CDS length (C) or larger background mutation rate (D). Grey histogram in the background represents CDS length or mutation rate of all genes in the genome. The blue curve for pLI shifts right, while the curves for Episcore and HE are similar to the distribution of all genes. (E-F) A combination of Episcore and pLI, the metascore, has better performance in variant prioritization when benchmarked using DDD exome sequencing data. Metascore is the output from a logistic regression model, using Episcore and pLI as input. Enrichment, true positive and precision were calculated similarly to (A-B).

Figure 4. Epigenetic features critical in the prediction. (A) Spearman correlation between epigenetic feature and Episcore. Features used in the Random Forest model, including H2A.Z, H3K27me3, H3K4me3, H3K9ac and the number of interacting enhancers, all have positive correlation with Episcore. Spearman correlation coefficients between gene expression level, measured in RPKM (reads per kilobase per million reads), and Episcore were also plotted for comparison. (B) Stem cells, and neural and fetal tissues are the most important ones in the prediction. The importance of each

tissue in generating Episcore is measured by average Z-score, which is converted from Spearman correlation coefficients between epigenetic feature and Episcore. (C) The epigenetic profile of an example HIS gene, *RBFOX2*, and a house-keeping gene, *CWC22*. Each small box represents 100bp region around TSS and the darkness of the color reflects averaged fold change of reads between ChIP-seq library and input.

References

1. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515**, 209-15 (2014).
2. Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216-21 (2014).
3. Hamdan, F.F. *et al.* De novo mutations in moderate or severe intellectual disability. *PLoS Genet* **10**, e1004772 (2014).
4. Deciphering Developmental Disorders, S. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* **519**, 223-8 (2015).
5. Homsy, J. *et al.* De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies. *Science (New York, N.Y.)* **350**, 1262-6 (2015).
6. McRae, J.F. *et al.* Prevalence, phenotype and architecture of developmental disorders caused by de novo mutation. *bioRxiv* (2016).
7. 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).
8. Huang, N., Lee, I., Marcotte, E.M. & Hurles, M.E. Characterising and predicting haploinsufficiency in the human genome. *PLoS Genet* **6**, e1001154 (2010).
9. Steinberg, J., Honti, F., Meader, S. & Webber, C. Haploinsufficiency predictions without study bias. *Nucleic Acids Res* **43**, e101 (2015).
10. Petrovski, S., Wang, Q., Heinzen, E.L., Allen, A.S. & Goldstein, D.B. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet* **9**, e1003709 (2013).
11. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-91 (2016).
12. Cassa, C.A. *et al.* Estimating the selective effects of heterozygous protein-truncating variants from human exome data. *Nat Genet* **49**, 806-810 (2017).
13. Chen, K. *et al.* Broad H3K4me3 is associated with increased transcription elongation and enhancer activity at tumor-suppressor genes. *Nat Genet* (2015).
14. Davoli, T. *et al.* Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell* **155**, 948-62 (2013).
15. Benayoun, B.A. *et al.* H3K4me3 breadth is linked to cell identity and transcriptional consistency. *Cell* **158**, 673-88 (2014).
16. Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-30 (2015).
17. Consortium, E.P. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
18. Dang, V.T., Kassahn, K.S., Marcos, A.E. & Ragan, M.A. Identification of human haploinsufficient genes and their genomic proximity to segmental duplications. *Eur J Hum Genet* **16**, 1350-7 (2008).
19. Shaikh, T.H. *et al.* High-resolution mapping and analysis of copy number variations in the human genome: a data resource for clinical and research applications. *Genome Res* **19**, 1682-90 (2009).
20. Zhu, Y. *et al.* Constructing 3D interaction maps from 1D epigenomes. *Nat Commun* **7**, 10812 (2016).
21. Zaidi, S. *et al.* De novo mutations in histone-modifying genes in congenital heart disease. *Nature* **498**, 220-3 (2013).

22. Sifrim, A. *et al.* Distinct genetic architectures for syndromic and nonsyndromic congenital heart defects identified by exome sequencing. *Nat Genet* **48**, 1060-5 (2016).
23. Samocha, K.E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nat Genet* **46**, 944-50 (2014).
24. Jin, S.C. *et al.* Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. *Nat Genet* (2017).
25. Krumm, N. *et al.* Excess of rare, inherited truncating mutations in autism. *Nat Genet* **47**, 582-8 (2015).
26. Gonzalez, A.J., Setty, M. & Leslie, C.S. Early enhancer establishment and regulatory locus complexity shape transcriptional programs in hematopoietic differentiation. *Nat Genet* (2015).
27. Vastenhouw, N.L. & Schier, A.F. Bivalent histone modifications in early embryogenesis. *Curr Opin Cell Biol* **24**, 374-86 (2012).
28. Stunnenberg, H.G., International Human Epigenome, C. & Hirst, M. The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. *Cell* **167**, 1145-1149 (2016).
29. Psych, E.C. *et al.* The PsychENCODE project. *Nat Neurosci* **18**, 1707-12 (2015).
30. Dekker, J. *et al.* The 4D Nucleome Project. *bioRxiv* (2017).
31. McLean, C.Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* **28**, 495-501 (2010).
32. Dixon, J.R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-80 (2012).

