

# Quantifying the regulatory potential of genetic variants via a hybrid sequence-oriented model with SVEN

Received: 10 May 2024

Accepted: 6 December 2024

Published online: 30 December 2024

 Check for updatesYu Wang<sup>1,2</sup>, Nan Liang<sup>1</sup> & Ge Gao<sup>1,2</sup> 

Deciphering how noncoding DNA determines gene expression is critical for decoding the functional genome. Understanding the transcription effects of noncoding genetic variants are still major unsolved problems, which is critical for downstream applications in human genetics and precision medicine. Here, we integrate regulatory-specific neural networks and tissue-specific gradient-boosting trees to build SVEN: a hybrid sequence-oriented architecture that can accurately predict tissue-specific gene expression level and quantify the tissue-specific transcriptomic impacts of structural variants across more than 350 tissues and cell lines. We further systematically screen a large-scale structural variants dataset derived from 3622 individuals and clinical structural variants from ClinVar, and provide an overview of transcriptomic impacts of structural variants in population. As a sequence-oriented model, SVEN is also able to predict regulatory effects for small noncoding variants. We expect that SVEN will enable more effective in silico analysis and interpretation of human genome-wide disease-related genetic variants.

Whole-genome sequencing enables the generation of high-resolution maps of genomic variations in the human genome<sup>1–5</sup>, revealing pervasive structural variants (SVs) in the human genome<sup>4,6,7</sup>. Operationally defined as large-scale genomic alterations (>50 bp), SVs have been shown to have prominent impacts on several complex diseases, including schizophrenia, rheumatoid arthritis, and type 1 and type 2 diabetes<sup>8–10</sup>. Typically, SVs are thought to function by influencing gene expression via effects on the regulatory regions of genes<sup>10–12</sup>. Determining the transcriptomic impact of SVs genome-wide remains a serious challenge.

Multiple attempts have been made to systematically characterize the cellular impacts of genetic variants. In addition to classical annotation-oriented approaches that based on existing functional annotations<sup>13–15</sup>, multiple sequence-oriented methods have been proposed<sup>16–18</sup>. By attempting to “learn and model” regulatory codes from DNA sequences directly via various deep learning networks, sequence-oriented methods have demonstrated notable performance in predicting the influence of genetic variants on gene expression in both well-annotated and poorly annotated genomic regions<sup>19–21</sup>. However, these

sequence-oriented methods are mainly developed for single nucleotide variants (SNVs) and small indels rather than large-scale SVs.

Here we present SVEN, a multi-modality sequence-oriented in silico model, for quantifying both small and large-scale genetic variants’ regulatory impacts in over 350 tissues and cell lines. In addition to its superior performance for tissue-specific gene expression prediction (mean Spearman  $R = 0.892$ ), SVEN was found to be highly accurate for assessing the impact of SVs on gene expression when being applied to a large-scale SV dataset (Spearman  $R = 0.921$ ), enabling to systematically characterize the effects of common and pathogenic SVs detected in large-scale population on gene expression in various tissues and cell lines. SVEN is available at <https://github.com/gao-lab/SVEN>.

## Results

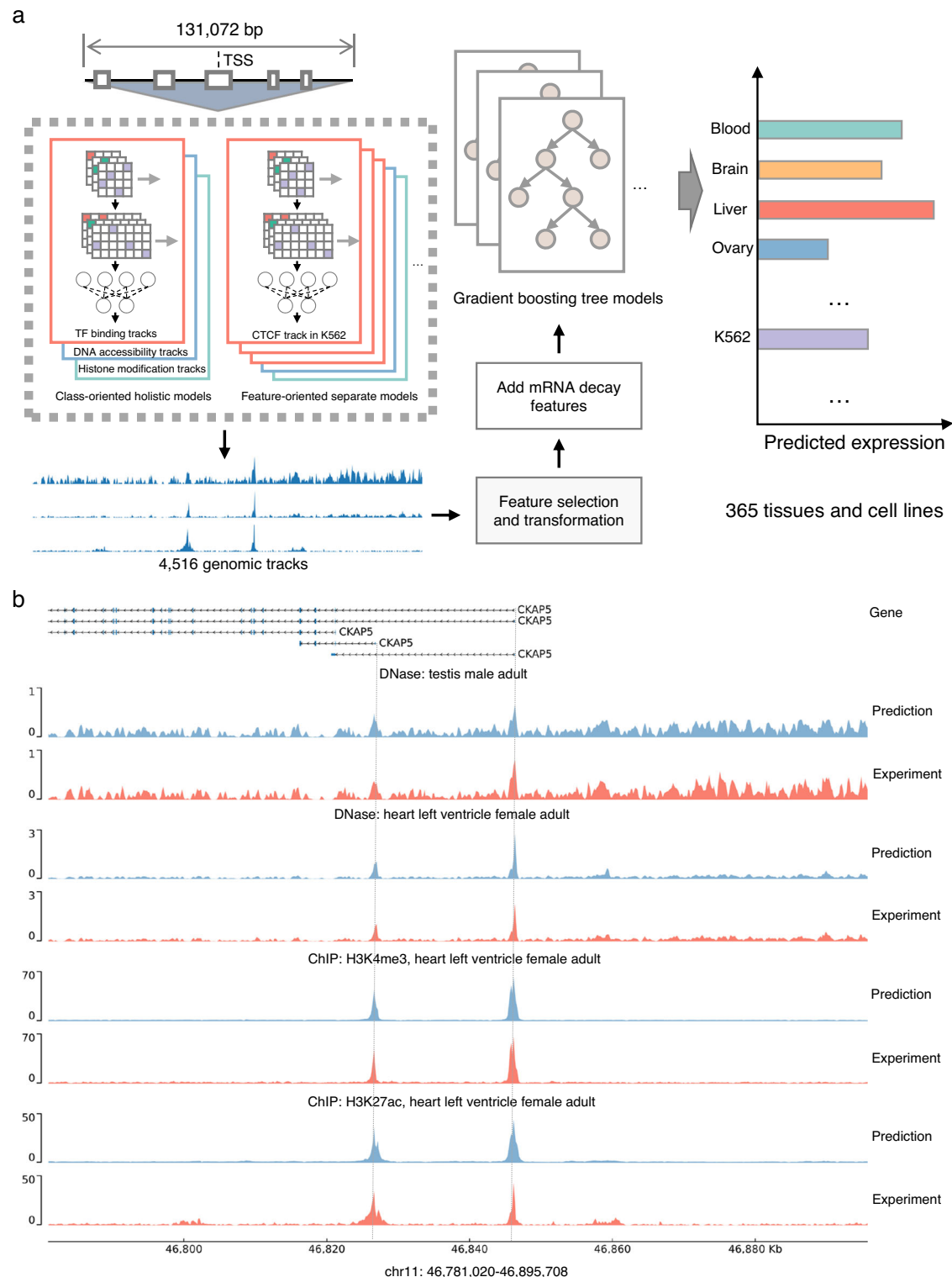
### SVEN predicts effects of SVs on tissue-specific gene expression accurately

SVEN employs a hybrid architecture to learn “regulatory codes” and infer the gene expression levels from transcription start site (TSS)-

<sup>1</sup>State Key Laboratory of Protein and Plant Gene Research, School of Life Sciences, Biomedical Pioneering Innovative Center (BIOPIC) and Beijing Advanced Innovation Center for Genomics (ICG), Center for Bioinformatics (CBI), Peking University, 100871 Beijing, China. <sup>2</sup>Changping Laboratory, 102206 Beijing, China. ✉ e-mail: [gaog@mail.cbi.pku.edu.cn](mailto:gaog@mail.cbi.pku.edu.cn)

