CREATE: cell-type-specific cis-regulatory elements identification via discrete embedding

Xuejian Cui[1], Qijin Yin[1], Zijing Gao[1], Zhen Li[1], Xiaoyang Chen[1], Shengquan Chen[2], Qiao Liu[3], Wanwen Zeng[3,*], and Rui Jiang[1,*]

[1] Ministry of Education Key Laboratory of Bioinformatics, Bioinformatics Division at the Beijing National Research Center for Information Science and Technology, Center for Synthetic and Systems Biology, Department of Automation, Tsinghua University, Beijing 100084, China

[2] School of Mathematical Sciences and LPMC, Nankai University, Tianjin 300071, China

[3] Department of Statistics, Stanford University, Stanford, CA 94305, USA

[*] Corresponding author: wanwen@stanford.edu, ruijiang@tsinghua.edu.cn

Identifying cis-regulatory elements (CREs) within non-coding genomic regions—such as enhancers, silencers, promoters, and insulators—is pivotal for elucidating the intricate gene regulatory mechanisms underlying complex biological traits. The current prevalent sequence-based methods often focus on singular CRE types, limiting insights into cell-type-specific biological implications. Here, we introduce CREATE, a multimodal deep learning model based on the Vector Quantized Variational AutoEncoder framework, designed to extract discrete CRE embeddings and classify multiple CRE classes using genomic sequences, chromatin accessibility, and chromatin interaction data. CREATE excels in accurate CRE identification and exhibits strong effectiveness and robustness. We showcase CREATE's capability in generating comprehensive CRE-specific feature spectrum, offering quantitative and interpretable insights into CRE specificity. By enabling large-scale prediction of CREs in specific cell types, CREATE facilitates the recognition of disease- or phenotype-related biological variabilities of CREs, thereby expanding our understanding of gene regulation landscapes.

Introduction

Gene regulation is a fundamental biological process that orchestrates gene expression through a sophisticated network of interactions among biomolecules, including regulatory factors and cis-regulatory elements (CREs) located in non-coding genomic regions[1, 2]. CREs, such as silencers, enhancers, promoters, and insulators[3, 4], typically locate in the chromatin accessible areas[5, 6] and play crucial roles in modulating gene expression by interacting with target genes through chromatin loops[7, 8] and other regulatory mechanisms. These characteristics are essential for controlling cell-type-specific gene expression patterns, which contribute to cellular diversity, tissue homeostasis, and the development of complex biological traits[9-11]. Consequently, identifying and characterizing cell-type-specific CREs is vital for advancing our understanding of gene regulation in normal physiology and disease states.

Silencers, enhancers, promoters, and insulators each have distinct roles in gene regulation. Silencers suppress gene transcription, enhancers boost transcriptional activity, promoters initiate transcription, and insulators act as boundary elements to regulate gene expression[3, 4]. Due to the restricted understanding of CRE-specific genetic signatures, identifying and validating CREs through biological experiments is cumbersome, time- and resource-consuming[9, 12]. Massive genomic and epigenomic data benefited from the rapid advancement of high-throughput sequencing technologies[13-16], has provided valuable opportunities for identifying cell-type-specific CREs using computational methods. For example, DeepSEA is a convolutional neural network (CNN) model based on genomic sequences, which can simultaneously predict chromatin-profiling data such as transcriptional factors (TFs) binding sites, histone modification sites, and chromatin accessible regions[17]. DanQ is a hybrid deep neural network that merges convolutional and recurrent architectures, aimed at quantifying the non-coding function of DNA sequences[18]. Enhancer-Silencer transition (ES-transition) is a deep learning model based on CNN for identifying enhancers and silencers specific to cell types in the human genome, and has been utilized to uncover the unique phenomenon of enhancer-silencer transitions[19]. DeepICSH integrates DNA sequences with various epigenetic features including histone modifications, chromatin accessibility and TF binding to predict cell-type-specific silencers[20].

However, recently innovated computational methods encounter many limitations and challenges. First, most current efforts are dedicated to the identification of a single type of CRE[20-23], particularly enhancers. In the past few decades, there has been extensive research on enhancers[24-30], while silencers, which generally share many properties with enhancers[31], have received little attention. Numerous undiscovered CREs and uncharacterized chromatin regions suggest an urgent need for a comprehensive and scalable method of multi-class CRE identification. Second, mainstream methods prevalently extract information from DNA sequences to distinguish CREs[17-19], overlooking the cell type specificity of CREs. Incorporating multi-omics data, including chromatin accessibility and chromatin interaction, for characterizing cell-type-specific CREs can provide valuable insights into gene regulatory mechanisms and cell heterogeneity. Third, deriving interpretable biological implications from conventional deep learning models remains challenging[32, 33], hindering the meaningful large-scale identification of CREs and the understanding of model-related biological variabilities of CREs.

To bridge these gaps, we propose CREATE (Cis-Regulatory Elements identificAtion via discreTe Embedding), a novel CNN-based supervised learning model that leverages the Vector Quantized Variational AutoEncoder (VQ-VAE)[34-36] framework. CREATE integrates genomic sequences with epigenetic features to offer a comprehensive approach for the identification and classification of multi-class CREs. The VQ-VAE framework is particularly suited for this task because it can distill genomic and epigenomic data into discrete CRE embeddings, capturing the nuanced differences between various CRE types. This discrete representation facilitates the generation of a CRE-specific feature spectrum, providing both quantitative and intuitive insights into CRE specificity. CREATE's ability to integrate multi-omics data enables it to overcome the limitations of previous methods by offering a more holistic view of CRE functionality and their cell-type-specific roles. Furthermore, CREATE demonstrates superior performance in accurately identifying CREs and exhibits robustness across diverse input types and hyperparameters. Its capability to perform large-scale predictions of CREs in specific cell types and to uncover disease- or phenotype-related biological variabilities in CREs underscores its potential as a powerful tool for constructing a comprehensive CRE atlas. In summary, CREATE represents a significant advancement in the computational identification of CREs,

and provides a robust foundation for future research in gene regulation and its implications for human health and disease.

Results

Overview of CREATE. CREATE is an advanced CNN model based on the VQ-VAE framework[34, 35] to predict and classify multi-class CREs from multi-omics data. Taking as input the one-hot encoded genomic sequence, the vector representing the chromatin accessibility scores for that sequence, and the vector representing the chromatin interaction scores for that sequence, CREATE is specifically crafted to capture discrete CRE embeddings, providing a comprehensive and interpretable characterization of CREs (Fig. 1a and Methods).