Figure 1

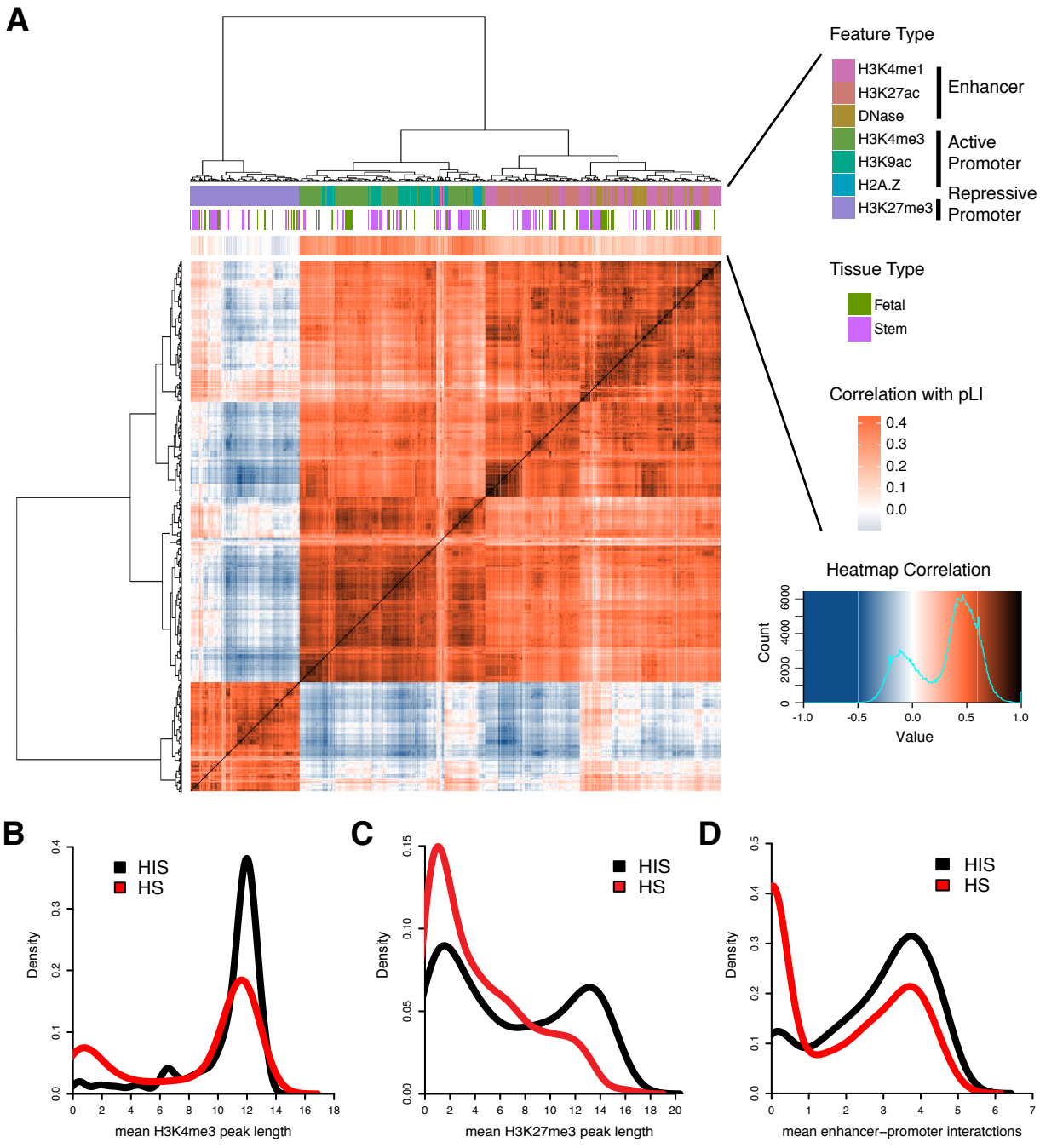

A
Feature Type
H3K4me1
H3K27ac
DNase
H3K4me3
H3K9ac
H2A.Z
H3K27me3
Enhancer
Active Promoter
Repressive Promoter
Tissue Type
Fetal
Stem
Correlation with pLI
0.4
0.3
0.2
0.1
0.0
Heatmap Correlation
Count
Value
B
HIS
HS
Density
mean H3K4me3 peak length
C
HIS
HS
Density
mean H3K27me3 peak length
D
HIS
HS
Density
mean enhancer–promoter interatctions

