Supplemental Materials

## Supplementary Note 1: The Training of the Dirichlet Process Mixture Model

The Dirichlet Process Mixture (DPM) model can automatically determine the meaningful classes. The training process is unsupervised: each variant is represented as a vector, which includes all of the functional annotations, without any label as to its functional status. DPM model learned different patterns from the input feature space based on the iterative algorithm using Variational Bayesian (VB) inference1,2. Although DPM is an infinite mixture model, the VB inference algorithm for the model training still works with a finite approximation. In particular, we used the stick-breaking representation for the Dirichlet Process and the truncation level denoting the upper bound of the cluster number was set to 8. Depending on the input of features, the actual number of components was finally stabilized to five, denoted as clusters C1 to C5, with the corresponding weights for C1~C5 being 0.4758, 0.1986, 0.1787, 0.0906 and 0.0563 (**Figure S7a**). Note that the calculation of the weights of clusters is the accumulation of posterior probabilities of the training data and is slightly different from simply counting the proportions of classes; we still use the proportion of variants a cluster occupies as the size of the cluster for simplicity. The cluster patterns changed very dynamically in the first 20 iterations and started to maintain stable ratios in the last 50 iterations. We also estimated the variational lower bound for each iteration to determine the stop of our algorithm. We found the training process converges very quickly and stopped at the 71st iteration with differential lower bound less than 0.1 (**Figure S7b**).

## Supplementary Note 2: Data used in the training and verification process of DVAR-cluster

The DVAR-cluster step makes use of genomics data across a large set of variants to discover inherent functional patterns. We randomly selected ~2,000,000 variants in the non-coding regions from the 1000 Genomes Project (2015 Aug) to serve as the DPM model’s training set. Non-coding variants were filtered with ANNOVAR UCSC known gene database (version 2016Feb01). None of the chosen variants have been labeled as functional or non-functional. After the model training, it totally identified five functional patterns. We also need to explore each variant cluster with known biological information in order to make sure that the discovered patterns are robust and distinguishable. To avoid getting over-fitted results on the training data, we re-sampled another ~2,000,000 variants using the random strategy with a different initial seed to show the pattern discovery results. Additionally, we used DVAR-score results, chromatin segmentation and gene-based annotation data to verify our clustering results. The chromatin state of each variant was generated by a 15-state ChromHMM model3 with five core marks in the ENCODE GM12878 cell line. Gene-based annotations (with values 'Intergenic', 'Intron', 'Promoter’,'UTR3', 'UTR5', 'Splice') were based on the R package TxDb.Hsapiens.UCSC.hg19.knownGene version 3.2.2 with Bioconductor version 3.2. We also added the annotation value ‘Enhancer’ utilizing the Fantom5 robust enhancers. For all the enrichment analysis, we used one-sided Fisher's exact test and obtain the p-values.

## Supplementary Note 3: Analysis of the clustering of the noncoding variants on the training dataset

In section 2.1, we only focus on the cluster patterns on the ~2 million noncoding variants of the testing dataset which have not been used in the training process of DPM. Relatively, in this section, we focus on the ~2 million noncoding variants in the training dataset. We expect that the five different patterns are robust and distinguishable in all the noncoding regions and the analysis results should be very similar to which in section 2.1. Since we have noticed in section 2.1 that the proportion of each cluster on the testing dataset and training dataset are very close (**Figure 1a**, **Figure S1a**), we also want to know whether the DVAR-score distributions on the training dataset is also similar to that in the testing dataset. The DVAR scores across C1~C5 are clearly different (**Figure S1b**). The sorted fictional average functional scores of the five clusters are: C5 (median = 0.936), C4 (median = 0.878), C3 (median =0.670), C2 (median =0.596) and C1 (median =0.242), suggested that the DVAR score distributions on the training dataset and testing dataset are very similar. We next analyzed the enrichment of training variants with the gene region based marks with heat map plot (**Figure S1c**,

**Table S12**). All of the noncoding variants were marked with 'Intergenic', 'Intron', ‘Enhancer’, 'Promoter’,'UTR3', 'UTR5', 'Splice'. The results showed that background variants in cluster C1 only enriched with Intergenic regions. Most of the functional elements like 3` and 5` UTRs, enhancers and promoters are consistently ordered into clusters C4 and C5. Splicing sites are only enriched in cluster C5 while C4 is associated with Intron regions. The variants in Intergenic regions also enriched in C3 and the variants in Intron regions also enriched in C2. The enrichment analysis suggests that clusters C1~C5 in training dataset shows the same pattern with that in testing dataset. We also analyzed the enrichment of training variants with segmentation annotations. We applied the same ChromHMM4 model across ENCODE GM12878 cell line and use the segmentation results for the enrichment analysis **(Figure S1d**, **Table S13**). The cluster C1 showed enrichment only for the E15 of ChromHMM. The cluster C2 were enriched for ChromHMM states E4, E5, E6, E8, E9, The cluster C3 enriched for ChromHMM states E7, E9, E13, E14, While C4 enriched for ChromHMM states E1~E8 and C5 enriched for ChromHMM states E1~E3, E7, E10~E14. Clusters C1~C5 for the training dataset enriched same ChromHMM states with that for the testing dataset. Finally, we analyzed the tissue specificity of the clusters C1-C5 on the training set. The tissue-specific annotations are used in the same way as mentioned in section 2.1. For the 127 epigenomes /17 tissue types across Encode and Roadmap, most variants in C1, C2, C3 even do not active in any epigenome while variants in C4 and C5 active highly in multiple epigenomes (median=7, 10, **Figure S1e**). In particular, most functional variants in C4 and C5 are active in one particular tissue group while most variants in C1, C2, and C3 are not active in any tissue groups (**Figure S1f**, **Table S14**). In summary, according to these analyses, the patterns discovered by DVAR-cluster are robust and distinguishable both in the training and testing dataset.

## Supplementary Note 4: DVAR score patterns across regulatory elements

Although enrichment analysis of the five clusters is informative (Section 2.1), simply restricting the analysis of functional elements on the cluster level is not enough, as variants in the same cluster are likely to have a different functional impact. Since we have demonstrated the high accuracy of DVAR scores (Section 2.2), we next investigated the patterns of DVAR scores across different REs. We still focused on the 2 million variants randomly selected from the background distribution. For each variant, we generated the gene-based annotation (e.g. UTRs, Promoters, etc.) and the DVAR score. The average functional scores for all types of REs are shown in **Figure 3a**. The DVAR score range is 0 to 1, reflecting the probability of a variant being functional. We sorted the annotated REs according to the average of DVAR scores, with the order of Splice sites (median = 0.978), Enhancers (median = 0.908), 5’ UTRs (median = 0.908), Promoters (median = 0.869), 3’ UTRs (median = 0.830), Intron (median = 0.604) and Intergenic region (median = 0.386). We mainly focused on introns and intergenic regions (IGRs) since they make up the vast majority of non-coding regions in the testing set (40.36% and 54.50%). The DVAR scores of introns are significantly higher than that of IGRs (t-test one-sided p-value<2.22e-308), indicating that intron variants are more likely to establish regularly function than IGR variants. Although the mechanism of how introns modify the gene expression is not yet clear5, there should be fundamental differences between introns and IGRs since the previous clustering analysis shows that they are enriched in distinct functional classes. Consistent with the previous findings6, known functional elements like Splice sites, Enhancers, Promoters, 5’ and 3’ UTRs showed more functional impact than introns and IGRs.

## Supplementary Note 5: Evaluation of the predictive performance on non-coding Indels

In addition to non-coding SNPs, Indels also represent an important type of variant that may likely to have more severe functional alterations. However, due to their complexity and mysticism, identifying non-coding functional Indels remains a bottleneck7. As an unsupervised learning approach, DVAR does not need prior knowledge of Indels, providing us with an opportunity to identify functional Indels. To handle non-coding Indels in our framework, we consider SNPs and Indels together in the model assumption. The 2,000,000 training samples randomly selected from the 1000 Genomes Project (1KGP, 2015 Aug, ALL sites) also include Indels. The background model learned from the training data has the ability to handle not only SNPs but also Indels. To evaluate the performance of our approach, we also need to construct proper testing datasets. For the original four functional SNPs testing dataset, Only Clinvar and GTEx eQTL datasets have enough functional Indels, and we, therefore, used these two as the testing datasets for Indels. For the Clinvar dataset, we collected 8,455 non-coding Indels from the Clinvar pathogenic variants (2017.1.30) for the positive standards and randomly sampled 8,455 non-coding Indels in the 1KGP with MAF >0.05 for the negative standards. Similarly, we constructed the eQTL testing set with 98 GTEx eQTL Indels and 90 negative samples from 1KGP. We show AUCs of PR curves and ROC curves to evaluate the performances of DVAR on these two sets (**Figure S8**). For the ClinVar indels dataset, the DVAR score performed exceptionally well (AUCPR=0.967, AUCROC=0.967), only slightly worse than that for ClinVar SNPs (AUCPR=0.981, AUCROC=0.980). For the eQTL indels dataset, the DVAR score also has a good prediction power (AUCPR=0.758, AUCROC=0.765), even a little better than that for eQTL SNPs (AUCPR=0.724, AUCROC=0.749).