The architecture of CREATE includes (Fig. 1b and Methods): 1) Encoder Module: Each type of input data is initially processed by dedicated omics-specific encoders that transform the raw data into feature representations suitable for integration. Following this, the processed features are concatenated and passed through the integration encoder module, which synthesizes information from all input modalities to create a unified representation of the genomic context. 2) Vector Quantization Module: In this module, the output embeddings of encoder module are substituted with the closest counterpart in the discrete embedding space called "codebook". In brief, the features in the codebook are concatenated to form the final CRE embeddings. Unlike traditional VAE-based models[32] with fixed prior distributions, CREATE's codebook is dynamic and updated during training. This flexibility allows the model to refine the discrete embeddings to better represent the underlying biological data. 3) Decoder Module: The decoder reconstructs the original multi-omics input data from the discrete embeddings. It consists of two stages: the integration decoder reconstructs the integrated feature representation from the discrete embeddings. The omics-specific decoders transform the integrated representation back into the respective omics data types, ensuring that the reconstructed data aligns with the original input features. 4) Classifier: To enhance the model's ability to distinguish between different CRE types, CREATE includes a classifier that enforces the separation of CREs into distinct vectors in the codebook. CREs of the same type are encouraged to map to similar vectors, while those of different types are spread out across

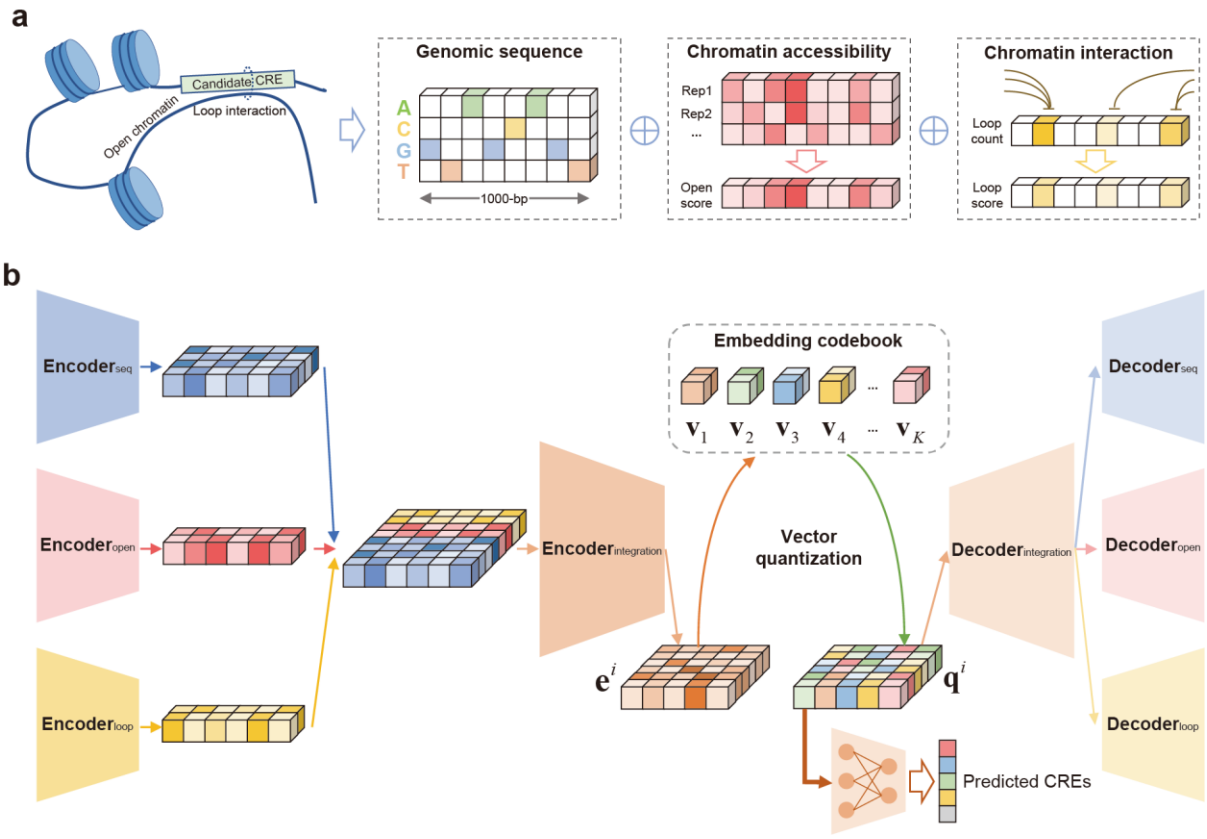

Fig. 1 | Overview of CREATE. a, The input of CREATE model. CREATE takes as input the genomic sequence, chromatin accessibility score and chromatin interaction score. **b**, The architecture of CREATE model. CREATE consists of encoders, a vector quantization module and decoders. The encoder module of CREATE combines encoders for multiple input-specific learning and an encoder for multiple input integration. For the *i*-th CRE, the encoder outputs the latent embedding $\mathbf{e}^i$ of dimension $L' \times D'$. By adopting split quantization, the latent embedding will be split into $L' \times M$ vectors $\mathbf{e}^i_{l,j}$ of dimension D and then quantized to $\mathbf{q}^i_{l,j}$ for the *i*-th CRE using embedding codebook with the size of K.

different vectors. This helps in achieving accurate and interpretable classifications.

CREATE offers several key advantages compared to existing methods: 1) Comprehensive Data Integration: By incorporating multiple omics data types, CREATE captures a more complete picture of the genomic context and CRE functionality. 2) Dynamic Codebook: The updateable codebook allows for flexible and accurate representation of CREs, overcoming limitations of fixed latent spaces in traditional VAE models. 3) Interpretable Embeddings: The discrete embeddings and their organization in the codebook provide clear and interpretable insights into CRE specificity and classification.

Overall, CREATE represents a significant advancement in computational CRE identification. Its ability to integrate multi-omics data, produce discrete embeddings, and offer interpretable results makes it a powerful tool for understanding gene regulation and its

implications in complex biological processes.