Figure 2

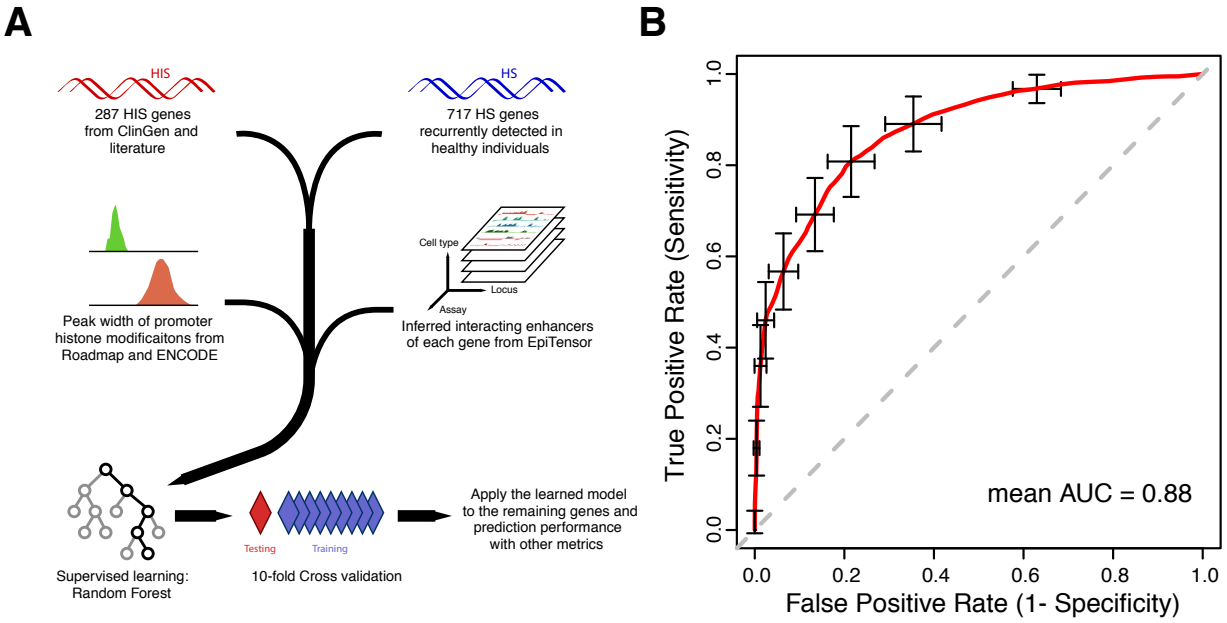

A
HIS
287 HIS genes
from ClinGen and
literature
HS
717 HS genes
recurrently detected in
healthy individuals
Peak width of promoter
histone modificaitons from
Roadmap and ENCODE
Cell type
Locus
Assay
Inferred interacting enhancers
of each gene from EpiTensor
Supervised learning:
Random Forest
Testing
Training
10-fold Cross validation
Apply the learned model
to the remaining genes and
prediction performance
with other metrics
B
True Positive Rate (Sensitivity)
False Positive Rate (1- Specificity)
0.0
0.2
0.4
0.6
0.8
1.0
mean AUC = 0.88