## Supplementary Note 6: DVAR scores across COSMIC somatic mutations

We evaluate how well DVAR can predict functional somatic mutations which are associated with human cancer. We focused on whether DVAR could distinguish recurrent somatic mutations from the non-recurrent somatic mutations in COSMIC database8. We collected 801 recurrent mutations from the COSMIC dataset used in the GWAVA paper9 as the positive instances. We also randomly selected 801 non-recurrent somatic mutations from the same dataset as the negative instances (for both sets, we removed the mutations in chromosome M). Since there are many indels in this dataset, we only compare the performance of DVAR with GWAVA. The GWAVA scores of the testing set are download from the GWAVA FTP site: <ftp://ftp.sanger.ac.uk/pub/resources/software/gwava/>. We use AUCROC to indicate the performance of DVAR (AUCroc=0.664) and GWAVA (AUCroc=0.667) on the COSMIC dataset. The results show that DVAR is competitive compared with the GWAVA (**Figure S9**). The performance of GWAVA on this dataset is consistent with the previous study9. On the COSMIC somatic mutations, compared to the two-class method (GWAVA), we found that DVAR does not show obvious advantages. This may be due to two reasons: 1) there are many passenger mutations which are not pathogenic but with positive labels in the testing set, which may decrease the prediction power. 2) DVAR is focused on the discovery of the functional patterns of germline mutations, which may be distinct from the patterns of somatic mutations. Pattern discovery on the somatic mutations with the integration of cancer-related annotation data should be considered in our future work.

## Supplementary Note 7: Explore further improvements in the performance of DVAR

We are interested in whether the optimization of the feature extraction module and scoring module of DVAR framework can continue to improve the performance. The first challenge we encountered during feature processing was the missing annotations of variants. We used the mean value imputation for each dimension of data in 2.2, which may ignore the correlations of annotations. In order to make better use of the other dimensions of data to impute the missing annotation, we used Multivariate Imputation by Chained Equations10 (MICE) method for all the variants in the default testing datasets. We choose the predictive mean matching as imputation method from the mice package. The iteration number of the algorithm is set to 40. We denote the scores of DVAR with MICE feature imputation as MICE-DVAR. The second attempt we made to improve the performance was feature transformation. DPM model used by DVAR is based on the assumption that the input features follow the infinite mixture of multivariate Gaussian distribution. However, the features we used ranged from 0 to 1, which may deviate from the model assumption. To address this problem, we used the logit function to ensure that the feature values fit the model assumption. We converted all the features of the variants in the training and test sets. The DPM model of DVAR was also re-trained. We denote the scores of DVAR with feature transformation as LOGIT-DVAR. The last attempt we made was to change the scoring algorithm. For the DVAR-score method, we used the cluster labels to determine the scoring weights of variants. This scoring method can be extended to support the input from clustering labels generated from other frameworks. For the DPM model that we used in the clustering process, we can replace the weights of the scoring method with the posterior probabilities for each variant of all the clusters to further improve the performance. We denote the scores generated by this scoring method as PW-DVAR. We used Clinvar, fine-mapped GWAS, GTEX eQTLs, and MPRA datasets to evaluate the prediction accuracy of MICE-DVAR, LOGIT-DVAR and PW-DVAR. The results show that all the scores are comparable with the original DVAR score on the Clinvar dataset (**Figure S10**). PW-DVAR (AUCPR=0.982) even performs slightly better than the original DVAR score (AUCPR=0.981). On the other datasets, the original DVAR method always maintains more stable performance than other methods (PW-DVAR is competitive compared with DVAR on the fine-mapped GWAS dataset). In summary, the original DVAR framework shows stable performance on various test sets. The optimization of the feature extraction method and the scoring method will continue in the future.

## Supplementary Note 8: DVAR scores across allele frequencies

We evaluated the functional consequences of variants with different MAFs and in particular, we were interested in whether and to what extent rare alleles are more likely to be functional than common ones. We pre-computed the functional scores of 2 million variants randomly select from 1KGP and separated them into four subgroups according to the different MAF bins ((0, 0.001), [0.001, 0.01), [0.01, 0.05), (0.05, 0.5)). We denote the subgroup with MAF > 0.05 as the common allele group while the subgroup with MAF < 0.001 as the rare allele group. Contrasting the rare vs. the common allele groups, it is clear that the rare alleles have significantly higher functional scores than the common alleles (one-sided P-value = 1.33e-104, Wilcoxon signed rank test). Evaluating alleles across these four groups, we found that the average scores increase with decreasing allele frequencies (**Figure S11**). We also evaluated variants in Intergenic, Intron and Enhancer regions separately, and in all of these three regions rare alleles have significantly higher functional prediction scores than common alleles (Wilcoxon signed rank test, one-sided P-values are: Intergenic = 0.012, Intron = 4.45e-64, Enhancer = 6.60e-4) with similarly weak differences across MAF bins.

Reference

1. Bishop, C.M. Pattern recognition. *Machine Learning* **128**(2006).

2. Blei, D.M. & Jordan, M.I. Variational inference for Dirichlet process mixtures. *Bayesian analysis* **1**, 121-143 (2006).

3. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* **9**, 215-6 (2012).

4. Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-30 (2015).

5. Chorev, M. & Carmel, L. The function of introns. *Front Genet* **3**, 55 (2012).

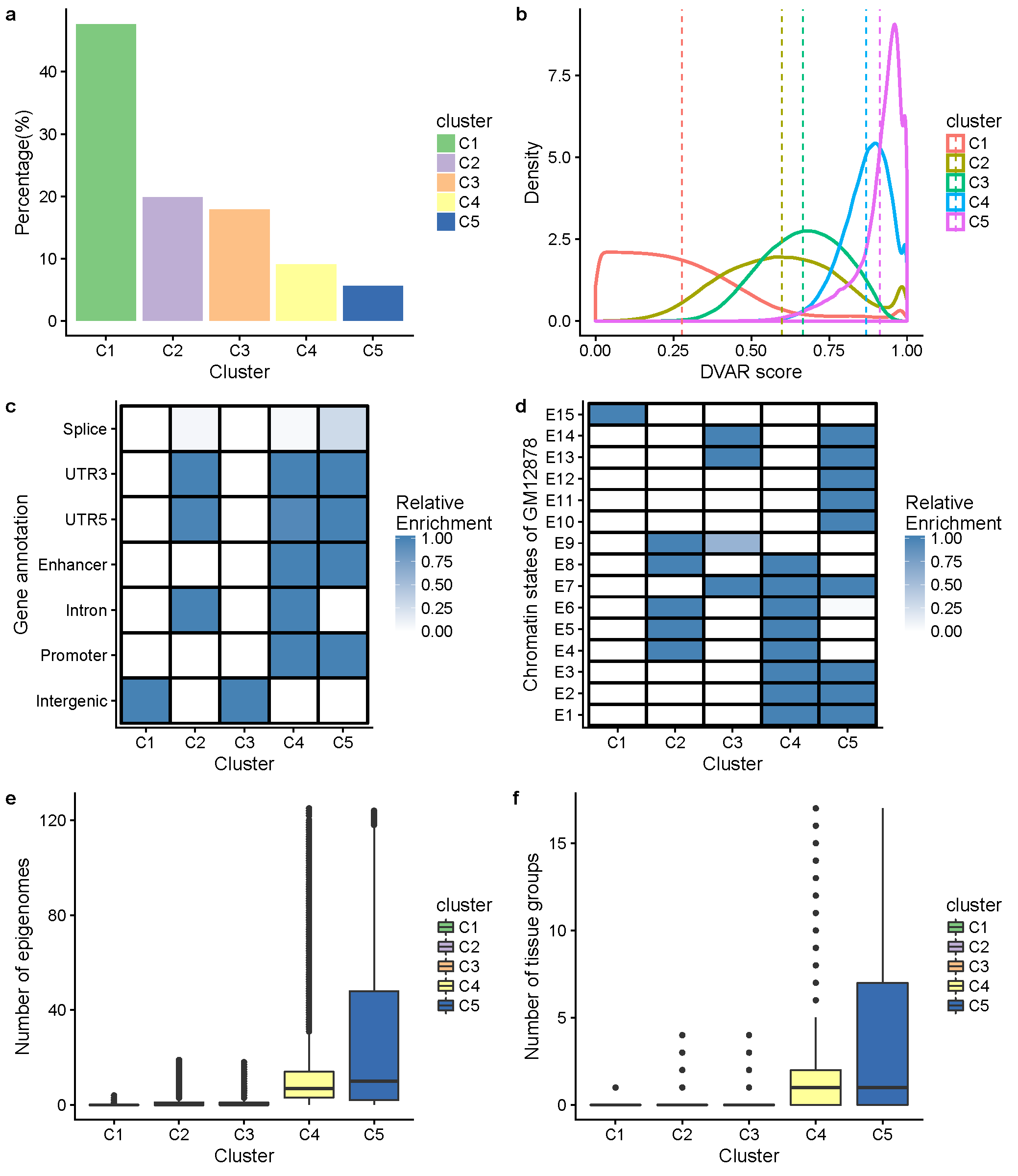
6. Huang, Y.F., Gulko, B. & Siepel, A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat Genet* (2017).

7. Ionita-Laza, I., McCallum, K., Xu, B. & Buxbaum, J.D. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat Genet* **48**, 214-20 (2016).