Cis-regulatory elements identification with CREATE. We comprehensively evaluated the performance of CREATE in identifying cell-type-specific CREs, including silencers, enhancers, promoters, insulators, and background regions, on the K562 and HepG2 cell types (Methods). To assess CREATE's effectiveness, we conducted 10-fold cross-validation experiments and compared its performance with four baseline methods, including DeepSEA[17], DanQ[18], ES-transition[19] and DeepICSH[20]. The primary evaluation metrics were area under the Receiver Operating Characteristic Curve (auROC), the area under the Precision-Recall Curve (auPRC) and the F1-score (Methods).

CREATE significantly surpasses the baseline methods by achieving the best classification performance on both K562 and HepG2 cell types (one-sided paired Wilcoxon signed-rank tests *P*-values < 1e-3), whereas the performance of baseline methods fluctuates across different cross-validation experiments (Fig. 2a-b and Supplementary Fig. 2a-b). For the K562 cell type, CREATE achieves a 10-fold macro-averaged auROC of 0.964 ± 0.002 (mean ± s.d.), outperforming the second-best method, ES-transition (0.928 ± 0.002) (Fig. 2c). Similarly, CREATE acquires a 10-fold macro-averaged auPRC of 0.848 ± 0.004, reflecting a substantial improvement of 10.5% compared to the second-best method, DeepICSH (0.743 ± 0.003) (Fig. 2d). A comparable performance enhancement is observed for the HepG2 cell type, with CREATE overtaking the baseline methods by a noticeable margin (Supplementary Fig. 2c-d).

Among the various CRE types, silencers and enhancers present unique challenges due to their similar epigenetic signatures[31]. Despite this, CREATE demonstrated a clear distinction between these difficult-to-differentiate elements. For the K562 cell type, CREATE achieves notable improvements in identifying silencers, with a mean auPRC that was 13.9% higher than the second-best method, which shows a greater advantage than the macro-averaged results of all CREs (Fig. 2f and Supplementary Fig. 2e). Similarly, CREATE attains a remarkable 22.1% improvement in mean auPRC for enhancers compared with the second-best method (Supplementary Fig. 3a-b). For other CRE types—promoters, insulators, and background regions—CREATE also demonstrates optimal classification performance, although baseline methods provide competitive results (Supplementary Fig. 3c-h). The performance trends for

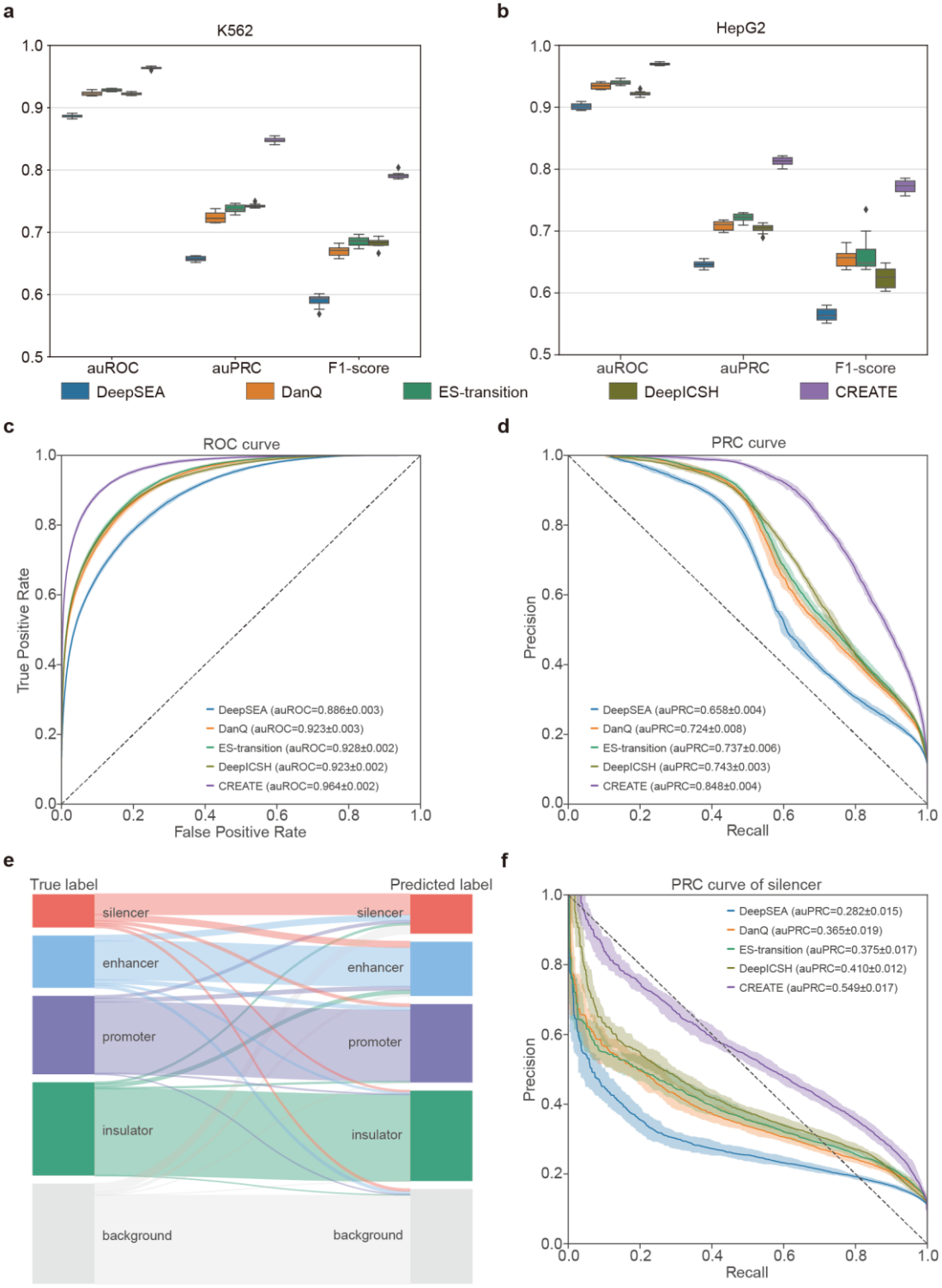

Fig. 2 | Evaluation of CREATE compared with the baseline methods. a-b, Boxplot of classification performance evaluated by auROC, auPRC and F1-score on the K562 cell type (**a**) and HepG2 cell type (**b**). Each box plot ranges from the upper to lower quartiles with the median as the horizontal line, whiskers extend to 1.5 times the interquartile range, and points represent outliers. **c-d**, Receiver Operating Characteristic curve (**c**) and Precision-Recall curve (**d**) comparing CREATE and baseline methods on the K562 cell type. **e**, The mapping between the true CRE labels and CREATE-predicted CRE labels on the testing data in one of the 10-fold cross-validation experiments of K562 cell type. **f**, Precision-Recall curve comparing CREATE and baseline methods for silencers in the K562 cell type. The mean and standard error of auROC or auPRC are reported in the legend. The confidence band shows ±1 s.d. for the averaged curve.

the HepG2 cell type mirrored those observed for the K562 cell type, further validating CREATE's robustness across different cell types (Supplementary Fig. 4).