Figure 3

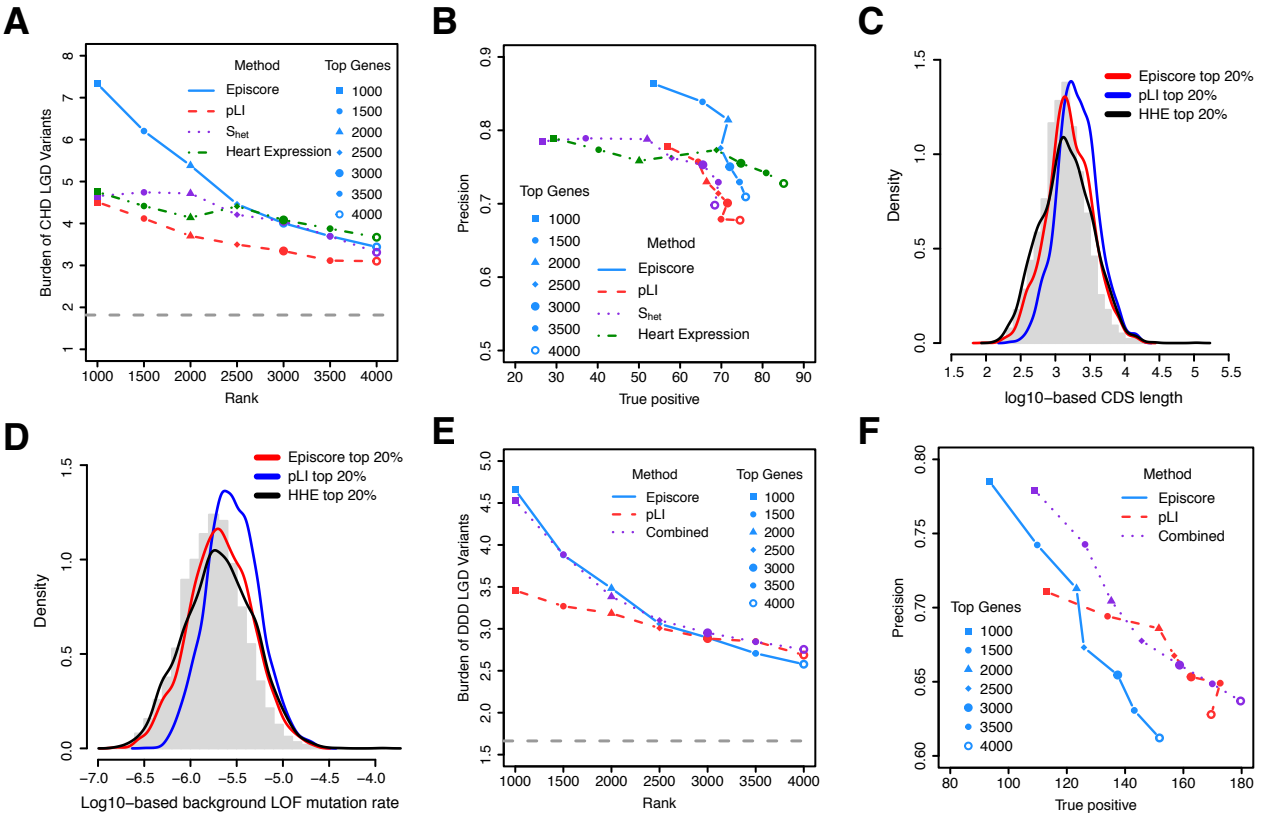