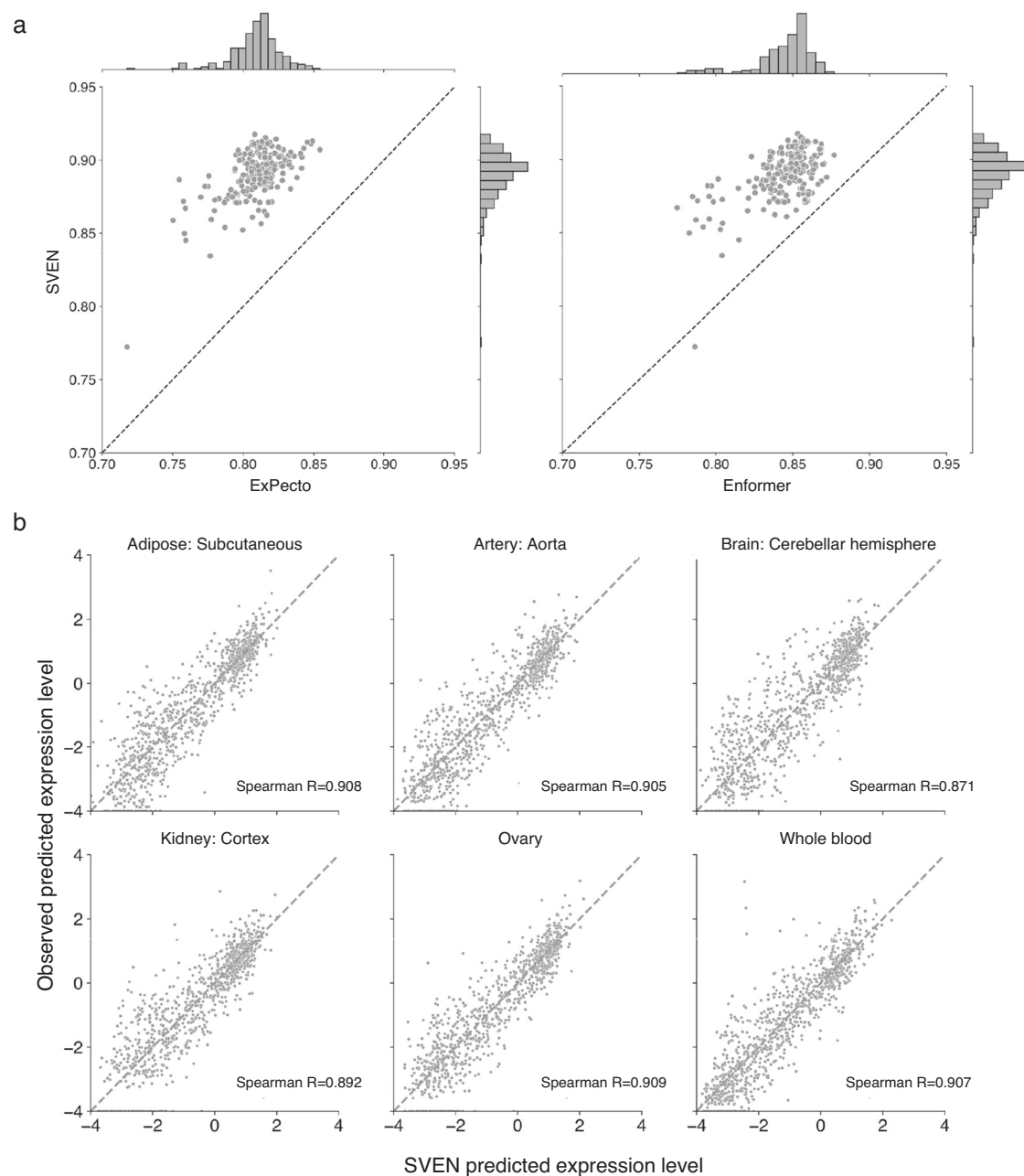
centered sequences in a tissue-specific manner (Fig. 1a and Methods). Briefly, we first trained multiple regulatory-specific neural networks based on 4516 transcription factor (TF) binding, histone modification and DNA accessibility features across over 400 tissues and cell lines

generated by ENCODE (Supplementary Fig. 1). Our evaluations suggested that these networks successfully learned the underlying regulatory codes from the inputs directly, with 0.660 mean Pearson correlation (Fig. 1b, Supplementary Fig. 2 and Supplementary Data 1). A



**Fig. 1 | Tissue-specific gene expression prediction framework. a** Schematic overview of SVEN. SVEN consists of three components that act sequentially: sequence-based deep neural networks to learn regulatory codes from sequences, feature selection and transformation to reduce the dimensionality of features, and gradient-boosting tree models to predict gene expression levels in a tissue-specific manner. TSS transcription start site, TF transcription factor. **b** Representative

examples of observed and predicted functional genomic features ( $\log_{10}$  scale) obtained from deep neural networks. The sequence (chr11:46,781,020–46,895,708) at the *CKAP5* locus was included in the test dataset of the networks. pyGenomeTracks<sup>62</sup> was used to plot genomic tracks. DNase deoxyribonuclease, ChIP chromatin immunoprecipitation. Source data are provided as a Source Data file.



**Fig. 2 | SVEN improves tissue-specific gene expression prediction. a** SVEN outperforms state-of-the-art tools in predicting tissue-specific gene expression on held-out sequences that were excluded from the training dataset for testing. The predicted  $\log_{10}$ RPKM values were compared with the observed  $\log_{10}$ RPKM values from RNA-seq data for overlapped 218 tissues and cell lines with Enformer and

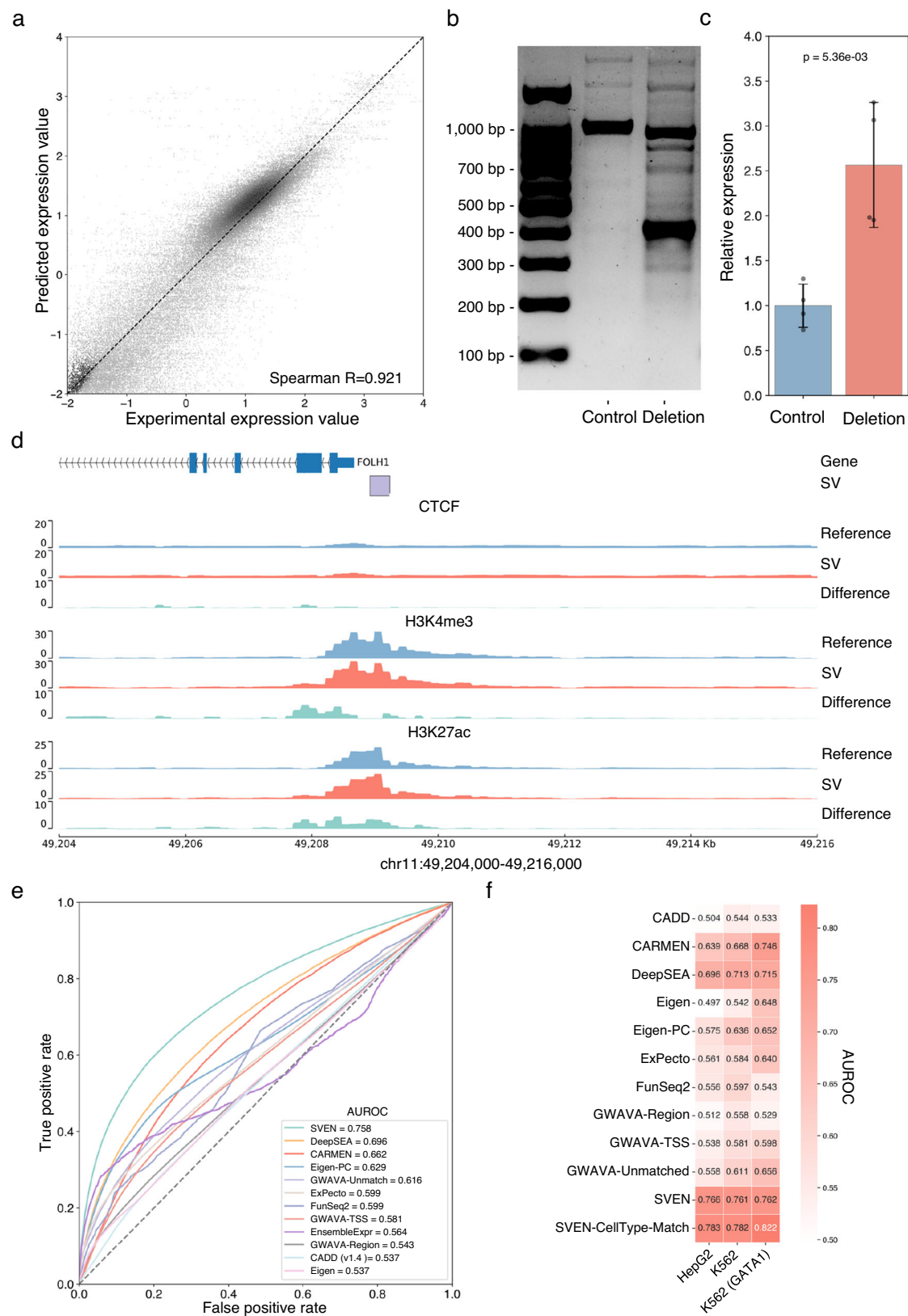
ExPecto. Spearman correlations between the predicted and observed values were calculated for comparison. **b** The values predicted by SVEN versus the observed values for all the testing sequences in the 6 example tissues reported by ExPecto and Enformer. Source data are provided as a Source Data file.