The results highlight CREATE's exceptional capability in accurately identifying and distinguishing between various CRE types, particularly those less studied or less abundant, such as silencers and enhancers. CREATE's superior performance in capturing CRE variability and cell type specificity underscores its potential as a powerful tool for advancing our understanding of gene regulation mechanisms.

Robustness and effectiveness of CREATE. CREATE integrates genomic sequences, chromatin accessibility, and chromatin interactions to deliver a thorough characterization of gene regulatory processes. To assess the contributions of these different inputs, we conducted extensive ablation experiments. We referred to the models employing a single type of omics data as CREATE(seq), CREATE(open) and CREATE(loop), and those incorporating two different types of omics data as CREATE(seq+open), CREATE(seq+loop) and CREATE(open+loop), respectively. Among the seven models evaluated, CREATE consistently demonstrates the highest classification performance, confirming the importance of incorporating chromatin accessibility and chromatin interactions for superior CRE identification (Fig. 3a and Supplementary Fig. 5a-b). Specifically, CREATE shows substantial improvements in identifying challenging CRE types such as silencers and enhancers (Supplementary Fig. 5c-d). Notably, CREATE(seq), which relies solely on genomic sequences, achieves a 10-fold macro-averaged auPRC of 0.800 ± 0.004, surpassing baseline methods by 5.7% in mean auPRC (Supplementary Fig. 5b). This underscores CREATE's robust performance even when using genomic sequences alone. Incorporating additional omics data, such as chromatin accessibility or chromatin interactions, further enhances performance, though the inclusion of only these inputs without genomic sequences results in relatively poorer outcomes (Fig. 3a and Supplementary Fig. 5a). This indicates that while chromatin accessibility and interactions are valuable, genomic sequences are indispensable for optimal CRE identification, particularly contributing to better identification of silencers and enhancers.

Based on the VQ-VAE framework[34, 35], discrete embedding allows the latent space of CREATE to be a learnable discrete distribution, as opposed to the fixed Gaussian distribution

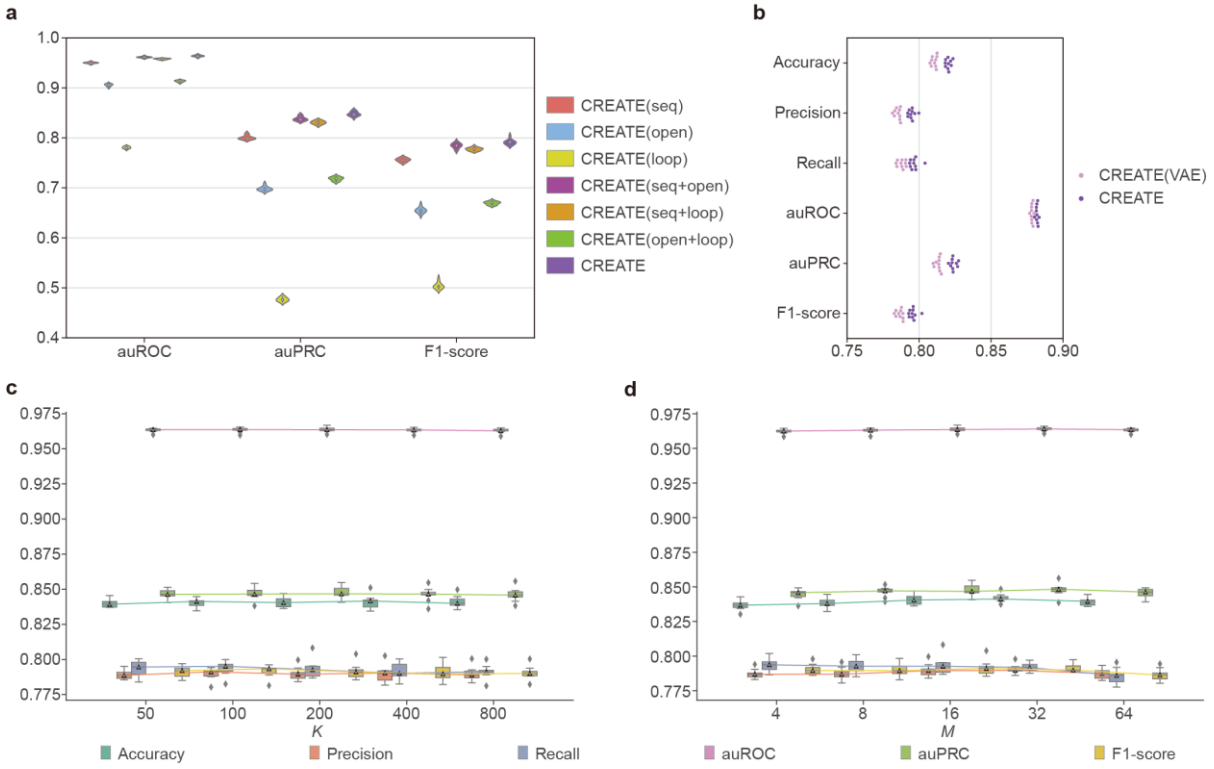

Fig. 3 | Robustness analysis of CREATE. a, Violin plot of classification performance evaluated by auROC, auPRC and F1-score for model ablation of CREATE on the K562 cell type. **b**, Swarm plot of classification performance evaluated by accuracy, precision, recall, auROC, auPRC and F1-score for CREATE compared with CREATE (VAE) on the K562 cell type. **c**, Classification performance of CREATE under different values of *K* (size of codebook) on the K562 cell type. **d**, Classification performance of CREATE under different values of *M* (time of split quantization) on the K562 cell type. Each box plot ranges from the upper to lower quartiles with the median as the horizontal line, whiskers extend to 1.5 times the interquartile range, and points represent outliers.

as in VAE models[32]. To verify the efficiency of discrete embedding in CREATE, we compared CREATE with a variant using VAE latent space (CREATE(VAE)) while keeping other modules and training strategies unchanged. CREATE significantly outperformed CREATE(VAE) across all evaluation metrics (one-sided paired Wilcoxon signed-rank tests *P*-values < 1e-3) (Fig. 3b). This result highlights the effectiveness of discrete embeddings in capturing complex CRE features.