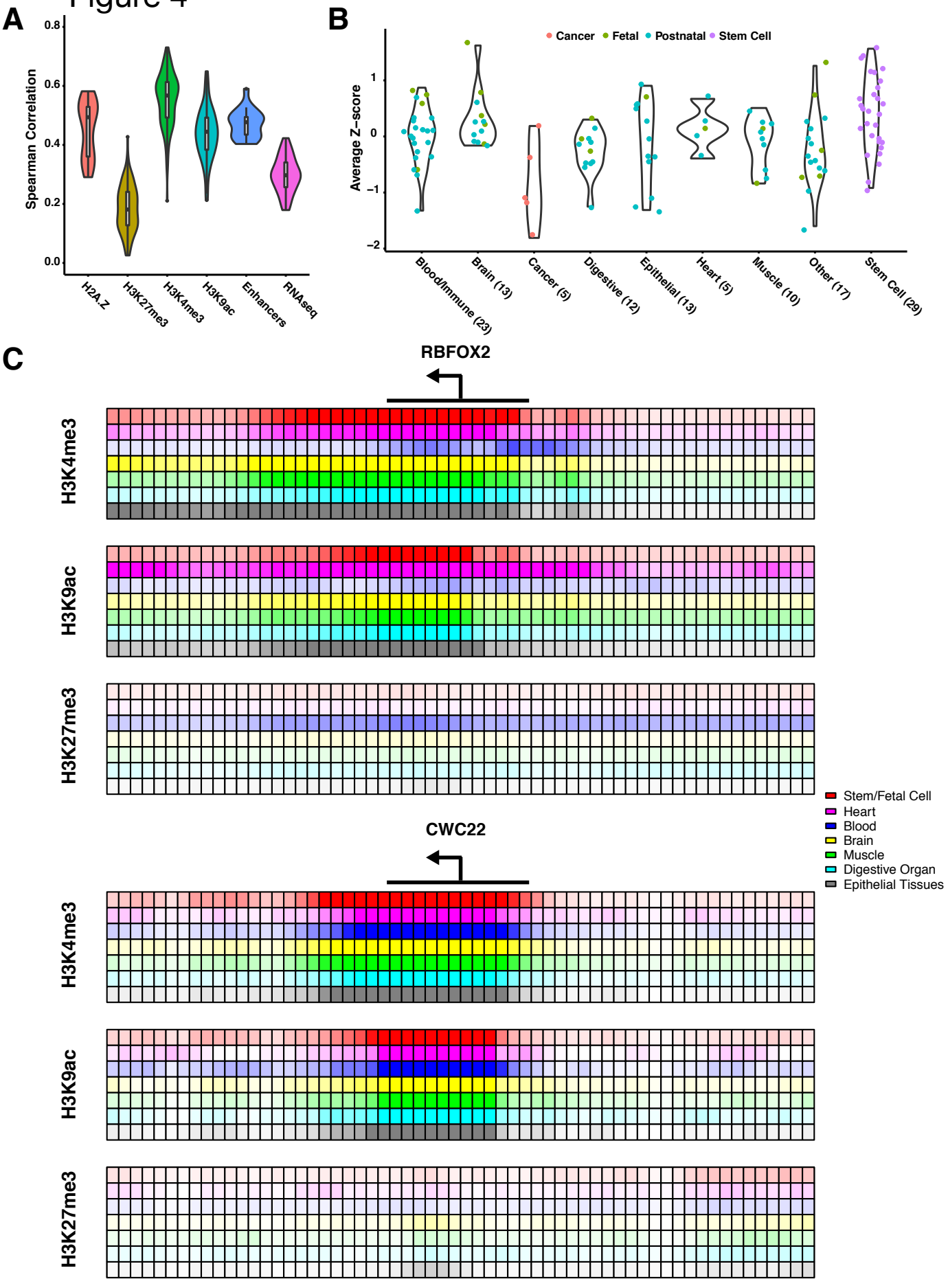
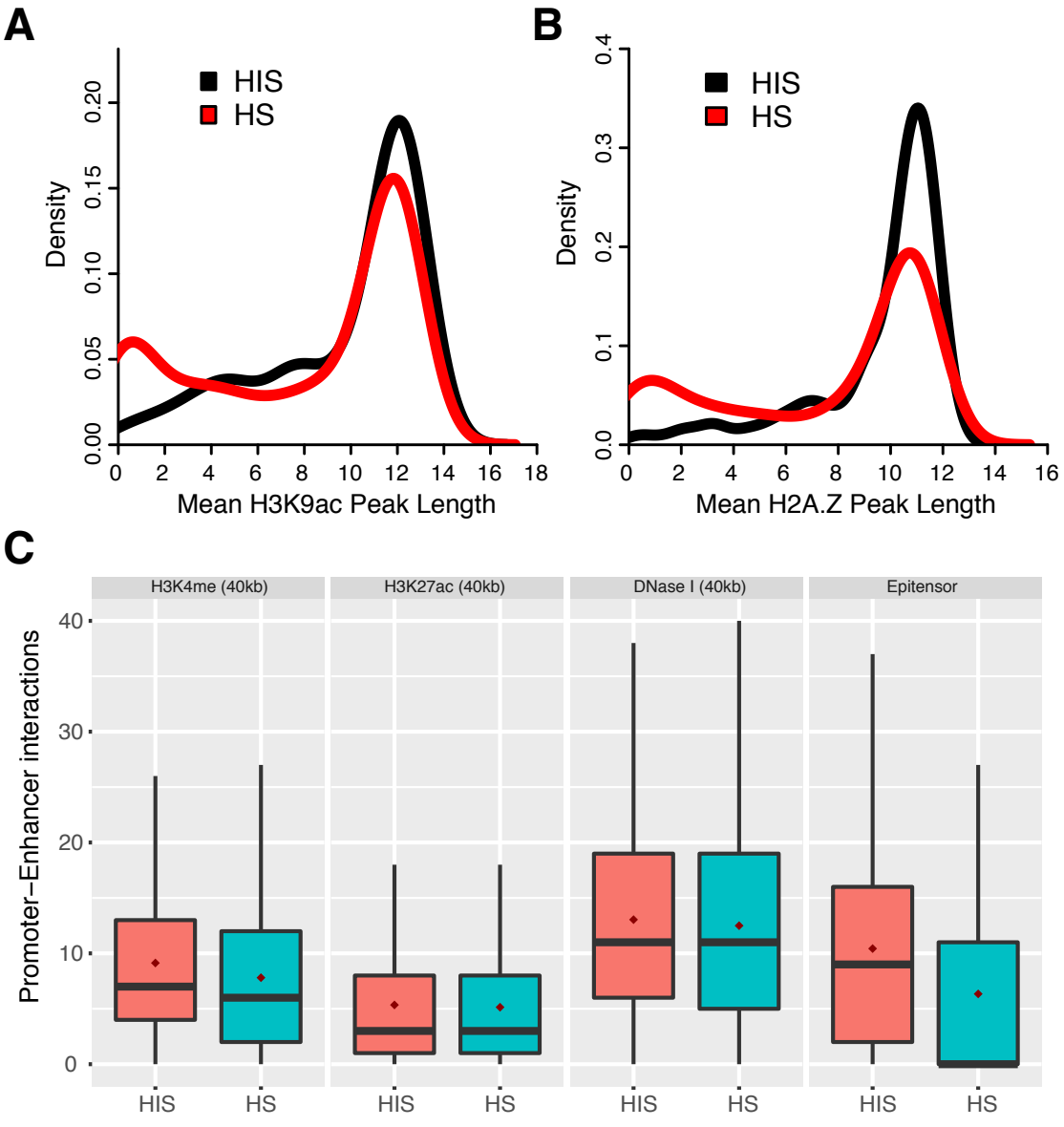