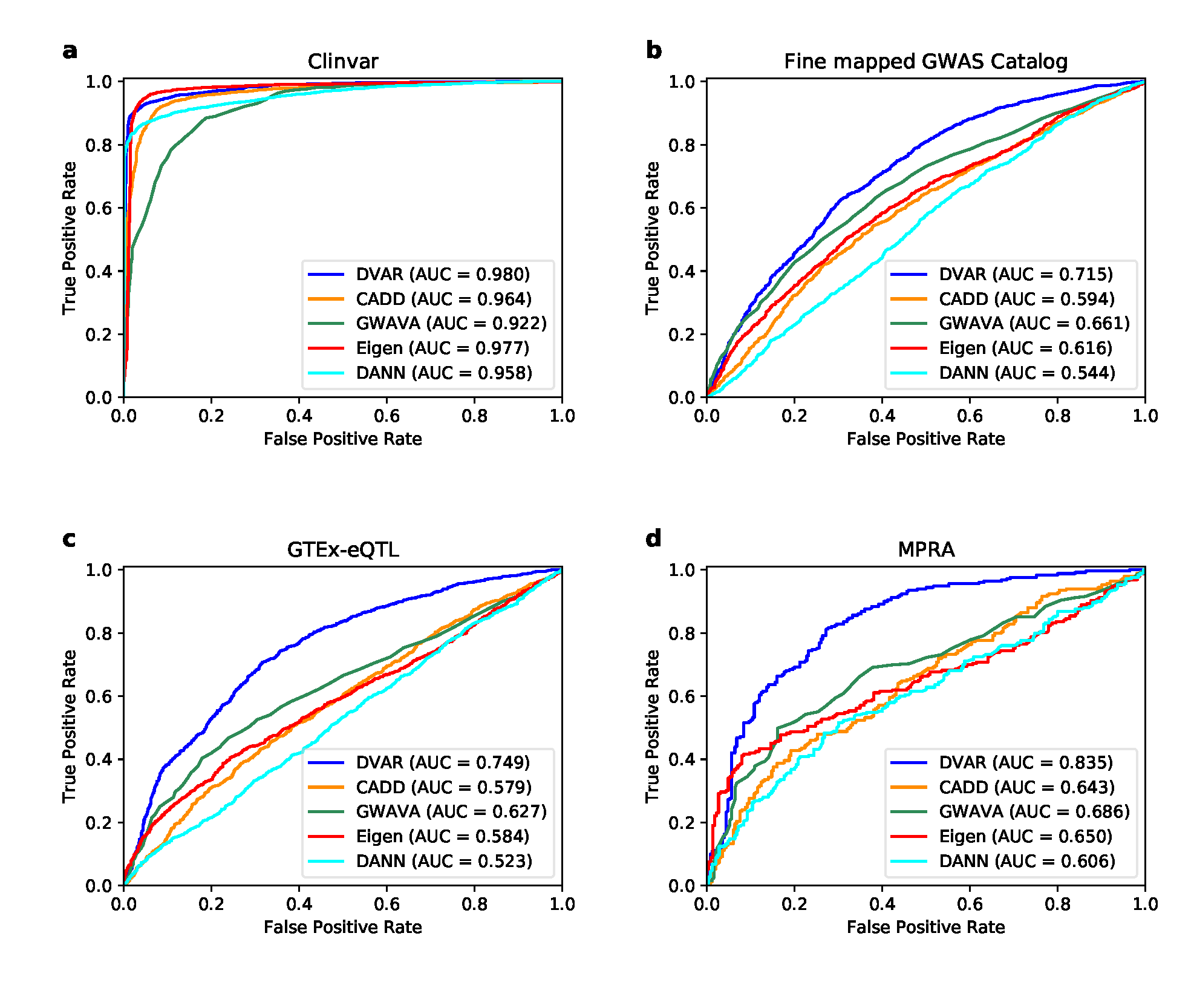
8. Forbes, S.A. *et al.* COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res* **45**, D777-D783 (2017).

9. Ritchie, G.R., Dunham, I., Zeggini, E. & Flicek, P. Functional annotation of noncoding sequence variants. *Nat Methods* **11**, 294-6 (2014).

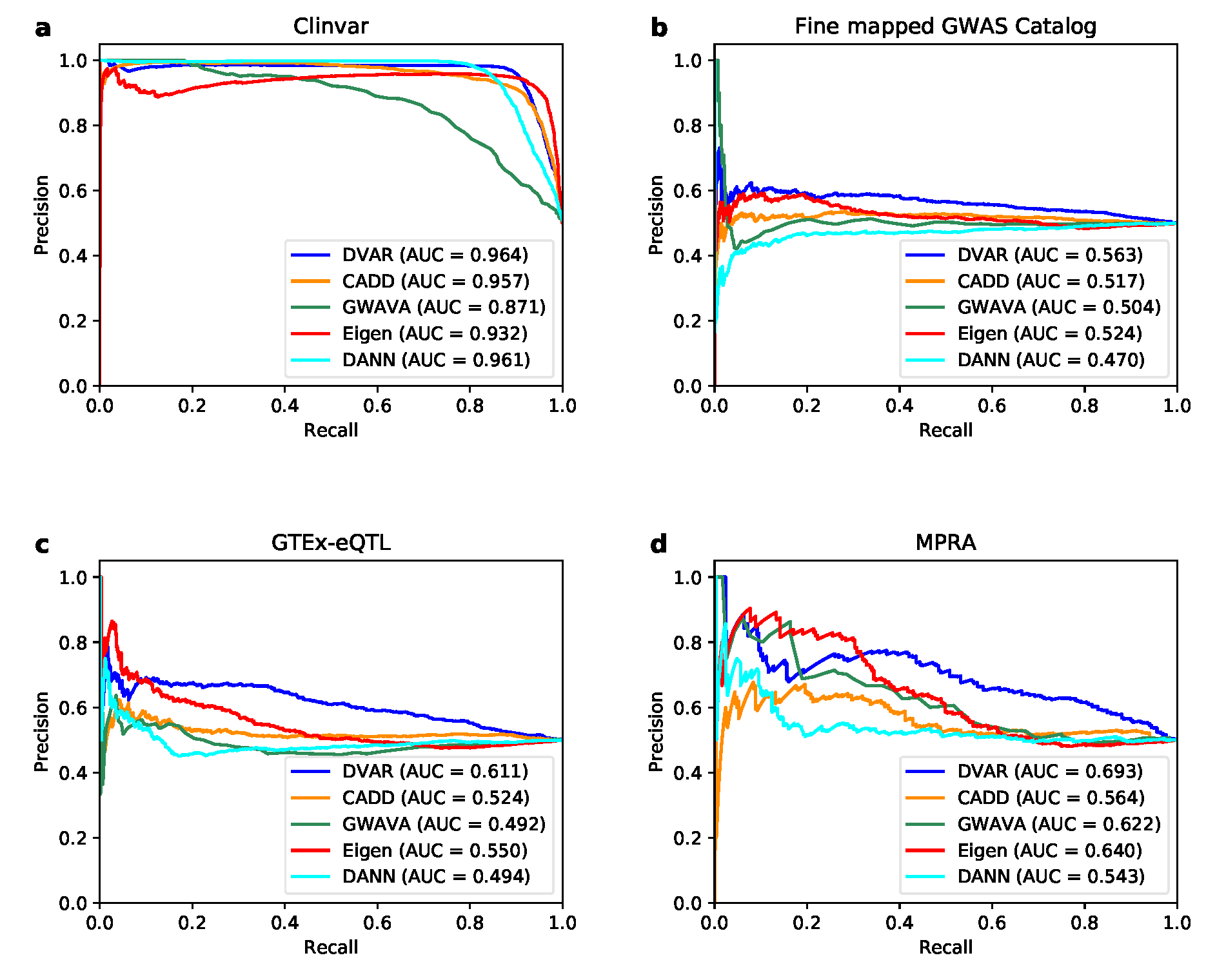
10. Buuren, S.v. & Groothuis-Oudshoorn, K. mice: Multivariate imputation by chained equations in R. *Journal of statistical software*, 1-68 (2010).