To validate the stability and effectiveness of CREATE, we designed comprehensive robustness analyses for the hyperparameters in CREATE, including *K* denoting the size of codebook, *M* denoting the time of split quantization[32, 36], α denoting the weight of $L_{encoder}$, μ denoting the update ratio of codebook. First, to evaluate the robustness of CREATE to the size of codebook, we trained CREATE with different values of *K* (50, 100, 200, 400 and 800) on the K562 cell type. CREATE exhibited consistent classification performance across these

values, demonstrating its insensitivity to codebook size variations (Fig. 3c). Taking into account the balance of CRE specificity preservation and codebook utilization, we set the default value of K to 200. Second, to evaluate the stability of CREATE to the time of split quantization, we trained CREATE with different values of M (4, 8, 16, 32 and 64) on the K562 cell type. The results show that CREATE attains highly stable classification performance across different values of M (Fig. 3d). Evidently, the lower the time of split quantization, the higher the dimension of codebook features. With the consideration that it is obviously challenging to look up the nearest neighbors for high-dimensional vectors, we set the default value of M to 16. Third, following the original studies of VQ-VAE, we aimed for the codebook to have less impact on the output of encoder so that we set the default value of α, the weight of $L_{encoder}$, to 0.25. To validate the robustness of CREATE with different weights of $L_{encoder}$, we trained CREATE with different values of α (0.05, 0.1, 0.25, 0.5 and 1.0) on the K562 cell type. The results demonstrate that CREATE consistently obtains stable classification performance under different values of α (Supplementary Fig. 5e). Fourth, similar to the original studies of VQ-VAE, we set the default value of μ, the update ratio of codebook, to 0.01. To assess the stability of CREATE with different update ratios, we trained CREATE under a series of μ, 0.001, 0.005, 0.01, 0.05 and 0.1, on the K562 cell type. The results demonstrate the stability of the classification performance under different values of μ (Supplementary Fig. 5f). To summarize, the effective integration of multiple omics inputs, stable hyperparameters, and efficient discrete embedding all demonstrate the robustness and usability of CREATE.

Feature spectrum for unveiling CRE specificity. Discrete latent embedding of CREATE can reveal biological insights in an interpretable and intuitive manner. Using the latent embeddings of CREs, we built a uniform manifold approximation and projection (UMAP)[37] plot (Fig. 4a). Clearly, promoters, insulators and background regions are effectively separated, while there is some degree of overlap between silencers and enhancers, which is consistent with the classification results. To further validate the capability of CREATE in quantitatively articulating CRE specificity, we obtained specific feature spectrum for each type of CRE (Supplementary Fig. 6a and Methods). Briefly, each element in the CRE-specific feature spectrum represents the probability of a codebook feature occurring in that particular CRE

embeddings. We can always discover a set of particular features that are uniquely associated with a specific CRE and have the highest probability scores on that CRE, and we refer to these features as CRE-specific features. Concretely, for the K562 cell type, there are specific features uniquely enriched in the feature spectrum of each CRE (Fig. 4b). For example, we definitely observe different sets of specific features corresponding to promoters, insulators and background regions, which are clearly separated in the UMAP visualization (Fig. 4a) and Sankey diagram (Fig. 2e) as well. For the most difficult-to-distinguish two types of CREs, the feature spectrum of silencers contains a set of features (to the left of the blue dashed line) with notably higher probability scores compared to their scores in the feature spectrum of enhancers. Similarly, there is a set of features (between the blue dashed line and the purple dashed line) with notably higher probability scores in the feature spectrum of enhancers than those of silencers. In short, there is a relatively clear difference between the feature spectra of silencers and enhancers while they are connected together in the UMAP visualization. A similar result also occurred on the HepG2 cell type (Supplementary Fig. 6b-c). The CRE-specific feature spectrum, derived from discrete latent embedding of CREATE, has the potential to depict the general and comprehensive patterns of a type of CRE, further unveiling the CRE specificity quantitatively and interpretably.

To demonstrate the potential of codebook features in the CRE-specific feature spectrum for capturing key biological patterns, we identified the codebook feature with the highest probability score in the CRE-specific feature spectrum of K562 cell type as the major feature of that CRE, and we then zeroed it out before passing the CRE embeddings through the decoder again to generate the reconstructed output. To better understand the relationship between multi-omics input and the major feature of silencers, we designed comparative experiments between the original and reconstructed genomic sequences, chromatin accessibility scores and chromatin interaction scores. First, we conducted motif enrichment analysis for the original and reconstructed silencers (Methods). It is worth noting that the motif enrichment significance ($-\log_{10}P$-value) of MAFA, LHX6 and PAX8, which were reported as repressors in the previous literature[38-40], is obviously higher in the original sequences compared to the reconstructed sequences (one-sided Wilcoxon rank-sum tests P-values < 7e-53) (Fig. 4c and Supplementary Fig. 6d-e), whereas similar comparison results were not observed for known activators, such