A
Burden of CHD LGD Variants
Rank
Method
Episcore
pLI
S_{het}
Heart Expression
Top Genes
1000
1500
2000
2500
3000
3500
4000
B
Precision
True positive
Top Genes
Method
C
Density
log10−based CDS length
Episcore top 20%
pLI top 20%
HHE top 20%
D
Density
Log10−based background LOF mutation rate
E
Burden of DDD LGD Variants
Rank
Combined
F
Precision
True positive

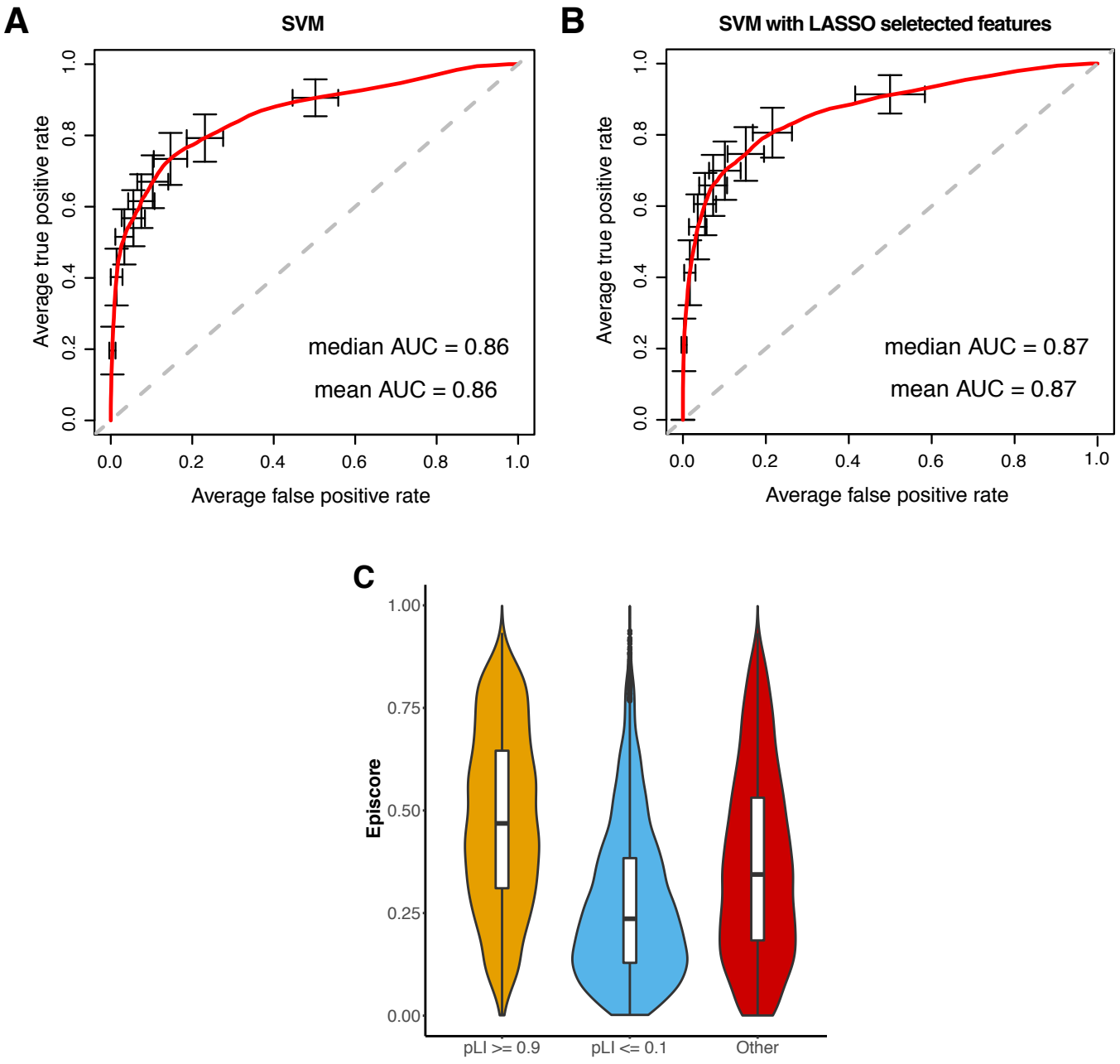
Figure 4


A
Spearman Correlation
0.8
0.6
0.4
0.2
0.0
H2A.Z
H3K27me3
H3K4me3
H3K9ac
Enhancers
RNAseq
B
Cancer
Fetal
Postnatal
Stem Cell
Average Z−score
1
0
−1
−2
Blood/Immune (23)
Brain (13)
Cancer (5)
Digestive (12)
Epithelial (13)
Heart (5)
Muscle (10)
Other (17)
Stem Cell (29)
C
RBFOX2
H3K4me3
H3K9ac
H3K27me3
CWC22
H3K4me3
H3K9ac
H3K27me3
Stem/Fetal Cell
Heart
Blood
Brain
Muscle
Digestive Organ
Epithelial Tissues