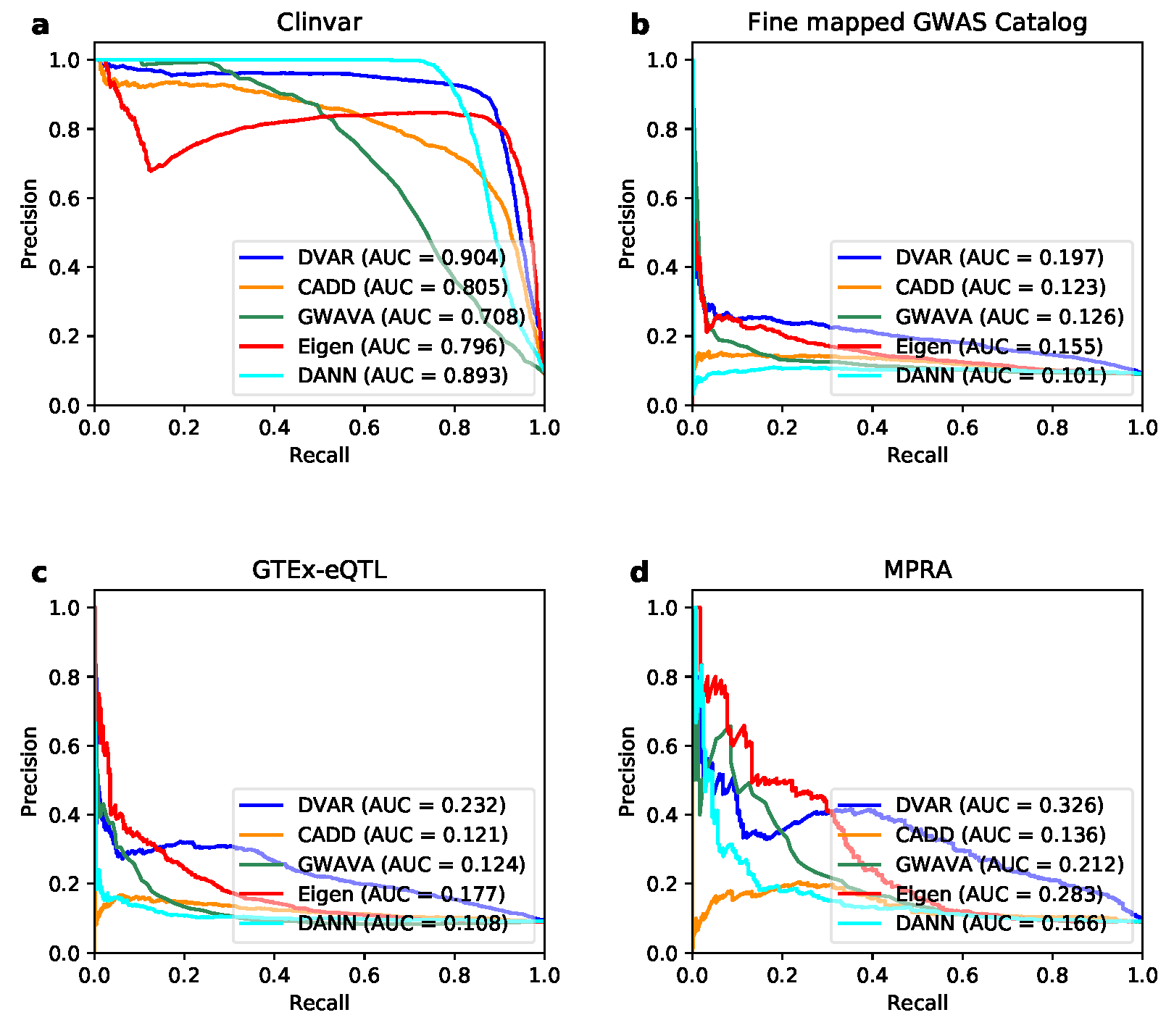
**Figure S1** Exploration of different patterns of DVAR clusters C1~C5 on the **training** **data**. **a**) The proportion of variants in clusters C1~C5: 47.61%, 19.87%, 17.85%, 9.05% and 5.59%. **b**) The DVAR score distribution of C1~C5. The median DVAR scores of C1-C5 are 0.242, 0.596, 0.670, 0.878 and 0.936. **c**) The heat map shows the relative enrichment of DVAR clusters with gene-based annotation regions. **d**) The heat map shows the relative enrichment of DVAR clusters with the 15-state ChromHMM model. **e**) The average number of activated epigenomes of DVAR clusters C1~C5. **f**) The average number of activated tissue groups of DVAR clusters C1~C5.



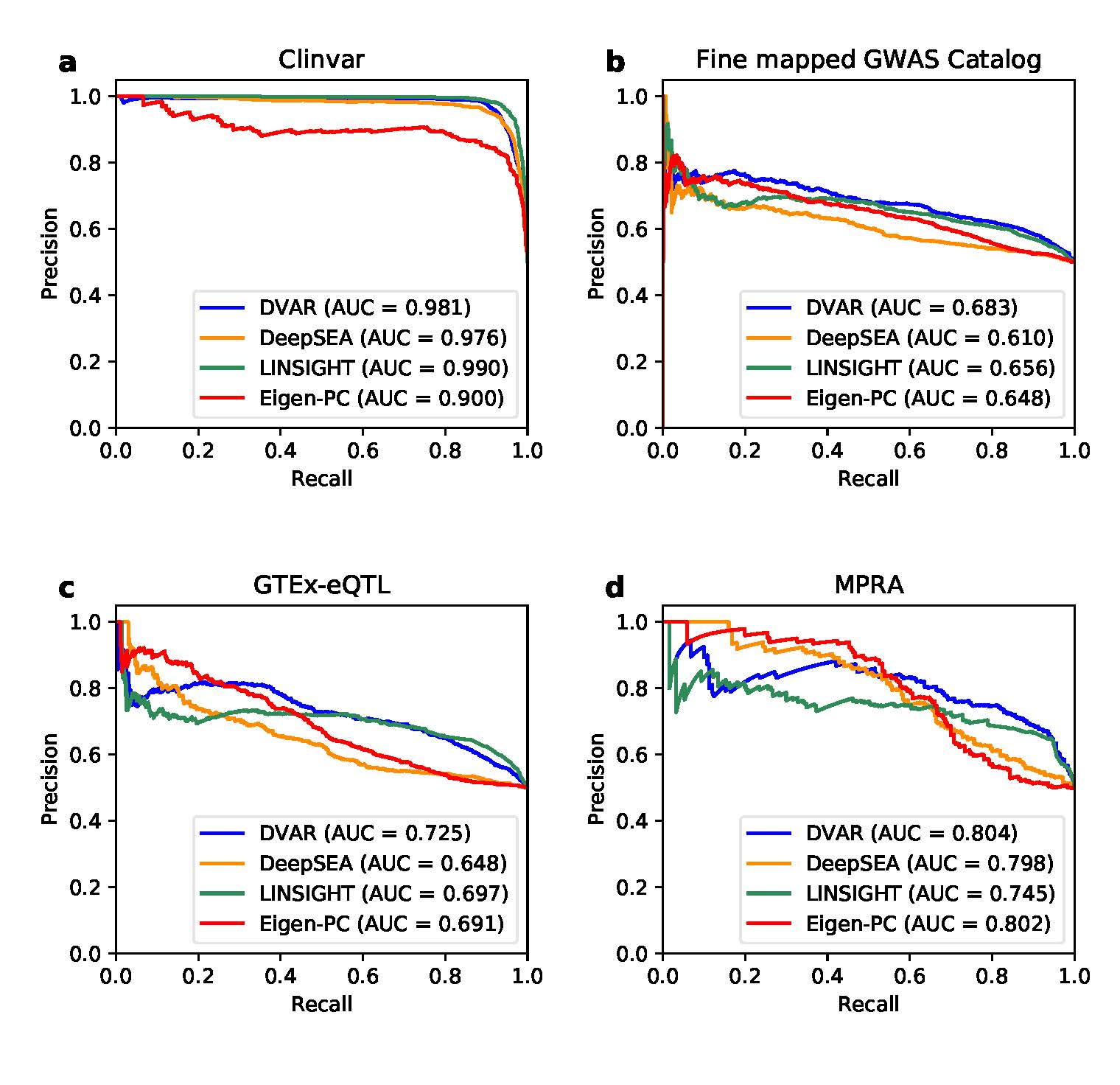
**Figure S2** Performance comparison of DVAR, CADD, GWAVA, Eigen, and DANN on **a**) Clinvar database, **b**) fine-mapped GWAS variants, **c**) GTEx-eQTL variants, and **d**) MPRA validated variants. The prediction power is evaluated based on the area under the receiver operating characteristic (ROC) curves.