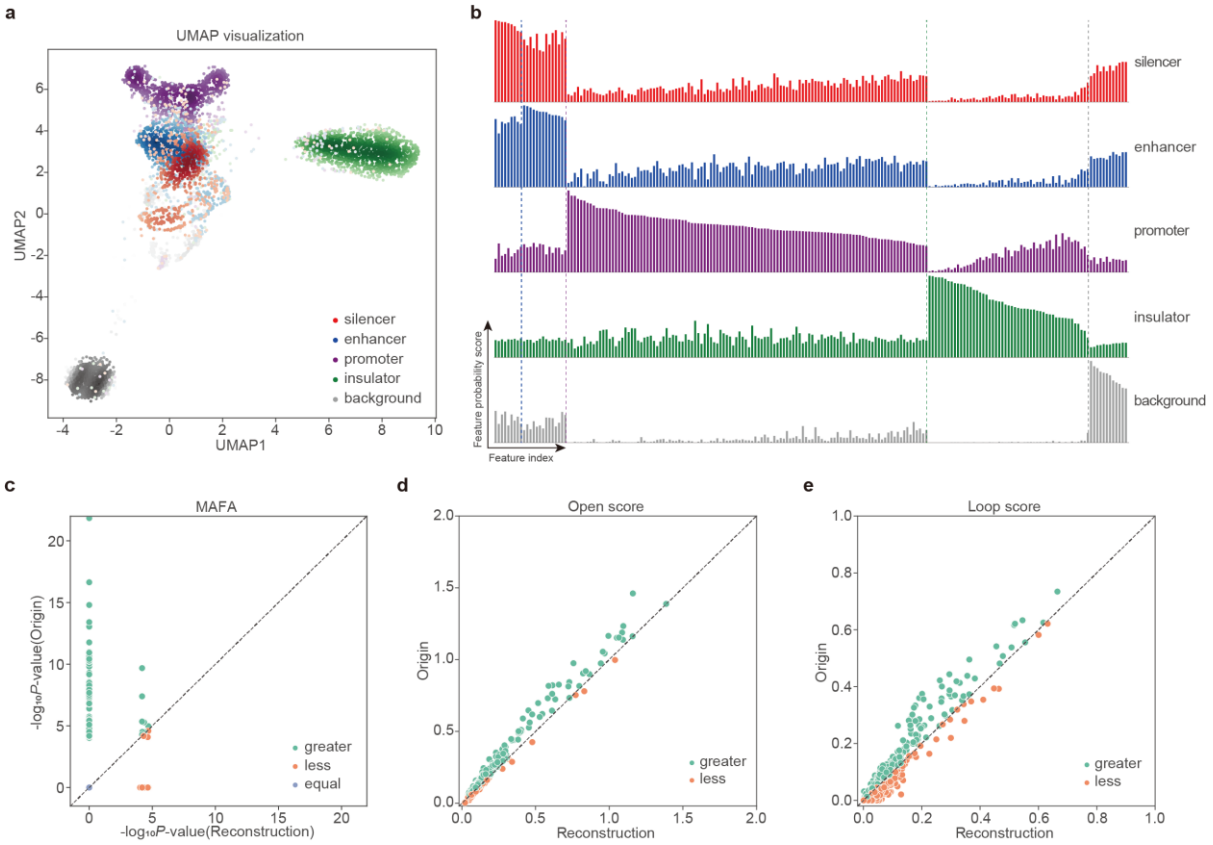

Fig. 4 | Generation and interpretation of CRE-specific feature spectrum. a, UMAP visualization of the CRE embeddings from CREATE on the testing data in one of the 10-fold cross-validation experiments of K562 cell type. **b**, CRE-specific feature spectrum. There are a distinct set of specific features that are enriched or depleted in the feature spectrum of each CRE on the K562 cell type. **c**, Comparison of MAFA motif enrichment significance (-$\log_{10}$*P*-value) between original input and reconstructed output when information derived from the major feature in the silencer-specific feature spectrum of K562 cell type is removed by zeroing it out before passing the CRE embeddings again through the decoder. **d**, Comparison of open scores between original input and reconstructed output when information derived from the major feature in the silencer-specific feature spectrum of K562 cell type is removed. **e**, Comparison of loop scores between original input and reconstructed output when information derived from the major feature in the silencer-specific feature spectrum of K562 cell type is removed.

as POU6F1[41] and MYC[42] (Supplementary Fig. 6f-g). This also demonstrates that the identified major feature of silencer-specific feature spectrum plays a crucial role in distinguishing between silencers and enhancers, as it indeed captures the motif information of some repressors, aligning with the repressive function of silencers. Simultaneously, TFs with the most significant difference between the original and reconstructed sequences, such as PRDM4, ZNF582 and SCRT2 (one-sided Wilcoxon rank-sum tests *P*-values < 6e-79) (Supplementary Fig. 7a-c), are considered to be novel silencer-related TFs. PRDM4 has been linked with recruiting chromatin modifiers, suggesting its involvement in establishing repressive chromatin states[43]. ZNF582 has been implicated in DNA methylation processes, which are crucial for

maintaining silencer function[44]. SCRT2 is less characterized, but its differential binding indicates a possible regulatory role in silencing mechanisms[45]. Similarly, the CRE-specific motif information is also harbored in the major feature of enhancers, promoters and insulators (Supplementary Fig. 7d-i), demonstrating that these features catch CRE-specific sequence patterns. Additionally, the unique motif patterns associated with silencers compared to enhancers, promoters, and insulators provide further evidence that these elements are distinct regulatory modules with specific TF associations. This distinction underscores the importance of considering a broader scope of CREs, including dual-function regulatory elements that might act as silencers under certain conditions and enhancers under others. Furthermore, the zeroing operation led to a reduction in the reconstructed chromatin accessibility scores and chromatin interaction scores (Fig. 4d-e), indicating that the major feature also captures silencer-specific epigenomic characteristics. Through the extensive comparative experiments that we designed, the CRE-specific feature spectrum generated by CREATE interpretably reveals the CRE specificity and is potentially involved in the gene regulation process in specific cell types. In conclusion, CREATE not only identifies known regulatory elements but also sheds light on less understood elements like silencers, filling a critical gap in the current landscape of gene regulation studies.