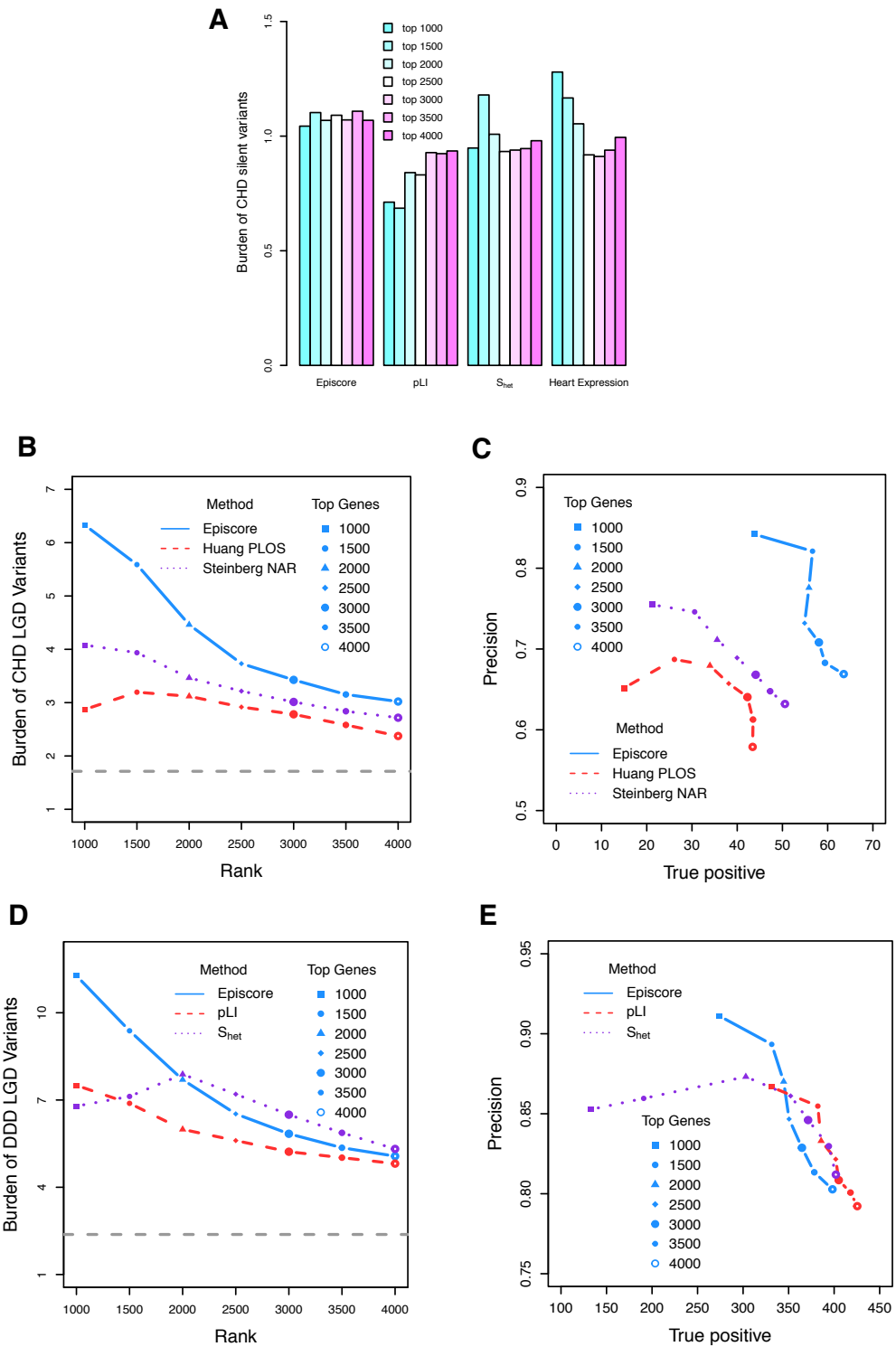
Sup. Figure 1


A
HIS
HS
Density
0.00
0.05
0.10
0.15
0.20
0 2 4 6 8 10 12 14 16 18
Mean H3K9ac Peak Length
B
HIS
HS
Density
0.0
0.1
0.2
0.3
0.4
0 2 4 6 8 10 12 14 16
Mean H2A.Z Peak Length
C
H3K4me (40kb)
H3K27ac (40kb)
DNase I (40kb)
Epitensor
Promoter–Enhancer interactions
0
10
20
30
40
HIS
HS