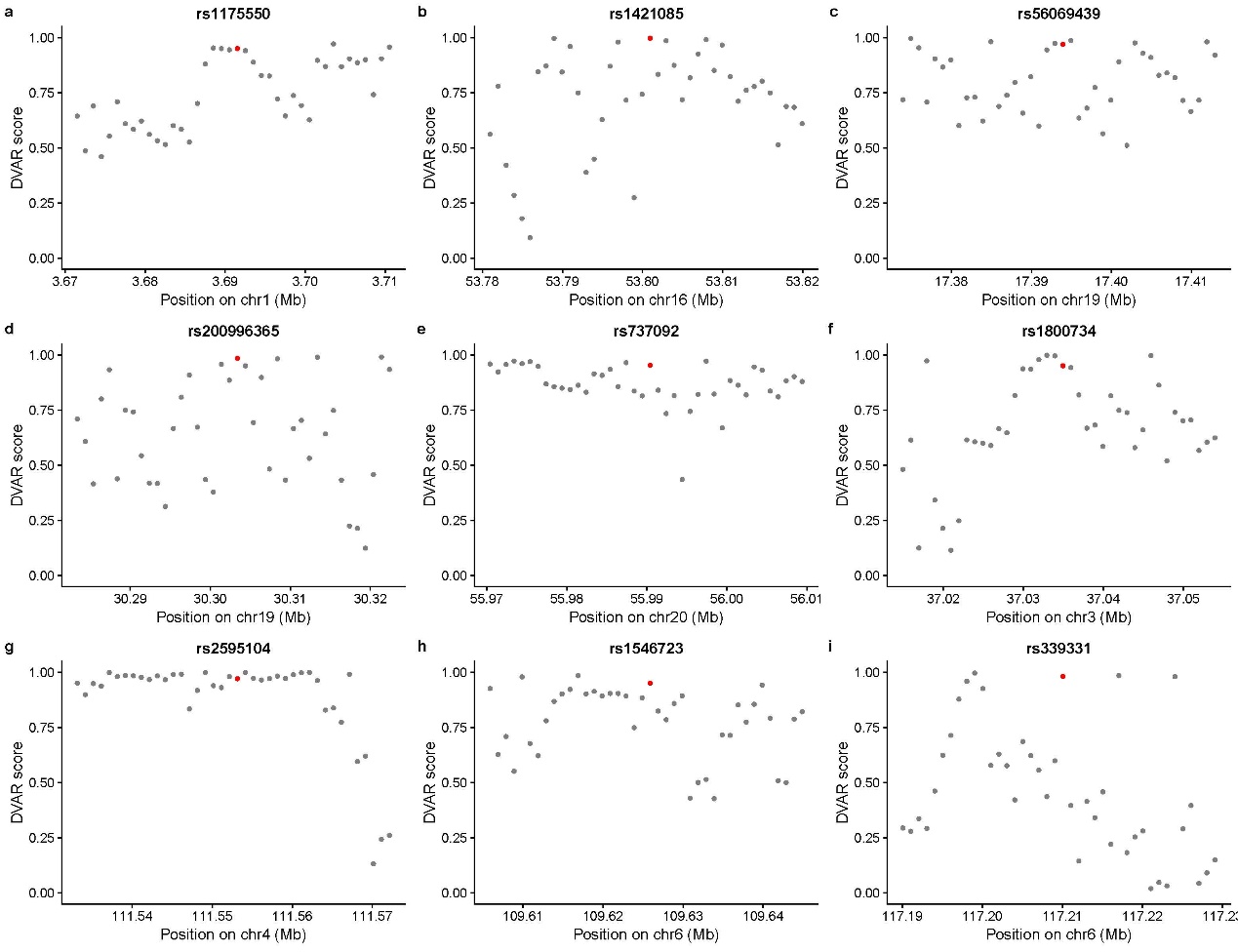
**Figure S3** Performance comparison of DVAR, CADD, GWAVA, Eigen, and DANN on **a**) Clinvar database, **b**) fine-mapped GWAS variants, **c**) GTEx-eQTL variants, and **d**) MPRA validated variants. The negative dataset is selected from 1000 Genomes based on the region-matched criterion. The prediction power is evaluated based on the area under the Precision-Recall curve.



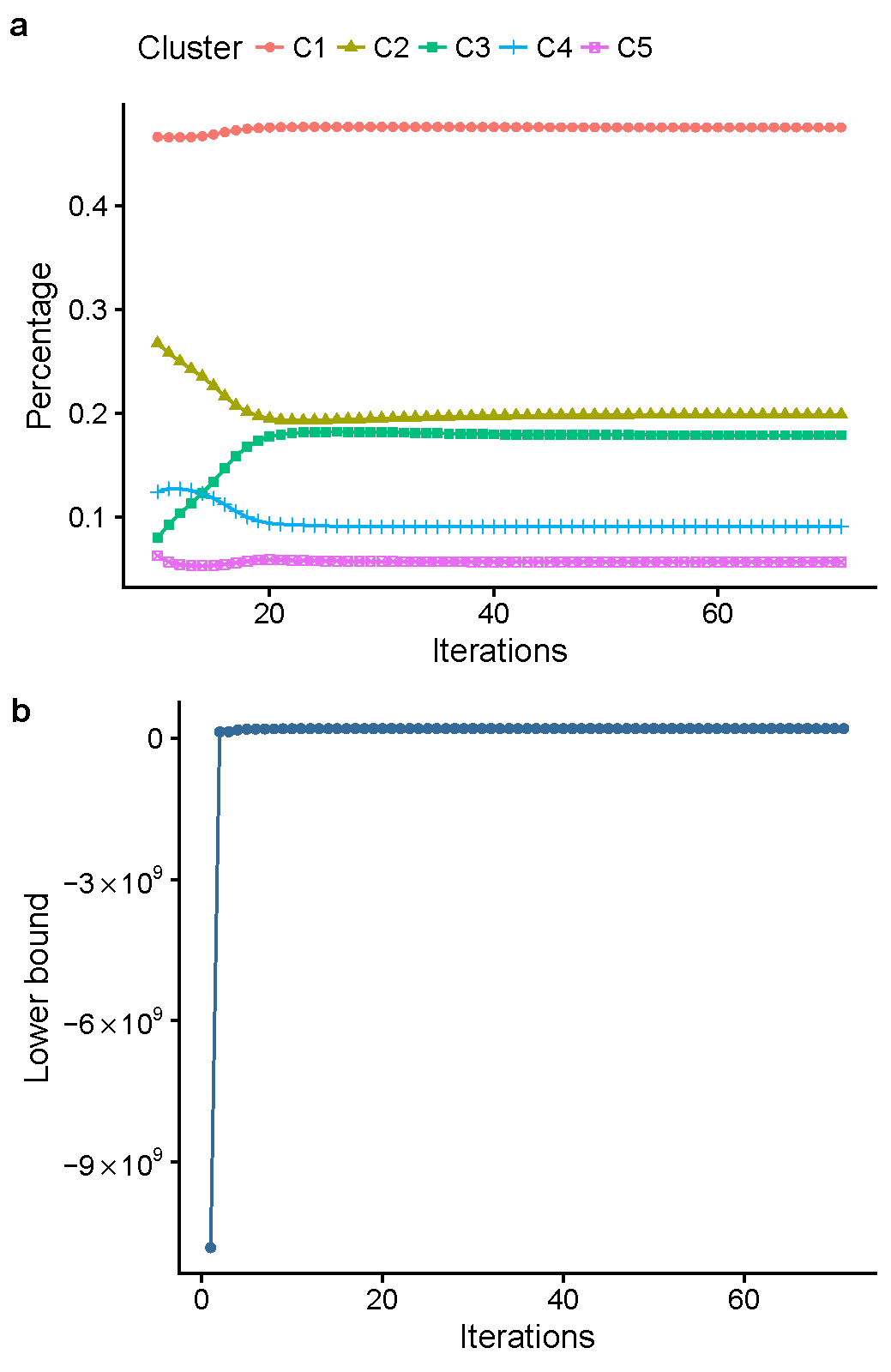
**Figure S4** Performance comparison of DVAR, CADD, GWAVA, Eigen, and DANN on **a**) Clinvar database, **b**) fine-mapped GWAS variants, **c**) GTEx-eQTL variants, and **d**) MPRA validated variants. The negative dataset is selected from 1000 Genomes based on the 1:10-imbalance criterion. The prediction power is evaluated based on the area under the Precision-Recall curve.



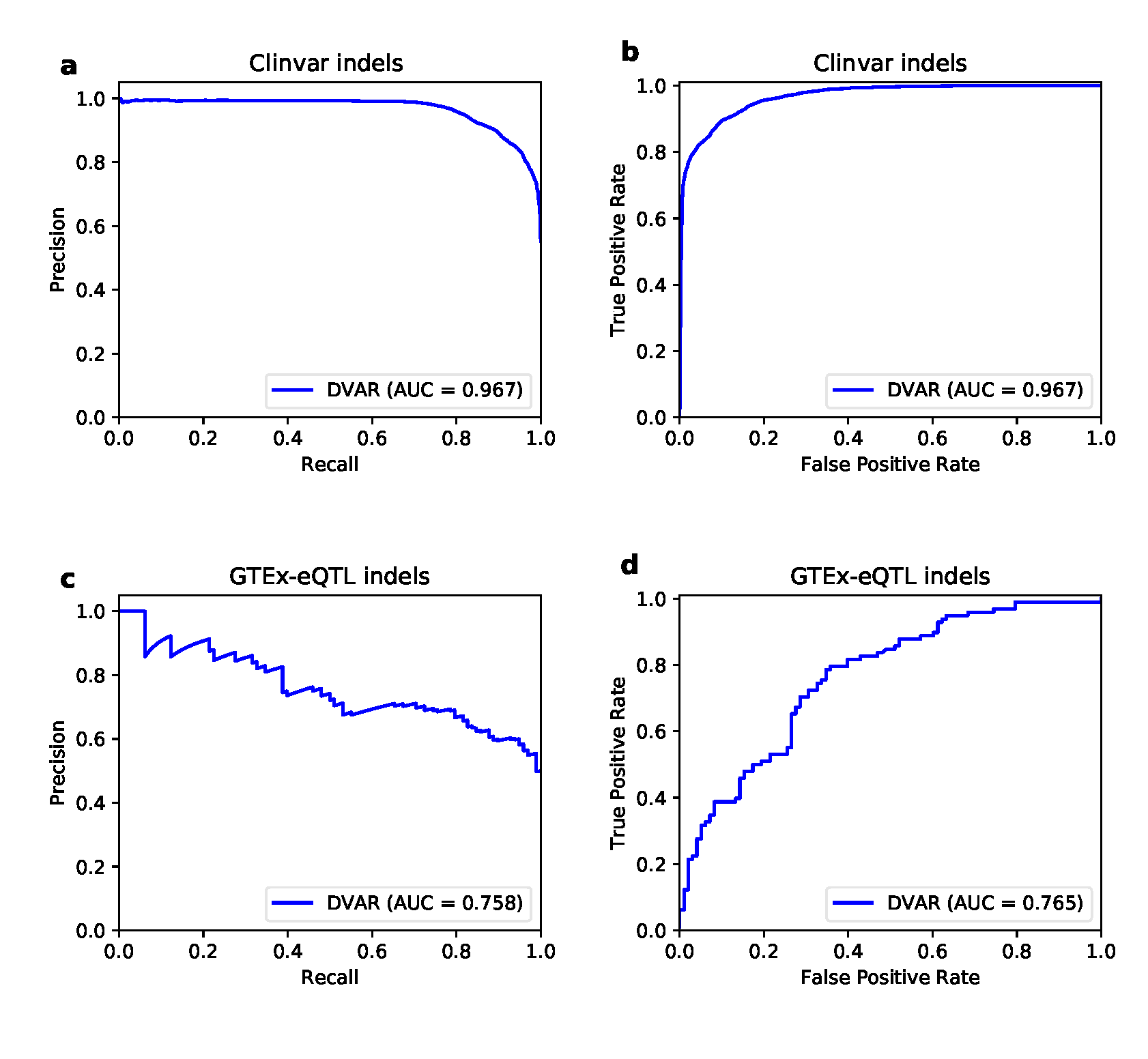
**Figure S5** Performance comparison of DVAR, DeepSEA, LINSIGHT, and Eigen-PC on **a**) Clinvar database, **b**) fine-mapped GWAS variants, **c**) GTEx-eQTL variants, and **d**) MPRA validated variants. The negative dataset is selected from 1000 Genomes based on the region-matched criterion. The prediction power is evaluated based on the area under the Precision-Recall curve.