data-oriented feature selection procedure was further employed, and 802 networks associated with less-informative regulatory features based on model performance were excluded (Supplementary Fig. 3; see the Methods for more details). The outputs of the remaining 3714 networks, along with a separate mRNA decay-related feature set<sup>22</sup>, were used to train 365 gradient-boosting trees to infer gene expression in 365 tissues and cell lines.

To evaluate the performance of the model in predicting gene expression, we assessed the performance of SVEN on testing sequences that were excluded from the training dataset. SVEN accurately predicted gene expression levels across 365 tissues and cell types, with a mean Spearman correlation of 0.892 (Supplementary Data 2). Notably, SVEN showed consistently better performance across

different tissues than ExPecto<sup>19</sup> and Enformer (Fig. 2, Wilcoxon rank-sum one-sided test  $p = 7.070 \times 10^{-72}$  and  $5.464 \times 10^{-67}$ , respectively).

We then evaluated the ability of SVEN to predict the regulatory impact of SVs by applying SVEN to a large-scale SV dataset from 1019 samples with paired RNA-seq data<sup>23</sup> (Methods). For 94,366 SV-gene pairs from 126 samples (corresponding to 42,034 SVs and 19,995 genes), SVEN accurately predicted their regulatory effects, with a Spearman correlation of 0.921 (0.907 after removal of unexpressed genes) between the predicted expression levels and the observed expression levels from paired RNA-seq data (Fig. 3a). Compared with Enformer, SVEN can predict effects of SVs on gene expressions more accurately and consistently across different SV types (Table 1, Supplementary Fig. 4 and Supplementary Table 1).



Serum albumin is the most abundant protein in human blood encoded by gene *ALB*, and the synthesis of albumin occurs in liver. Serum albumin has important biological functions<sup>24</sup> in human and it suggests that low serum albumin has close relation with the emergence and worsening of cardiovascular diseases<sup>25</sup>. It has been found that a 4 kb deletion upstream of *ALB* affecting the promoter region

(Supplementary Fig. 5a) was associated with lower serum albumin level. We predicted the effect of this deletion with SVEN and it was predicted to decrease the expression of *ALB* in liver (log2 fold change = -1.30) and HepG2 cell line (log2 fold change = -3.29) significantly, which is consistent with serum albumin level of individuals carried with this deletion<sup>26</sup>.

**Fig. 3 | SVEN can accurately quantify the regulatory potential of genetic variants.** **a** Evaluation of SVEN in assessing the effects of SVs on gene expression. The predicted  $\log_{10}$ RPKM value obtained from the GMI2878 model was compared with the observed  $\log_{10}$ RPKM from RNA-seq data for the LCL cell line. Spearman correlation is shown. **b** DNA fragments generated by PCR using genomic DNA from non-targeting control A375 cells and those subjected to CRISPR-based deletion of the FOLHI-SV region. **c** Effect of CRISPR-based deletion on the relative abundance of *FOLHI* mRNA, as determined by real-time quantitative PCR (qPCR) analysis. All qPCRs were performed in 96-well plates, and for each set of measurements, 4 wells were used as technical replicates. Relative mean expression levels  $\pm$  standard deviation of all technical replicates are presented. Two-sided independent sample t-test was conducted for statistical analysis. Since there was a difference between

the target deletion and the designed deletion in the CRISPR experiment, we also used SVEN to predict the expected deletion in the CRISPR experiment, and the same conclusion was reached. **d** Annotations of FOLHI-SV from the SVEN annotation model. Differences in the annotation tracks revealed differences between the signal of the reference sequence and the signal of the deleted sequence ( $Signal_d - Signal_r$ ). Since the SVEN annotation did not include annotations for the A375 cell line, the CTCF, H3K4me3 and H3K27ac signals shown here are the features with the largest feature contributions in the SVEN model (A375). **e** ROC (Receiver Operating Characteristic) curves of small noncoding variants measured by in vitro massively parallel reporter assays. AUROC: Area Under the Receiver Operating Characteristic curve. **f** Model performance on cell line-specific small noncoding variants. Source data are provided as a Source Data file.

To further demonstrate the predictive ability of SVEN, we selected deletions from the large-scale SV datasets for experimental validation (Methods). For the top 5 deletions predicted by SVEN to have the strongest regulatory impacts (Supplementary Data 3), we ran independent CRISPR-based assays. SVEN successfully predicted the direction of the impact on gene expression for 4 deletions, and 2 of the results were statistically significant (Fig. 3b and c, Supplementary Fig. 5b and 6). Notably, the deletion upstream of the cancer biomarker PSMA-encoding gene *FOLHI*<sup>27–29</sup> disrupts the promoter region (Supplementary Fig. 5c) and the annotation-based algorithm predicted that this deletion would barely affect gene transcription (regulatory disruption score =  $-0.02$ )<sup>30</sup>; however, SVEN correctly predicted an increase in expression, partly because its annotation module indicated that the variant effectively increases expression-activating H3K4me3 and H3K27ac signals rather than the deleting known silencers<sup>31</sup> or insulators, which further suggested a plausible underlying mechanism for the effect of this deletion (Fig. 3d).

**Screening of SVs detected in large-scale population and clinical SVs**

Large-scale whole-genome sequencing studies have identified a number of SVs, however, their impacts on gene expression remain unclear. To estimate the regulatory effects of SVs genome-wide, we systematically screened a large-scale dataset derived from 3,622 samples<sup>32</sup>. For 159,018 curated SV-gene pairs corresponding to 70,749 SVs (Methods), SVEN predicted that most SVs did not significantly affect gene expression (Supplementary Fig. 7a). One of possible reasons is that all these SVs were detected in the samples without disease-related phenotypes.

In addition, we further assessed the impact of pathogenic SVs on gene expression. For pathogenic deletions (18,620 SV-gene pairs) and duplications (428 SV-gene pairs) derived from ClinVar (Methods), we noticed that known pathogenic deletions were more likely to affect gene expression than benign and population ones (Wilcoxon rank-sum one-sided test  $p = 1.878 \times 10^{-17}$  and 0, respectively), especially exon-disrupted SVs with longer length (Supplementary Fig. 7b, c and d). No statistically significant difference was detected for duplications (Wilcoxon rank-sum one-sided test  $p = 0.785$  and 0.212, respectively),

which may be attributed to the relatively lower statistical power caused by limited number of duplications. Interestingly, the results obtained by SVEN suggested that pathogenic deletions tended to decrease gene expression, and loss of function of target genes might be one reason for the pathogenic effects of these variants. For example, the SV nssv17171470 (chrX:139,530,701–139,530,862) is a 161 bp deletion located at the locus of the gene *F9*, which encodes coagulation factor IX and is associated with hemophilia B. SVEN predicted that this SV would decrease the expression of *F9* ( $\log_2$ -fold change =  $-10.83$  in the liver). This deletion disrupts the first exon and the promoter region of *F9* (Supplementary Fig. 7e), including the TF binding sites of HNF4A that has been reported to have a positive and significant correlation with the level of *F9* in the liver<sup>33</sup>.