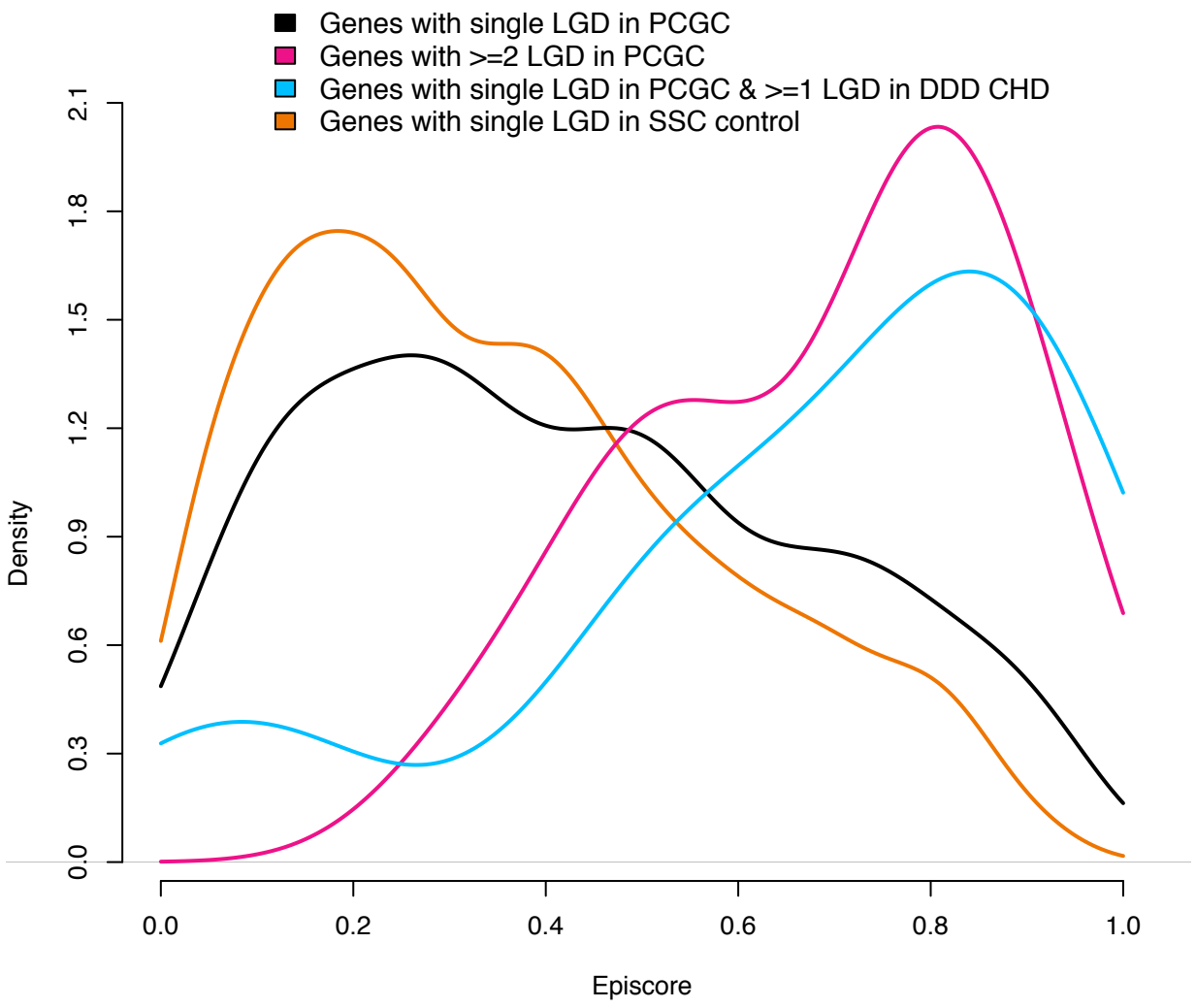
Sup. Figure 2


A
SVM
Average true positive rate
Average false positive rate
median AUC = 0.86
mean AUC = 0.86
B
SVM with LASSO seletected features
Average true positive rate
Average false positive rate
median AUC = 0.87
mean AUC = 0.87
C
Episcore
1.00
0.75
0.50
0.25
0.00
pLI >= 0.9
pLI <= 0.1
Other

Sup. Figure 3


A
Burden of CHD silent variants
top 1000
top 1500
top 2000
top 2500
top 3000
top 3500
top 4000
Episcore
pLI
Shet
Heart Expression
B
Burden of CHD LGD Variants
Method
Episcore
Huang PLOS
Steinberg NAR
Top Genes
Rank
C
Precision
True positive
D
Burden of DDD LGD Variants
E

Sup. Figure 4


Genes with single LGD in PCGC
Genes with >=2 LGD in PCGC
Genes with single LGD in PCGC & >=1 LGD in DDD CHD
Genes with single LGD in SSC control
Density
0.0
0.3
0.6
0.9
1.2
1.5
1.8
2.1
0.0
0.2
0.4
0.6
0.8
1.0
Episcore