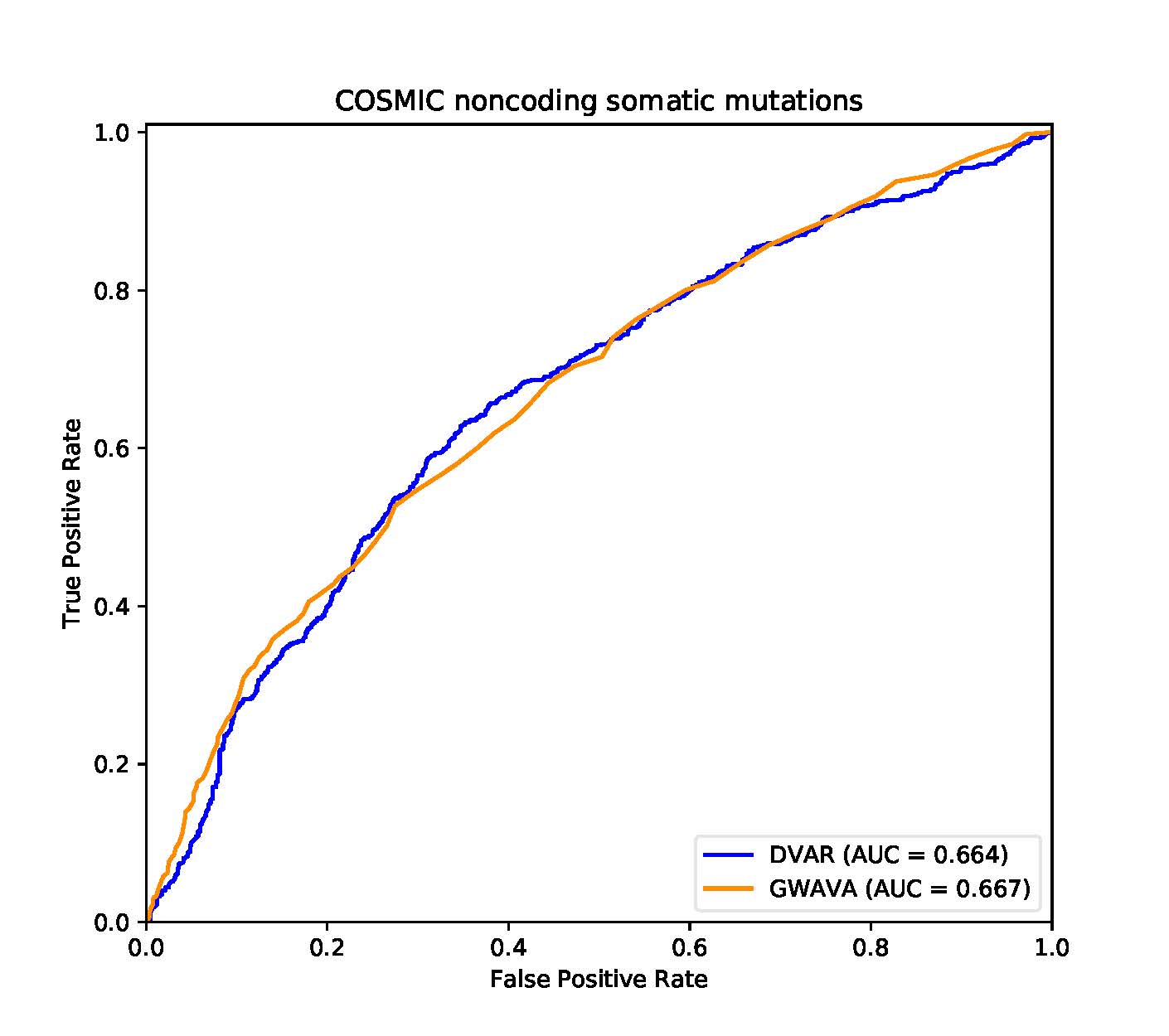
**Figure S6** Scores of positions across 40kb regions surrounding the genome-editing validated variants (the red dots denote the validated variants (rs1175550, rs1421085, rs56069439, rs200996365, rs737092, rs1800734, rs2595104, rs1546723, rs339331) and the grey dots denote the positons corresponding to the validated variants).



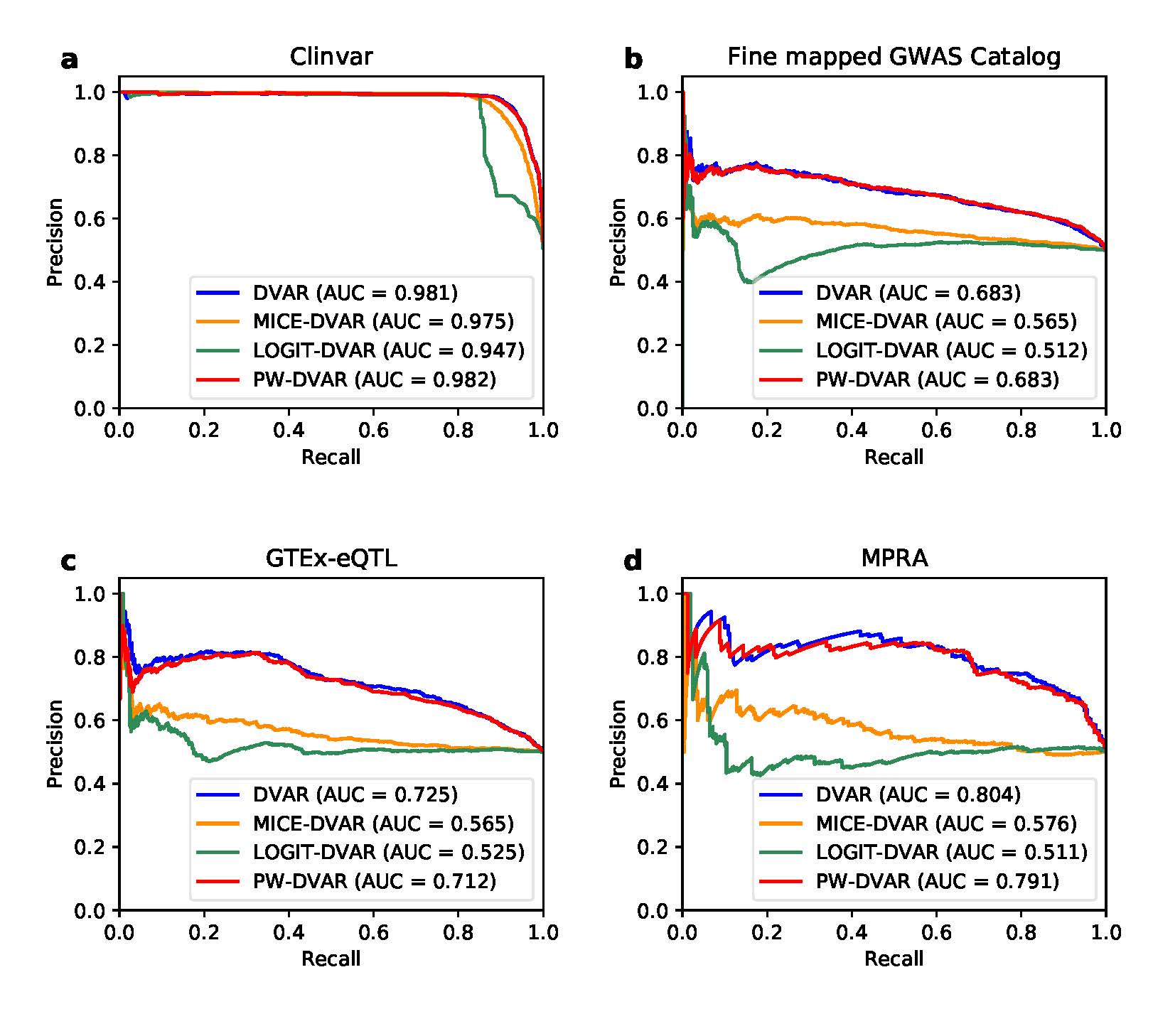
**Figure S7** A summary of the training process of the Dirichlet Process Mixture Model. **a**) The change of the proportion of variants in each cluster during the training process. **b**) The change of the lower bound during the training process.