**SVEN improves small noncoding variant effects prediction**

As a sequence-oriented model, SVEN should also be able to predict regulatory effects for small noncoding variants (length  $\leq 50$  bp). Therefore, we further evaluated the performance of SVEN with variants whose functional effects were directly measured by in vitro massively parallel reporter assays (MPRAs)<sup>34</sup>. We used 5,248,124 variants tested in a variety of cell types, including 30,121 positive variants and 5,218,003 negative variants. Here, we would call a variant “positive” as long as it is considered to be an expression-altering variant according to MPRA experiments. For each variant, we calculated predicted score  $S$  without any additional training to evaluate variant effects. Briefly,

$$S = \frac{1}{M} \sum_m \sum_n^N |p_{mn}^{alt} - p_{mn}^{ref}| \tag{1}$$

where  $M$  represents the number of features ( $M = 4516$  in this case, considering all functional annotations),  $N$  represents the number of 128-bp bins of each feature ( $N = 896$ ),  $p_{mn}^{alt}$  and  $p_{mn}^{ref}$  is predicted genomic signal for feature  $m$  at bin  $n$ . Compared with several state-of-the-art methods, SVEN showed the best performance, with the AUROC (Area Under the Receiver Operating Characteristic curve) improving by 8.9% (Fig. 3e).

Genetic variants can function in tissue and cell-line-specific manner<sup>35</sup>. To evaluate performance for cell line-specific variants, we extracted variants tested in the HepG2 and K562 cell lines from the benchmarking dataset, selecting 265 and 483 cell line-matched features in the HepG2 and K562 cell lines, respectively, from 4516 SVEN annotations. SVEN-CellType-Match used the same K562 annotations for the K562 and K562 (GATA1) variants. Notably, SVEN cell-line matched predictions outperformed cell-line agnostic predictions (Fig. 3f), further demonstrating the importance of the context-specific design of SVEN.

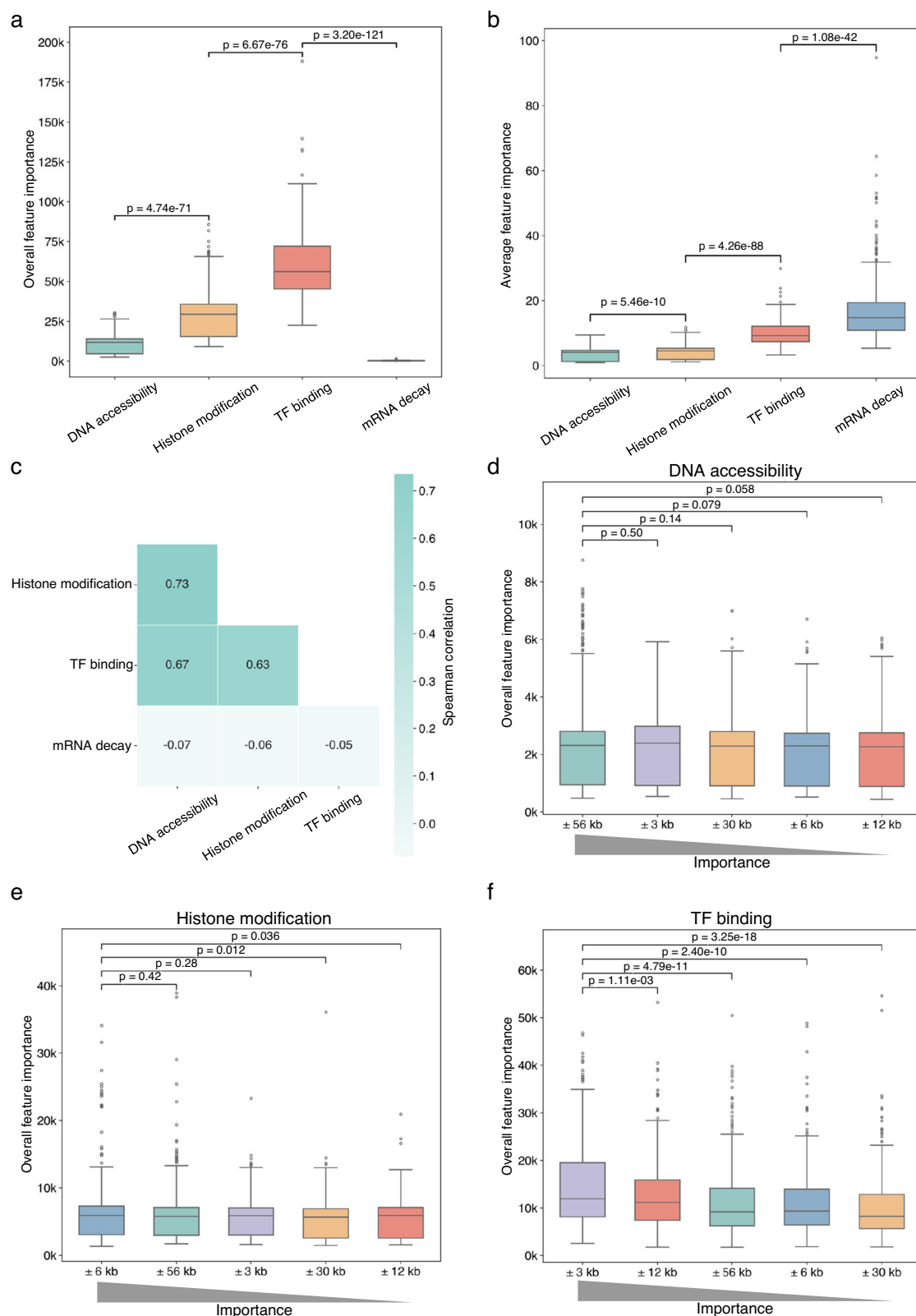
**Feature importance for the prediction of gene expression**

We further evaluated the importance of selected features for the prediction of expression levels (Methods). The results showed that TF binding features, as a class, made the largest overall contribution to expression outcomes (Fig. 4a), DNA accessibility features had the

**Table 1 | Model performance in assessing the effects of SVs on gene expression (Spearman R)**

SV type	Number of SV-gene pair	SVEN	Enformer
All	94,366 (61,811)	0.921 (0.907)	0.841 (0.786)
Insertion	31,352 (20,360)	0.921 (0.910)	0.841 (0.785)
MEI	2109 (1401)	0.928 (0.915)	0.869 (0.810)
Deletion	36,450 (24,186)	0.923 (0.905)	0.844 (0.785)
Duplication	8671 (5805)	0.923 (0.909)	0.839 (0.792)
Inversion	88 (62)	0.923 (0.899)	0.846 (0.794)
Complex	15,696 (9997)	0.914 (0.903)	0.831 (0.785)

Numbers in bracket are results after removal of unexpressed genes.



lowest mean importance across all 365 models, which may be partly due to their bidirectional biochemical effects: an accessible region of DNA can be bound by either activating or repressing regulatory factors. Notably, we found that mRNA decay features had the highest mean feature importance (Fig. 4b and c), likely owing to the close relationship between the mRNA decay rate and the regulation of gene expression<sup>36</sup>.