Large-scale prediction of cis-regulatory elements. The emergence of extensive epigenomic sequencing data across various cell types has enabled us to leverage a wealth of information for identifying cell-type-specific CREs on a large scale and establishing regulatory elements maps. CREATE proves to be a powerful tool for a comprehensive characterization of gene regulatory processes, revealing CREs with high accuracy and interpretability. In our study, we collected 270,259 candidate CREs on the K562 cell type and 232,456 candidate CREs on the HepG2 cell type for large-scale prediction (Supplementary Table 1 and Methods). For each cross-validation experiment, based on the trained CREATE model, we calculated a cutoff score for each type of CRE according to the validation set with a false positive rate (FPR) not exceeding 0.01. Candidate regions in Supplementary Table 1 exceeding the silencer cutoff score are marked as predicted silencers, and other CREs are labeled similarly. This approach led to the identification of 26,012 predicted silencers, 29,423 predicted enhancers, 2,057

predicted promoters, and 10,558 predicted insulators in the K562 cell type using the models trained on the K562 cell type, and the remaining sequences were classified as background regions. Similarly, in the HepG2 cell type, we identified 16,000 predicted silencers, 49,145 predicted enhancers, 4,422 predicted promoters and 13,122 predicted insulators using the models trained on the HepG2 cell type. The predicted CREs showed strong correlations with known epigenomic markers. For example, H3K27me3, a key histone modification associated with silencers[46-48], was found at higher proportions in predicted silencers (9.0% in K562 and 16.0% in HepG2) compared to other CREs (average 2.2% in K562 and average 4.1% in HepG2) (highlighted with red dashed box; Fig. 5a and Supplementary Fig. 8a). Similarly, histone modifications associated with enhancers, such as H3K9ac[49, 50], H3K27ac[50-52], H3K4me1[50, 53, 54], H3K4me2[54] and H3K4me3[54, 55], were more prevalent in predicted enhancers compared to other CRE types (average 18.5% in K562 and average 36.6% in HepG2) (highlighted with blue dashed box; Fig. 5a and Supplementary Fig. 8a).

To further validate the epigenomic characteristics of predicted CREs on a large scale, we conducted extensive comparative analyses. First, TFs play a crucial regulatory role in gene transcription by binding to CREs, and sequence-specific TF motifs can be considered key factors in identifying CREs[56-58]. We performed motif enrichment analysis on both true CREs (experimentally validated CREs) and predicted CREs (Methods). Compared to other true CREs and background regions, true silencers are enriched with the binding motifs of repressive TFs previously reported in the literature (Fig. 5b), such as FOXD1[59], TFAP2A[60], MAFA[38], MAFB[61], LHX6[39], PAX8[40], NFIA[62] and PRDM6[63], which are also enriched in the predicted silencers (Fig. 5c). Motifs belonging to active TFs, including POU6F1[41], MYC[42], ZFHX3[64] and SOX8[65], are enriched consistently across the true and predicted enhancers (Supplementary Fig. 8b-c). Notably, silencers and enhancers, the two most similar types of CREs, are enriched with the same TFs, such as MYC and ZFHX3. These TFs have been validated to act as either activators or repressors[42, 64, 66, 67], which aligns with the potential conversion between silencers and enhancers under different conditions[19, 68]. Similar motif enrichment results, consistent between true and predicted CREs, have also been observed in promoters (Supplementary Fig. 8d-e). These results indicate that CREATE effectively captured the CRE-specific sequence characteristics.