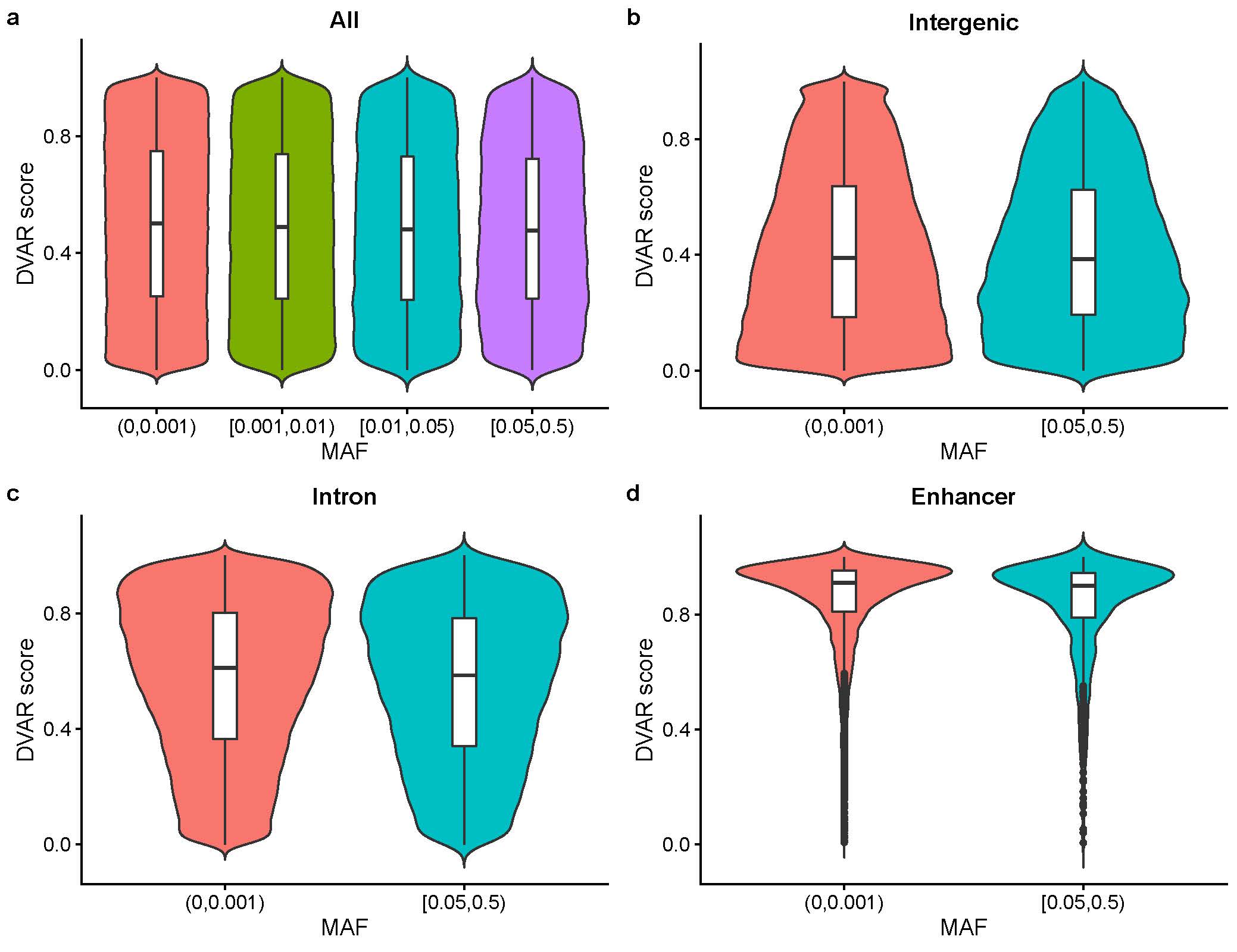
**Figure S8** Prediction performance of DVAR on Clinvar disease variants and GTEx-eQTL variants. We focused on the statistics of the area under precision-recall curves (**a**, **c**) and the area under the ROC curves (**b**, **d**).



**Figure S9** ROC curves of DVAR and GWAVA for the performance comparison on the cosmic dataset.



**Figure S10** Performance comparison of DVAR, MICE-DVAR, LOGIT-DVAR, and PW-DVAR on **a**) Clinvar database, **b**) fine-mapped GWAS variants, **c**) GTEx-eQTL variants, and **d**) MPRA validated variants. The prediction power is evaluated based on the area under the Precision-Recall curve.



**Figure S11** Distributions of VAR scores of the rare allele group (MAF < 0.001) and common allele group (MAF >0.05) for a) All the non-coding regions, b) Intergenic regions, c) Intron regions and d) Enhancer regions.

**Table S1** Description of the parameters and variables of DVAR-cluster and DVAR-score.

|  |  |
| --- | --- |
|  | Description |
| Variables: | Random sample distribution drawn from a Dirichlet process  Prior distribution (we use Normal-Inverse-Wishart distribution here)  The total number of variants in the observed dataset  The dimension of the annotation feature  Cluster number scaling random variable distribution  For annotation feature of variant  For cluster label of variant  For random variables of the mixture k  For random variables of the mixture k |
| Parameters: | For k=1,2,…∞ weight of the mixture k  Parameters of Beta distribution  Parameters of Gamma distribution  Parameters of distribution of the mixture k  Parameters of distribution of the mixture k |
| Other:  DP() | Dirichlet process  Weight factors  Between-class functional score  Within-class functional score  Total functional score  Probability score of variants |

**Table S2** P-values (Wilcoxon signed rank test, one side with ‘less’) for the analysis of DVAR Scoring of different DVAR clusters on the testing dataset.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | C1 | C2 | C3 | C4 | C5 |
| C1 | - | **<2.22E-308** | **<2.22E-308** | **<2.22E-308** | **<2.22E-308** |
| C2 | - | - | **<2.22E-308** | **<2.22E-308** | **<2.22E-308** |
| C3 | - | - | - | **<2.22E-308** | **<2.22E-308** |
| C4 | - | - | - | - | **<2.22E-308** |
| C5 | - | - | - | - | - |

**Table S3** P-values (Fisher's exact test) for the Enrichment analysis of DVAR Clusters and gene-based annotations on the testing set.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | C1 | C2 | C3 | C4 | C5 |
| Intergenic | **<2.22E-308** | 1 | **<2.22E-308** | 1 | 1 |
| Intron | 1 | **<2.22E-308** | 1 | **<2.22E-308** | 9.87E-1 |
| Enhancer | 1 | 1 | 1 | **<2.22E-308** | **<2.22E-308** |
| Promoter | 1 | 1 | 1 | **<2.22E-308** | **<2.22E-308** |
| UTR3 | 1 | **<2.22E-308** | 1 | **<2.22E-308** | **6.35E-168** |
| UTR5 | 1 | **1.34E-69** | 1 | **3.79E-78** | **<2.22E-308** |
| Splice | 9.99E-1 | **4.56E-4** | 9.95E-1 | **5.23E-3** | **1.42E-15** |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | C1 | C2 | C3 | C4 | C5 |
| E1 (Active TSS) | 1 | 1 | 1 | **<2.22E-308** | **<2.22E-308** |
| E2 (Flanking active TSS) | 1 | 1 | 1 | **<2.22E-308** | **<2.22E-308** |
| E3 (Transcr. at gene 5′ and 3′) | 1 | 1 | 1 | **<2.22E-308** | **4.54E-108** |
| E4 (Strong transcription) | 1 | **<2.22E-308** | 1 | **<2.22E-308** | 1 |
| E5 (Weak transcription) | 1 | **<2.22E-308** | 1 | **<2.22E-308** | 1 |
| E6 (Genic enhancers) | 1 | **2.99E-212** | 1 | **<2.22E-308** | **1.44E-2** |
| E7 (Enhancers) | 1 | 1 | **4.36E-211** | **<2.22E-308** | **<2.22E-308** |
| E8 (ZNF genes + repeats) | 1 | **<2.22E-308** | 1 | **1.11E-100** | 7.88E-1 |
| E9 (Heterochromatin) | 1 | **<2.22E-308** | **2.20E-46** | 1 | 1 |
| E10 (Bivalent/poised TSS) | 1 | 1 | 1 | 9.99E-1 | **<2.22E-308** |
| E11 (Flanking bivalent TSS/Enh) | 1 | 1 | 1 | 9.99E-1 | **<2.22E-308** |
| E12 (Bivalent enhancer) | 1 | 1 | 9.96E-1 | 9.99E-1 | **<2.22E-308** |
| E13 (Repressed Polycomb) | 1 | 1 | **4.57E-97** | 1 | **<2.22E-308** |
| E14 (Weak repressed Polycomb) | 1 | 1 | **<2.22E-308** | 1 | **<2.22E-308** |
| E15 (Quiescent/low) | **<2.22E-308** | 1 | 1 | 1 | 1 |

**Table S4** P-values (Fisher's exact test) for the Enrichment analysis of DVAR Clusters and Chromatin annotations on the testing dataset.

|  |  |  |
| --- | --- | --- |
|  | # activity tissue groups | # activity epigenomes |
| C1 | 0 | 0 |
| C2 | 0 | 0 |
| C3 | 0 | 0 |
| C4 | 1 | 7 |
| C5 | 1 | 10 |