We used decay constants to control the receptive field of the transformed features (Methods and Supplementary Fig. 3b), and the preferences of different categories of features were diverse. For DNA accessibility features, there was no clear preference for features with different receptive field (Fig. 4d). However, the feature importance of histone modification features showed a different pattern: features from TSS-proximal regions and those with the largest receptive field

**Fig. 4 | Feature importance for predicting gene expression.** **a** Overall feature importance across all 365 models by feature category. **b** Mean feature importance across all 365 models by feature category. **c** Correlations between the feature contributions of different categories across all 365 models. The mean SHAP value was calculated for each model on all training and testing sequences. **d**, **e** and **f** Overall feature importance of **(d)** DNA accessibility, **(e)** histone modification and **(f)** TF binding features across all 365 models according to the decay constant used in feature transformation. We used different decay constants to control the receptive field of the transformed features: 0.01,  $\pm$  56 kb; 0.02,  $\pm$  30 kb; 0.05,  $\pm$

12 kb; 0.10,  $\pm$  6 kb; 0.20,  $\pm$  3 kb. Boxes were sorted based on the mean values of the data within each category. In each boxplot (**a**, **b** and **d–f**), the lower and upper boundaries of the box represent the first (Q1) and third quartiles (Q3), with the median indicated by a line inside the box. The whiskers typically extend to the most extreme data points within 1.5 times the interquartile range (IQR) from the quartiles. Data points outside this range are considered outliers and are plotted individually by circles. Wilcoxon rank-sum one-sided test (without adjustment) was conducted for statistical analysis (**a**, **b** and **d–f**). TF transcription factor. Source data are provided as a Source Data file.

had higher importance (Fig. 4e), further confirming the importance of histone modifications<sup>37</sup>. As expected, the TF binding features from TSS-proximal regions had the highest performance (Fig. 4f).

## Discussion

SVs are a common form of genetic variation that may contribute to diverse complex diseases; however, we still have a limited understanding of the functional impact of these variants. Previous studies try to assess impacts of SVs through classical annotation-oriented approaches<sup>38–42</sup> which highly rely on known functional annotations. While current sequence-oriented approaches (like ExPecto) for characterizing impacts of genetic variants mainly focus on small variants other than large-scale SVs<sup>16–19,43,44</sup>. In this study, we introduced SVEN, a multi-modality sequence-oriented model for quantifying genetic variants' regulatory impacts. By integrating deep learning networks to learn regulatory grammar from sequence ab initio and gradient boosting tree models to predict gene expression levels, SVEN was found to predict tissue-specific gene expression and the impact of large-scale SVs on gene expression accurately across more than 350 cell lines and tissues. SVEN is able to assess both SVs and small variants' impact at a whole-genome level and provide mechanism hints.

To achieve the high accuracy in predicting the impact of large-scale SVs on tissue-specific gene expression, we introduced a hierarchical hybrid architecture to build the model. On the one hand, SVEN consists of three components that act sequentially, which makes SVEN can benefit from existing genomic functional data rather than sequence only (such as Enformer<sup>21</sup>); on the other hand, we used hybrid architecture to train annotation component: class-oriented holistic models and feature-oriented separate models. Class-oriented holistic models could benefit from related features (such as same modification across different cell lines and tissues or biological-related features) and feature-oriented separate model paid more attention on sequence information alone. By combining these two kinds of models, annotation component could utilize available data more effectively and achieve better performance with relatively simple model structure.

SVs are of clinical interest because they have shown to have prominent impacts on several complex diseases. Several methods have been developed to predict pathogenicity of SVs<sup>38,40,42,45–48</sup>. We'd noted that SVEN was designed for quantifying the tissue-specific transcriptomic impacts of SVs, and the fact that gene expression alteration does not necessarily mean disease causing<sup>49–51</sup>, so we would expect that the SVEN will hardly surpass these pathogenicity-oriented algorithms in predicting the pathogenicity of SVs (Supplementary Table 2). Instead, we'd suggest that SVEN and these algorithms could be rather complementary, enabling a possibility to jointly identify pathogenic SVs, which function through transcriptomic changes.

There are still several paths for further improving the accuracy and scalability. Three-dimensional structure of human genome mediates the interaction between regulatory regions and regulates gene expression, and we could further combine the prediction of the three-dimensional structure of the human genome<sup>52,53</sup> to model distal regulation better as well as expand the scope of applicable genetic variants. CAGE (cap analysis gene expression) and RNA-seq are two major technologies used to quantify the transcription level of genes. Although the data from them are not fully comparable, as

complementary technologies, they can be used to improve gene expression prediction models<sup>54</sup>.

We believe that SVEN, an accurate and flexible sequence-oriented model, will enable more effective and efficient mining of disease-related genetic variants in the human genome. Thus, we constructed a whole SVEN package, which includes tutorials and demo cases and is publicly available online at <https://github.com/gao-lab/SVEN> for the community.

## Methods

### Framework architecture of SVEN

The SVEN framework consists of three components that act sequentially. The first component is a set of sequence-based deep neural networks that learn regulatory codes from sequences to predict functional genomic features such as TF binding, histone modification, and DNA accessibility. The second component is a feature selection and transformation approach to reduce the dimensionality of the generated features obtained from deep neural networks. Finally, the third component is a set of gradient-boosting tree models that are trained with selected features as well as mRNA decay features and used to make tissue-specific gene expression predictions.

The first component is a set of deep neural networks that were used to predict 4516 functional genomic features, including 1896 TF binding features, 1976 histone modification features, and 684 DNA accessibility features. Inspired by previous work<sup>20</sup>, the basic neural network consists of three parts: (1) 7 residual convolutional blocks with pooling layers, (2) 11 residual dilated convolutional blocks, and (3) a cropping layer followed by a pointwise convolution block and fully connected layer for the output (Supplementary Fig. 1a). The input sequence of the models was a one-hot encoded DNA sequence of 131,072 bp, and the output was (predicted) functional genomic features with a length of 896 corresponding to 114,688 centered base pairs aggregated into 128-bp bins. The residual convolution blocks with pooling layers were used to extract sequence motifs and learn the interactions between them, and the dimensionality was reduced to 1024. Then, we used residual dilated convolutional blocks to learn long-range interactions across sequences. Finally, we applied a cropping layer to trim off 64 units at the beginning and end of the sequence due to the potential loss of information in these regions, where only one-sided information can be observed. Then, we used a pointwise convolution block to change the number of channels, followed by a fully connected layer with a soft plus function as the activation function as the final output layer.

Inspired by our previous work<sup>55</sup>, we introduced a hybrid architecture including class-oriented holistic models and feature-oriented separate models to train deep neural networks. Specifically, we trained 3 class-oriented holistic models for each type of chromatin profile (TF binding, histone modification and DNA accessibility); this approach could benefit from related features (such as the same binding protein or modification across different cell lines and tissues). The outputs of three class-oriented models were  $N \times 896 \times 1896$ ,  $N \times 896 \times 1976$  and  $N \times 896 \times 684$  respectively ( $N$ : number of input sequences). We also trained feature-oriented separate models, which paid more attention to sequence information alone, with each model corresponding to one feature (the output was  $N \times 896 \times 1$ ). Finally, we selected the best models

according to their performance on the validation set and combined the outputs of these models as the final output.

In addition, we found that DNA accessibility information could improve the model performance of TF binding and histone modification models; therefore, we incorporated a pretrained DNA accessibility model into the TF binding and histone modification models (Supplementary Fig. 1b). Specifically, we first trained a holistic DNA accessibility model as a pretrained model. Then, we incorporated the pretrained model into the TF binding and histone modification model; the pretrained model was not frozen and was involved in further model training. The pretrained model was trained on the same sequences as the TF binding and histone modification models.

The second component is feature selection and transformation. First, we filtered all 4516 genomic features according to the model performance (Supplementary Fig. 3a). We removed 802 features with Pearson  $R < 0.5$  and retained 3714 genomic features. Then, we transformed these features via the method described in ExPecto<sup>19</sup>. Briefly, to reduce the dimensionality of the features, we transformed features with 10 exponential functions to weight the upstream and downstream regions of the TSS separately based on the assumption that regions with longer distances to the TSS usually have weaker effects on gene expression (Supplementary Fig. 3b). We also use 5 different decay constants {0.01, 0.02, 0.05, 0.10, and 0.20} to control the receptive field of the transformed features. The number of features decreased from 3,327,744 ( $3714 \times 896$ ) to 37,140 ( $3714 \times 10$ ).

In addition, we incorporated mRNA decay features for subsequent model training. We extracted the GC content and the lengths of the 3'UTRs and 5'UTRs from Ensembl (104, GRCh38). We calculated the minimum, maximum, median, and mean values of all 3'UTRs and 5'UTRs for the target genes and generated 16 features for each gene. If there was no known 3'UTR or 5'UTR for the target gene, the value of the feature was set to 0.

We combined the 37,140 transformed chromatin features and 16 mRNA decay features for further feature selection. We used all 37,156 features to train 365 extreme gradient-boosting (XGBoost) regression tree models, and each model corresponded to one tissue or cell line. Then, we selected the features used in the trained models, retrained the models with the selected features and repeated this procedure several times until the model performance decreased or the number of features no longer decreased. We found that the number of features of all models stopped decreasing after the third round of selection (Supplementary Fig. 3c). Therefore, we used the features obtained after the third-round selection as the final feature set.

The third component is a set of 365 gradient-boosting tree models, each corresponding to one tissue or cell type. We used the selected features to retrain all XGBoost regression tree models as the final models.

### Model training and evaluation of SVEN

For the first component of SVEN, the deep neural networks were trained, evaluated, and tested on the same sequences of 4516 ENCODE chromatin features extracted from 5313 features used in Basenji2. The dataset contained 34,021 training, 2213 validation, and 1937 test sequences. The length of the sequence was 131,072 bp, and all sequences were based on the GRCh38 reference genome.

We used the Poisson negative log-likelihood function as the loss function and Adam as the optimizer, with default parameters. All models were implemented on TensorFlow (v2.5.0) and trained on 8 NVIDIA Tesla A100 GPUs with a batch size of 32 for 1000 epochs with early stopping. The validation set was used for hyperparameter selection and model selection, and the performance on the test set was reported as the final performance of all models. Pearson correlation was used to evaluate model performance; this is identical to the metric used by Basenji2 and Enformer. We used pretrained Basenji2 and Enformer models for model performance comparison.

For the third component of SVEN and for feature selection, we used the same XGBoost (v2.0.1) regression tree models. We used the representative TSSs of genes (lifted to GRCh38 coordinates) used by ExPecto to construct the training and evaluation dataset. Briefly, the expression profiles of 365 tissues and cell lines were obtained from the GTEx, Roadmap and ENCODE projects. We only used expression profiles of protein-coding genes (18,632) and lincRNA genes (5068). Then, we added a pseudocount of 0.01 (0.0001 for GTEx tissues due to high coverage) and applied log10 transformation for model training. We used trained deep neural networks to annotate all sequences centered on the TSS of each gene for all genomic signals for subsequent feature selection and model training.

All genes from chromosome 8 were excluded from training for testing (987 genes), and all other genes were used for model training (22,713 genes). However, to determine the hyperparameters of the regression tree models, we further selected all genes from chromosome 9 (939 genes) from the training set as the temporary validation set, and the remaining genes (21,774 genes) were used to train the models. Then, we evaluated the performance of the model on the temporary validation set for hyperparameter selection. After hyperparameter selection, we used fixed hyperparameters to retrain the models for feature selection with the whole training dataset (22,713 genes) and evaluated model performance on the held-out test set as the final model performance. All the models were trained on NVIDIA Tesla A100 GPUs. Spearman correlation was used to evaluate model performance; this is identical to the metric used by ExPecto and Enformer.

For comparison with ExPecto, we retrained 218 ExPecto gene expression prediction models with official codes on the original training and testing dataset; for comparing with Enformer, we trained 218 Elastic Net models with all Enformer CAGE predictions (ten 128-bp bins centered with TSS) on training and testing sequences used by SVEN. The model performance of retrained models was comparable with reported performance in their papers (median Spearman correlations: 0.812 for ExPecto and 0.850 for Enformer).

To test each module's contribution of SVEN, we made several tests based on gene expression profiles across 218 tissues and cell lines used by ExPecto, Enformer and SVEN. We replaced SVEN annotation module with Enformer annotations (same 4,516 functional genomic features), removed mRNA decay features and replaced gradient-boosting trees with linear models, respectively (Supplementary Data 4).

### Feature importance analysis

We used the built-in function of XGBoost to determine the importance of features in the trained regression tree models. The importance type we used was "gain" (the average gain of splits that use the feature). To investigate the contributions of features to the prediction, we calculated the SHAP<sup>56</sup> value of each feature on all training and testing sequences. Then, we calculated the mean feature importance of each category or decay constant in each model and combined the feature importance or feature contributions from all 365 models for comparison. Single-tailed Wilcoxon tests and Spearman correlations were used for statistical analysis.

### Prediction of the effects of SVs on gene expression

The effects of SVs on gene expression were predicted on the basis of the differences between the predicted expression levels of sequences with reference alleles and alternative alleles. Here, we used the log2 fold change to measure the effects of SVs on target genes. First, we checked the reference alleles and alternative alleles of the SVs. The full sequence of SVs is necessary for SVEN, which is a sequence-based framework. Then, we checked the position of the SV to determine whether the whole SV (all bases) fell within the region of 131,072 bp centered on the TSS of any gene. If there was no overlapping gene, we excluded this SV from further prediction of expression effects. The TSSs of genes were the same as the representative TSSs used in model

training. For the included genes, we constructed SV-gene pairs and predicted the effects of SVs on all included genes. Specifically, we extracted the sequences of each gene from the 5' to 3' ends and generated reference sequences and alternative sequences. If no reference allele was specified, we used the sequences from the reference genome (GRCh38) as reference sequences. Then, we used SVEN to predict the expression levels of the included genes with reference and alternative sequences. As the output of SVEN was the log-transformed value of the expression level, we transformed the predicted values to the original values and then calculated the log<sub>2</sub> fold change in gene expression for the included genes as the final output.

### Evaluation of the ability of SVEN to predict the effects of SVs on gene expression

To evaluate the ability of SVEN to predict the effects of SVs on tissue-specific gene expression, we applied SVEN to a large-scale SV dataset from 1019 samples from the 1000 Genome Project generated by Schloissnig et al.<sup>23</sup> via long-read sequencing, which contains 164,625 SVs. We first filtered SVs (lifted to GRCh38 coordinates) falling into the 131 kb sequences centered on the TSSs of lincRNAs and protein-coding genes and constructed SV-gene pairs. We obtained 205,438 SV-gene pairs corresponding to 92,943 SVs and 22,453 genes.

We also downloaded paired-end RNA-seq fastq files from Geuvadis project<sup>57</sup> and single-end RNA-seq fastq files from Human Genome Structural Variation Consortium Phase 2 (HGSVC2). RNA-seq data were mapped to the reference genome GRCh38 with HISAT2 (v2.2.1)<sup>58</sup>. The hisat2 index was built with the reference genome GRCh38 and transcript information extracted from the GENCODE (v24) annotation. Then, we mapped the RNA-seq reads using HISAT2 with default parameters. The gene abundance data was generated by StringTie (v2.2.1)<sup>59</sup>. We used the expression estimation mode to estimate the coverage of the transcripts on the basis of the GENCODE (v24) annotation. For replicated experiments, we calculated the mean value of all replicates as sample RPKM value (per gene). We used the log<sub>10</sub>-transformed expression levels for subsequent analysis.

To match SVs with gene expression levels, we used the expression values of the lead sample (the first sample with paired RNA-seq data, if any) as the target expression level. We filtered SVs with RNA-seq data and obtained 94,366 SV-gene pairs from 126 samples corresponding to 42,034 SVs and 19,995 genes. The RNA-seq data was generated from EBV-transformed lymphoblastoid cell lines (LCLs), therefore, we used SVEN (GM12878) to predict the effects of these SVs on gene expression and compared the predicted values with the observed values. Spearman correlation was used to evaluate the model performance.

We also used an SV dataset from 32 diverse human genomes generated by Ebert et al.<sup>60</sup> to evaluate the ability of SVEN. We obtained 129,815 SV-gene pairs with paired RNA-seq data from 26 samples corresponding to 57,893 SVs and 20,805 genes for following assessment.