**Table S5** Median of the total number of activity tissue groups/epigenomes of each cluster on the testing set.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | DVAR | CADD | GWAVA | Eigen | DANN |
| rs1175550 | 0.9509 | -0.1637 | 0.59 | 1.2927 | 0.500 |
| rs1421085 | 0.9976 | 1.7631 | 0.41 | 1.7886 | 0.866 |
| rs56069439 | 0.9702 | 0.4429 | 0.34 | 0.8704 | 0.667 |
| rs737092 | 0.9536 | 0.0408 | 0.54 | 0.6011 | 0.624 |
| rs1800734 | 0.9512 | 0.6369 | 0.91 | 1.7268 | 0.768 |
| rs2595104 | 0.9711 | 0.8215 | 0.35 | 0.3383 | 0.359 |
| rs1546723 | 0.9517 | 0.2961 | 0.53 | 0.6059 | 0.533 |
| rs339331 | 0.9820 | 2.3910 | 0.28 | 1.2820 | 0.892 |

**Table S6** Scores of functional variants (rs1175550, rs1421085, rs56069439, rs737092, rs1800734, rs2595104, rs1546723, rs339331) evaluated by genome editing. (DVAR, CADD, GWAVA, Eigen, and DANN)

**Table S7** P values (Wilcoxon signed rank test, one-sided with ‘greater’) for the analysis of DVAR high/low score groups with EHR-based phenotype association counts on the BioVU and UK biobank datasets.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| PheWAS  P value  Cutoff | BioVU | | | UK Biobank | | |
| MAF >0 | MAF >0.05 | MAF >0.1 | MAF >0 | MAF >0.05 | MAF >0.1 |
| 1E-3 | **5.68E-7** | **3.50E-7** | **3.76E-7** | **<2.22E-308** | **<2.22E-308** | **<2.22E-308** |
| 1E-4 | **2.58E-20** | **4.06E-32** | **2.33E-21** | **<2.22E-308** | **<2.22E-308** | **<2.22E-308** |
| 1E-5 | **9.27E-20** | **9.44E-42** | **2.05E-29** | **<2.22E-308** | **<2.22E-308** | **<2.22E-308** |
| 1E-6 | **6.74E-23** | **1.81E-36** | **9.52E-33** | **<2.22E-308** | **<2.22E-308** | **<2.22E-308** |

**Table S8** P values (Wilcoxon signed rank test, one-sided with ‘greater’) for the analysis of GWAVA high/low score groups with EHR-based phenotype association counts on the BioVU and UK biobank datasets.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| PheWAS  P value  Cutoff | BioVU | | | UK Biobank | | |
| MAF >0 | MAF >0.05 | MAF >0.1 | MAF >0 | MAF >0.05 | MAF >0.1 |
| 1E-3 | 1 | 1 | 9.81e-1 | **4.031-3** | 1 | 1 |
| 1E-4 | **3.70E-9** | **2.46E-14** | **9.18E-15** | **3.15E-41** | 8.52E-1 | 9.96E-1 |
| 1E-5 | **7.83E-22** | **2.91E-46** | **1.17E-44** | **4.07E-22** | **5.73E-6** | **1.73E-2** |
| 1E-6 | **3.21E-16** | **1.19E-27** | **5.44E-27** | **2.50E-5** | **1.73E-2** | 2.24E-1 |

**Table S9** P values (Wilcoxon signed rank test, one-sided with ‘greater’) for the analysis of CADD high/low score groups with EHR-based phenotype association counts on the BioVU and UK biobank datasets.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| PheWAS  P value  Cutoff | BioVU | | | UK Biobank | | |
| MAF >0 | MAF >0.05 | MAF >0.1 | MAF >0 | MAF >0.05 | MAF >0.1 |
| 1E-3 | 7.50E-1 | 1 | 9.97E-1 | **<2.22E-308** | **2.35E-100** | **<9.42E-89** |
| 1E-4 | 1 | 1 | 1 | **2.21E-301** | **9.80E-9** | **1.05E-13** |
| 1E-5 | 1 | 1 | 1 | **2.01E-14** | 9.99E-1 | 9.91E-1 |
| 1E-6 | 1 | 1 | 1 | 1 | 1 | 1 |

**Table S10** P values (Wilcoxon signed rank test, one-sided with ‘greater’) for the analysis of Eigen high/low score groups with EHR-based phenotype association counts on the BioVU and UK biobank datasets.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| PheWAS  P value  Cutoff | BioVU | | | UK Biobank | | |
| MAF >0 | MAF >0.05 | MAF >0.1 | MAF >0 | MAF >0.05 | MAF >0.1 |
| 1E-3 | **6.27E-4** | 9.66E-1 | 8.91E-1 | **<2.22E-308** | **<2.22E-308** | **<2.22E-308** |
| 1E-4 | 1 | 1 | 1 | **<2.22E-308** | **5.91E-86** | **2.15E-110** |
| 1E-5 | 1 | 1 | 1 | **5.43E-57** | 3.77E-1 | **2.40E-5** |
| 1E-6 | 1 | 1 | 1 | 1 | 1 | 1 |

**Table S11** P values (Wilcoxon signed rank test, one-sided with ‘greater’) for the analysis of DANN high/low score groups with EHR-based phenotype association counts on the BioVU and UK biobank datasets.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| PheWAS  P value  Cutoff | BioVU | | | UK Biobank | | |
| MAF >0 | MAF >0.05 | MAF >0.1 | MAF >0 | MAF >0.05 | MAF >0.1 |
| 1E-3 | **7.06E-4** | 9.97E-1 | 9.98E-1 | **<2.22E-308** | **4.02E-15** | **8.59E-7** |
| 1E-4 | 8.37E-1 | 1 | 1 | **<2.22E-308** | 9.95E-1 | 8.24E-1 |
| 1E-5 | 1 | 1 | 1 | **4.79E-7** | 1 | 9.99E-1 |
| 1E-6 | 1 | 1 | 1 | 1 | 1 | 9.99E-1 |

**Table S12** P-values (Fisher's exact test) for the Enrichment analysis of DVAR Clusters and gene-based annotations on the training set.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | C1 | C2 | C3 | C4 | C5 |
| Intergenic | **<2.22E-308** | 1 | **<2.22E-308** | 1 | 1 |
| Intron | 1 | **<2.22E-308** | 1 | **<2.22E-308** | 8.96E-1 |
| Enhancer | 1 | 1 | 1 | **<2.22E-308** | **<2.22E-308** |
| Promoter | 1 | 1 | 1 | **<2.22E-308** | **<2.22E-308** |
| UTR3 | 1 | **<2.22E-308** | 1 | **<2.22E-308** | **9.35E-175** |
| UTR5 | 1 | **9.91E-64** | 1 | **3.64E-61** | **<2.22E-308** |
| Splice | 1 | **6.67E-5** | 9.99E-1 | **8.20E-4** | **5.64E-19** |

**Table S13** P-values (Fisher's exact test) for the Enrichment analysis of DVAR Clusters and

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | C1 | C2 | C3 | C4 | C5 |
| E1 (Active TSS) | 1 | 1 | 1 | **<2.22E-308** | **<2.22E-308** |
| E2 (Flanking active TSS) | 1 | 1 | 1 | **<2.22E-308** | **<2.22E-308** |
| E3 (Transcr. at gene 5′ and 3′) | 1 | 1 | 1 | **<2.22E-308** | **6.16E-104** |
| E4 (Strong transcription) | 1 | **<2.22E-308** | 1 | **<2.22E-308** | 1 |
| E5 (Weak transcription) | 1 | **<2.22E-308** | 1 | **<2.22E-308** | 1 |
| E6 (Genic enhancers) | 1 | **1.10E-259** | 1 | **<2.22E-308** | **3.55E-3** |
| E7 (Enhancers) | 1 | 1 | **1.19E-229** | **<2.22E-308** | **<2.22E-308** |
| E8 (ZNF genes + repeats) | 1 | **<2.22E-308** | 1 | **1.66E-128** | 9.99E-1 |
| E9 (Heterochromatin) | 1 | **<2.22E-308** | **1.78E-38** | 1 | 1 |
| E10 (Bivalent/poised TSS) | 1 | 1 | 1 | 1 | **<2.22E-308** |
| E11 (Flanking bivalent TSS/Enh) | 1 | 1 | 1 | 9.99E-1 | **<2.22E-308** |
| E12 (Bivalent enhancer) | 1 | 1 | 8.92E-1 | 8.86E-1 | **<2.22E-308** |
| E13 (Repressed Polycomb) | 1 | 1 | **6.31E-109** | 1 | **<2.22E-308** |
| E14 (Weak repressed Polycomb) | 1 | 1 | **<2.22E-308** | 1 | **<2.22E-308** |
| E15 (Quiescent/low) | **<2.22E-308** | 1 | 1 | 1 | 1 |

Chromatin annotations on the training set.

**Table S14** Median of the total number of activity tissue groups/epigenomes of each cluster on the training set.

|  |  |  |
| --- | --- | --- |
|  | # activity tissue groups | # activity epigenomes |
| C1 | 0 | 0 |
| C2 | 0 | 0 |
| C3 | 0 | 0 |
| C4 | 1 | 7 |
| C5 | 1 | 10 |