### Estimation of the effects of SVs in large populations and pathogenic SVs

To better estimate the effects of SVs on gene expression in large populations, we applied SVEN to a SV dataset from 3622 samples generated by Beyter et al.<sup>32</sup> using long-read sequencing. The dataset contains 133,886 SVs, including 68,332 insertions, 26,370 deletions, 28,571 complex SVs and 10,613 sequence-unresolved SVs. We filtered the SVs as described above and obtained 159,018 SV-gene pairs corresponding to 70,749 SVs and 21,998 genes. To estimate the effects of these SVs in the population, we predicted the effects of these SVs in 53 GTEx tissues and calculated the maximum effect (absolute value) across all tissues.

Similarly, for disease-related SVs, we used SVs with clinical annotations ("pathogenic" and "benign") from dbVar (nstd102, 20231012). We filtered the SVs and selected deletions (11,509), duplications (573) and inversions (14) due to the lack of sequence information for other

types of SVs. Inversions were excluded owing to the limited number of SVs and finally we obtained 19,262 SV-gene pairs (pathogenic: 19,048; benign: 214). We predicted the effects of pathogenic SVs in 53 GTEx tissues and calculated the maximum effect (absolute value) across all 53 GTEx tissues, and compared with reported benign deletions (175 SV-gene pairs) and duplications (39 SV-gene pairs), as well as deletions (84,881 SV-gene pairs) and duplications (18,777 SV-gene pairs) detected in large-scale population<sup>23</sup>.

### Experimental validation

We selected deletions from the SV datasets generated by Ebert et al. and Beyter et al. as candidates for experimental validation. We filtered deletions with length less than 1 kb and got 61,114 SV-gene pairs. To evaluate the generalization ability of SVEN, we focused on deletions predicted to cause the upregulation of gene expression in four cell lines (HepG2, MCF-7, A375 and A549). We selected the top 5 deletions (protein-coding gene, gene expression RPKM > 1) with available CRISPR target sites according to their effects on the target gene in the target cell line and validated these 5 deletions in A375 cell line (obtained from Cell Resource Center, Peking Union Medical College).

The guide RNA pairs that were closest to the targeted SV boundaries were selected from the UCSC genome browser CRISPR Targets track.

DNA fragments carrying guide RNA pairs (pgRNA) were generated by performing PCR on the mU6-pgRNA-4.0 plasmid with the primers pgRNA-F and pgRNA-R (Supplementary Data 5). These fragments were then subjected to agarose gel electrophoresis, purified with a DNA Clean Kit (Zymo #D4003), and assembled into the sgRNA-SV40-PURO plasmid using the Golden Gate method (NEB #E1602). The success of these assemblies was confirmed by Sanger sequencing.

The pgRNA lentiviruses were produced by transfecting 293T cells at 70% confluence with 1 µg of sgRNA-SV40-PURO, 1 µg of pCMVR8.74 (Addgene #22036), and 0.1 µg of pVSV-G (Addgene #138479) in each well of a 6-well plate. Transfections were carried out using PEI (Proteintech #PR40001). The lentiviruses were harvested by collecting supernatants from the 293T cell culture at 72 hours post-transfection.

Monoclonal cell lines with constitutive Cas9 expression were generated by transduction with a lentivirus carrying a Cas9-2A-mCherry construct. These cells were then transduced with the target pgRNA lentiviruses when they reached 70% confluence. 0.5 µg/mL Puromycin (InvivoGen #ant-pr-1) was added to the culture medium 24 hours post-transduction. The cells were cultured until all negative control cells (non-transduced) had died, and the cells were allowed to reach 70% confluence again (~6 days). The cells were then cultured in fresh medium without puromycin for 24 hours before being collected. All cells were maintained in DMEM (HyClone #SH30243.01) supplemented with 10% fetal bovine serum (Gibco #10091148) and 1× penicillin–streptomycin–amphotericin B solution (Solarbio #P7630).

Nucleic acids from the cells were obtained by direct lysis on plates using an AllPure DNA/RNA kit (Magen #R5111) following the manufacturer's instructions.

We converted total RNA to cDNA by using HiScript III RT Super-Mix for qPCR (Vazyme #R323-01) following the manufacturer's instructions. We selected *RPL41* as an internal reference gene because it has the most consistently high expression levels in the HPA database of 1055 cell lines. All qPCRs were performed in 96-well plates using a Roche LightCycler 480 machine. Each well contained 10 µL of ChamQ Universal SYBR qPCR Master Mix (Vazyme #Q711-03), 0.4 µL of 10 µM forward primer, 0.4 µL of reverse primer, and 9.2 µL of cDNA (diluted 1:25). For each set of measurements, 4 wells were used as technical replicates.

We performed melt curve analyses to ensure that the amplicons produced by the same pair of primers were specific and consistent each time. Each quantification cycle value was calculated using the second derivative maximum method recommended in the manual of

the instrument. All the above steps were carried out according to the manufacturers' instructions.

### Statistics and reproducibility

The detailed statistical tests were explained in each figure legend. Sample data were obtained from public repositories. No statistical method was used to predetermine the sample size. No data were excluded from the analyses. The experiments were not randomized. The Investigators were not blinded to allocation during experiments and outcome assessment. More information was provided in the Reporting Summary file.

### Reporting summary

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

### Data availability

Basenji2 training, validation and testing sets, as well as the trained model, were obtained from [https://console.cloud.google.com/storage/browser/basenji\\_barnyard](https://console.cloud.google.com/storage/browser/basenji_barnyard). ExPecto training and testing sequences for tissue-specific gene expression prediction, TSS annotations of genes as well as training codes of ExPecto were obtained from <https://github.com/FunctionLab/ExPecto>. Trained Enformer was obtained from <https://github.com/deepmind/deepmind-research/tree/master/enformer>. SVs and paired RNA-seq data for 1019 and 26 samples were obtained from the 1000 Genome Project (<https://www.internationalgenome.org/>). SVs from 3622 Icelanders were obtained from [https://github.com/DecodeGenetics/LRS\\_SV\\_sets](https://github.com/DecodeGenetics/LRS_SV_sets). Pathogenic and benign SVs were obtained from dbVar at <https://www.ncbi.nlm.nih.gov/dbvar/studies/nstd102/>. Gene annotations were obtained from <https://www.genecodegenes.org/> (v24, GRCh38). Benchmark dataset of small noncoding variants is available at <https://reva.gao-lab.org/>. Source data are provided with this paper.

### Code availability

All codes for training and evaluating models, as well as trained models and detailed parameters, are available at <https://github.com/gao-lab/SVEN> and <https://doi.org/10.5281/zenodo.14281154><sup>61</sup>.

### References

- Abecasis, G. R. et al. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
- Abecasis, G. R. et al. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
- Sudmant, P. H. et al. An integrated map of structural variation in 2504 human genomes. *Nature* **526**, 75–81 (2015).
- Collins, R. L. et al. A structural variation reference for medical and population genetics. *Nature* **581**, 444–451 (2020).
- Byrsk-Bishop, M. et al. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**, 3426–3440 (2022).
- Abel, H. J. et al. Mapping and characterization of structural variation in 17,795 human genomes. *Nature* **583**, 83–89 (2020).
- Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* **12**, 363–376 (2011).
- Feuk, L., Carson, A. R. & Scherer, S. W. Structural variation in the human genome. *Nat. Rev. Genet.* **7**, 85–97 (2006).
- Hurles, M. E., Dermizakis, E. T. & Tyler-Smith, C. The functional impact of structural variation in humans. *Trends Genet.* **24**, 238–245 (2008).
- Weischenfeldt, J., Symmons, O., Spitz, F. & Korbel, J. O. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat. Rev. Genet.* **14**, 125–138 (2013).
- Stranger, B. E. et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**, 848–853 (2007).
- Chiang, C. et al. The impact of structural variation on human gene expression. *Nat. Genet.* **49**, 692–699 (2017).
- Ritchie, G. R. S., Dunham, I., Zeggini, E. & Flicek, P. Functional annotation of noncoding sequence variants. *Nat. Methods* **11**, 294–296 (2014).
- 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**, 214–220 (2016).
- Huang, Y., Gulko, B. & Siepel, A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat. Genet.* **49**, 618–624 (2017).
- Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
- Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12**, 931–934 (2015).
- Zeng, H., Edwards, M. D., Guo, Y. & Gifford, D. K. Accurate eQTL prioritization with an ensemble-based framework. *Hum. Mutat.* **38**, 1259–1265 (2017).
- Zhou, J. et al. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat. Genet.* **50**, 1171–1179 (2018).
- Kelley, D. R. Cross-species regulatory sequence activity prediction. *PLoS Comput. Biol.* **16**, e1008050 (2020).
- Avsec, Ž. et al. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **18**, 1196–1203 (2021).
- Agarwal, V. & Shendure, J. Predicting mRNA abundance directly from genomic sequence using deep convolutional neural networks. *Cell Rep.* **31**, 107663 (2020).
- Schloissnig, S. et al. Long-read sequencing and structural variant characterization in 1019 samples from the 1000 Genomes Project. Preprint at <https://www.biorxiv.org/content/10.1101/2024.04.18.590093v1> (2024).
- Chien, S., Chen, C., Lin, C. & Yeh, H. Critical appraisal of the role of serum albumin in cardiovascular disease. *Biomark. Res.* **5**, 31 (2017).
- Arques, S. Human serum albumin in cardiovascular diseases. *Eur. J. Intern. Med.* **52**, 8–12 (2018).
- Chen, L. et al. Association of structural variation with cardiometabolic traits in Finns. *Am. J. Hum. Genet.* **108**, 583–596 (2021).
- Noss, K. R., Wolfe, S. A. & Grimes, S. R. Upregulation of prostate specific membrane antigen/folate hydrolase transcription by an enhancer. *Gene* **285**, 247–256 (2002).
- Ren, H. et al. Prostate-specific membrane antigen as a marker of pancreatic cancer cells. *Med. Oncol.* **31**, 857 (2014).
- Ciappuccini, R. et al. PSMA expression in differentiated thyroid cancer: association with radioiodine, 18FDG uptake, and patient outcome. *J. Clin. Endocrinol. Metab.* **106**, 3536–3545 (2021).
- Han, L. et al. Functional annotation of rare structural variation in the human brain. *Nat. Commun.* **11**, 2990 (2020).
- Doni Jayavelu, N., Jajodia, A., Mishra, A. & Hawkins, R. D. Candidate silencer elements for the human and mouse genomes. *Nat. Commun.* **11**, 1061 (2020).
- Beyter, D. et al. Long-read sequencing of 3622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nat. Genet.* **53**, 779–786 (2021).
- Salloum-Asfar, S. et al. MiRNA-based regulation of hemostatic factors through hepatic nuclear factor-4 alpha. *PLoS One* **11**, e0154751 (2016).
- Wang, Y., Shi, F., Liang, Y. & Gao, G. REVA as a well-curated database for human expression-modulating variants. *Genom. Proteom. Bioinforma.* **19**, 590–601 (2021).
- Barbeira, A. N. et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat. Commun.* **9**, 1825 (2018).

36. Yang, E. et al. Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes. *Genome Res* **13**, 1863–1872 (2003).
37. Zhang, Y. et al. edited by D. Fang and J. Han, 1283, pp. 1–16 (Springer Singapore, Singapore, 2020).
38. Ganel, L., Abel, H. J. & Hall, I. M. SVScore: an impact prediction tool for structural variation. *Bioinformatics* **33**, btw789 (2017).
39. Gurbich, T. A. & Ilinsky, V. V. ClassifyCNV: a tool for clinical annotation of copy-number variants. *Sci. Rep.* **10**, 20375 (2020).
40. Kumar, S., Harmanci, A., Vytheswaran, J. & Gerstein, M. B. SVFX: a machine learning framework to quantify the pathogenicity of structural variants. *Genome Biol.* **21**, 274 (2020).
41. Zhang, L. et al. X-CNV: genome-wide prediction of the pathogenicity of copy number variations. *Genome Med* **13**, 132 (2021).
42. Sharo, A. G., Hu, Z., Sunyaev, S. R. & Brenner, S. E. StrVCTVRE: a supervised learning method to predict the pathogenicity of human genome structural variants. *Am. J. Hum. Genet.* **109**, 195–209 (2022).
43. Fu, Y. et al. FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol.* **15**, 480 (2014).
44. Caron, B., Luo, Y. & Rausell, A. NCBoost classifies pathogenic non-coding variants in Mendelian diseases through supervised learning on purifying selection signals in humans. *Genome Biol.* **20**, 32 (2019).
45. Sánchez-Gaya, V. & Rada-Iglesias, A. POSTRE: a tool to predict the pathological effects of human structural variants. *Nucleic. Acids. Res.* (2023).
46. Hertzberg, J., Mundlos, S., Vingron, M. & Gallone, G. TADA—a machine learning tool for functional annotation-based prioritisation of pathogenic CNVs. *Genome Biol.* **23**, 67 (2022).
47. Kleinert, P. & Kircher, M. A framework to score the effects of structural variants in health and disease. *Genome Res* **32**, 766–777 (2022).
48. Danis, D. et al. SvAnna: efficient and accurate pathogenicity prediction of coding and regulatory structural variants in long-read genome sequencing. *Genome Med* **14**, 44 (2022).
49. Corbett, A. H. Post-transcriptional regulation of gene expression and human disease. *Curr. Opin. Cell. Biol.* **52**, 96–104 (2018).
50. Sun, B. B. et al. Genetic associations of protein-coding variants in human disease. *Nature* **603**, 95–102 (2022).
51. Wang, S. & Sun, S. Translation dysregulation in neurodegenerative diseases: a focus on ALS. *Mol. Neurodegener.* **18**, 58 (2023).
52. Fudenberg, G., Kelley, D. R. & Pollard, K. S. Predicting 3D genome folding from DNA sequence with Akita. *Nat. Methods* **17**, 1111–1117 (2020).
53. Schwessinger, R. et al. DeepC: predicting 3D genome folding using megabase-scale transfer learning. *Nat. Methods* **17**, 1118–1124 (2020).
54. Kawaji, H. et al. Comparison of CAGE and RNA-seq transcriptome profiling using clonally amplified and single-molecule next-generation sequencing. *Genome Res* **24**, 708–717 (2014).
55. Shi, F. et al. Computational assessment of the expression-modulating potential for non-coding variants. *Genom. Proteom. Bioinforma.* **21**, 662–673 (2023).
56. Lundberg, S. M. et al. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2**, 56–67 (2020).
57. Lappalainen, T. et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
58. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
59. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
60. Ebert, P. et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, eabf7117 (2021).
61. Wang, Y., Liang, N. & Gao, G. Quantifying the regulatory potential of genetic variants via a hybrid sequence-oriented model with SVEN. SVEN Model, <https://doi.org/10.5281/zenodo.14281154> (2024).
62. Lopez-Delisle, L. et al. pyGenomeTracks: reproducible plots for multivariate genomic datasets. *Bioinformatics* **37**, 422–423 (2021).

## Acknowledgements

This work was supported by funds from the National Science and Technology Major Project (grant no. 2022ZD0115004), the Changping Laboratory, the State Key Laboratory of Protein and Plant Gene Research and the Beijing Advanced Innovation Center for Genomics at Peking University. Part of the analysis was carried out on the Computing Platform of the Center for Life Sciences of Peking University and supported by the High-performance Computing Platform of Peking University and Changping Laboratory.

## Author contributions

G.G. conceived the study and supervised the research. Y.W. designed and implemented the computational framework and conducted benchmarks and case studies with guidance from G.G. N.L. conducted the experimental validation. Y.W. and G.G. wrote the manuscript with inputs from all the authors.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-024-55392-7>.

**Correspondence** and requests for materials should be addressed to Ge Gao.

**Peer review information** *Nature Communications* thanks Kai Wang, who co-reviewed with Zhuoran Xu, Kenta Nakai and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024