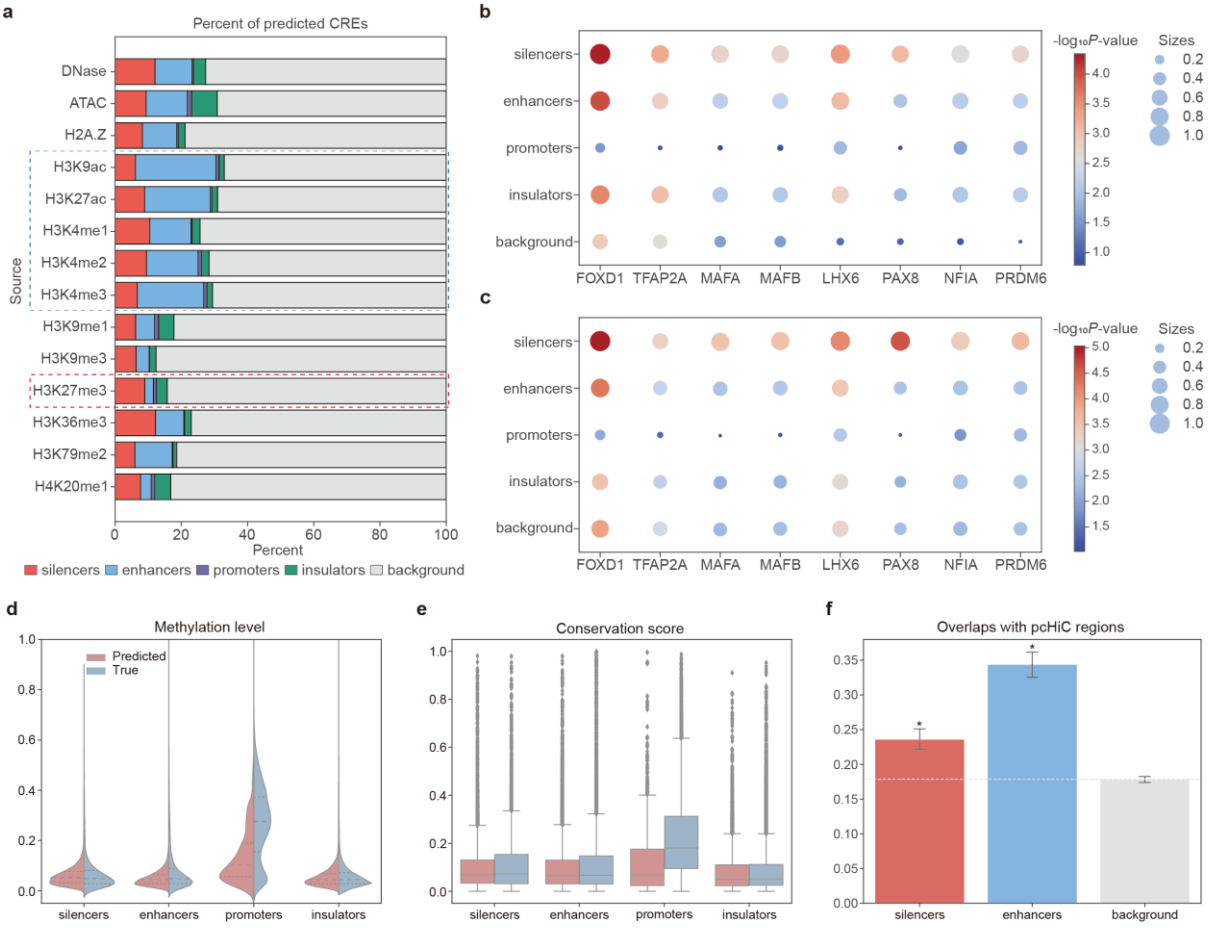

Fig. 5 | Characteristics of predicted CREs by CREATE. a, Percentage of predicted CREs and background regions from different candidate sources in the K562 cell type. **b-c**, Bubble plot of motif enrichment significance (-$\log_{10}$*P*-value) of repressive TFs (silencer-related TFs) at true CREs (**b**) and predicted CREs (**c**) on the K562 cell type. The size of bubbles represents the proportion of CREs with *P*-value < 0.01. **d**, Violin plot of methylation levels at true CREs and predicted CREs on the K562 cell type. Each violin plot contains three horizontal dashed lines denoting the median, the upper quartile, and the lower quartile. **e**, Box plot of conservation scores at true CREs and predicted CREs on the K562 cell type. Each box plot ranges from the upper to lower quartiles with the median as the horizontal line, whiskers extend to 1.5 times the interquartile range, and points represent outliers. **f**, Bar plot of overlaps between the pcHiC regions and predicted silencers, enhancers or background regions on the K562 cell type. The asterisks above the bars indicate the significant enrichments compared with the background regions. (∗) *P*-value < 2e-6. The error bars denote the 95% confidence interval, and the centers of error bars denote the average value.

Second, DNA methylation is an important epigenetic modification involved in gene regulation, particularly gene silencing[69, 70]. We calculated the methylation levels for both true CREs and predicted CREs (Methods), and observed the consistency between them except for promoters (Fig. 5d), which may be due to the complete collection of experimentally validated promoters and the limited number of predicted promoters. Specifically, 65.1% of the predicted promoters are adjacent to the promoters of non-coding genes, which is much higher than the

9.6% of randomly sampled genomic regions. In addition, the methylation levels of predicted CREs are significantly higher than those of predicted background regions (one-sided Wilcoxon rank-sum tests *P*-values < 2e-6) (Supplementary Fig. 8f).

Third, CREs are usually conserved in the evolutionary process of vertebrates, and the conserved regions are essential for deciphering the landscapes of gene regulation[71-73]. We computed the phastCons scores for the true CREs and the predicted CREs (Methods), and noticed that the conservation scores exhibit strong consistency between them except for promoters (Fig. 5e). Compared with the predicted background regions, the predicted CREs except insulators are significantly more conserved (one-sided Wilcoxon rank-sum tests *P*-values < 1e-13) (Supplementary Fig. 8g).

Fourth, CREs frequently regulate gene expression by connecting promoters through chromatin loops, which can be identified by promoter-capture HiC (pcHiC)[74]. We counted the number of overlaps between the pcHiC regions and the true CREs or the predicted CREs (Methods), and perceived the predicted silencers and enhancers harbor significantly more overlaps with pcHiC regions than the predicted background regions (one-sided Wilcoxon rank-sum tests *P*-values < 2e-6) (Fig. 5f), which aligns with the functional roles of these CREs in influencing target genes.

We further tested CREATE's cross-cell type prediction capabilities by using models trained on the K562 cell type to predict CREs in the HepG2 cell type. The results (auROC of 0.964 ± 0.002 and auPRC of 0.792 ± 0.005) demonstrate that CREATE maintains high performance across different cell types, confirming its robustness and generalizability (Supplementary Fig. 2c-d). Collectively, CREATE precisely extracts the CRE-specific epigenomic characteristics, enabling the construction of a comprehensive CRE atlas.

Characterization of dual-function regulatory elements. Dual-function regulatory elements (DFREs) are regions that exhibit dual roles as either silencers or enhancers depending on the cellular context[9, 19, 68, 75]. Understanding DFREs is crucial for unraveling the complexity of gene regulation, as these elements can significantly impact gene expression by switching functions based on the cellular environment. A K562 silencer overlapping a HepG2 enhancer

by more than 600 bp was considered a DFRE, resulting in 2,409 DFREs (9.3% of all predicted silencers) and 23,603 normal silencers. Conversely, we identified 36,448 HepG2 enhancers that do not function as enhancers or silencers in the K562 cell type, categorizing them as normal enhancers. Reasonably, the ability of CREATE to effectively characterize DFREs is demonstrated by its capacity to assign higher CREATE enhancer scores to DFREs than normal silencers (one-sided Wilcoxon rank-sum test *P*-value < 6e-47) (Fig. 6a), and higher CREATE silencer scores than normal enhancers (one-sided Wilcoxon rank-sum test *P*-value < 2e-5) (Fig. 6b). This differentiation underscores CREATE's effectiveness in distinguishing between multifunctional and context-specific regulatory elements.

To further explore the biological significance of DFREs, we conducted several comparative analyses. First, DFREs exhibit the highest conservation scores (one-sided Wilcoxon rank-sum tests *P*-values < 4e-3) and the background regions have the lowest conservation scores (one-sided Wilcoxon rank-sum tests *P*-values < 2e-31) (Fig. 6c), suggesting that these elements are evolutionarily preserved due to their critical roles in gene regulation. This high conservation underscores their functional importance across species and reinforces the value of identifying these elements for understanding gene regulatory mechanisms. Second, DFREs possess higher methylation levels in the K562 cell type compared to normal silencers (one-sided Wilcoxon rank-sum test *P*-value < 7e-4) (Fig. 6d). This observation highlights the unique epigenomic signatures of DFREs, suggesting that their dual functionality is associated with distinct methylation patterns, which may influence their regulatory roles. Third, DFREs show a strong preference with more overlaps with pcHiC regions of K562 cell type than normal silencers (Fig. 6e). This indicates that DFREs are actively involved in chromatin looping interactions, which are critical for mediating gene expression and regulatory network organization. Fourth, we computed the number of overlaps with expression quantitative trait loci (eQTLs) from whole-blood or liver tissues from GTEx[76, 77] for DFREs, normal silencers, normal enhancers and the background regions (Methods). DFREs are significantly enriched with more whole-blood eQTLs than normal silencers (one-sided Wilcoxon rank-sum test *P*-value < 3e-5) (Fig. 6f), and more liver eQTLs than normal enhancers (one-sided Wilcoxon rank-sum test *P*-value < 2e-3) (Fig. 6g). This enrichment demonstrates the tissue-specific regulatory potential of DFREs, highlighting their role in fine-tuning gene expression across different biological contexts.

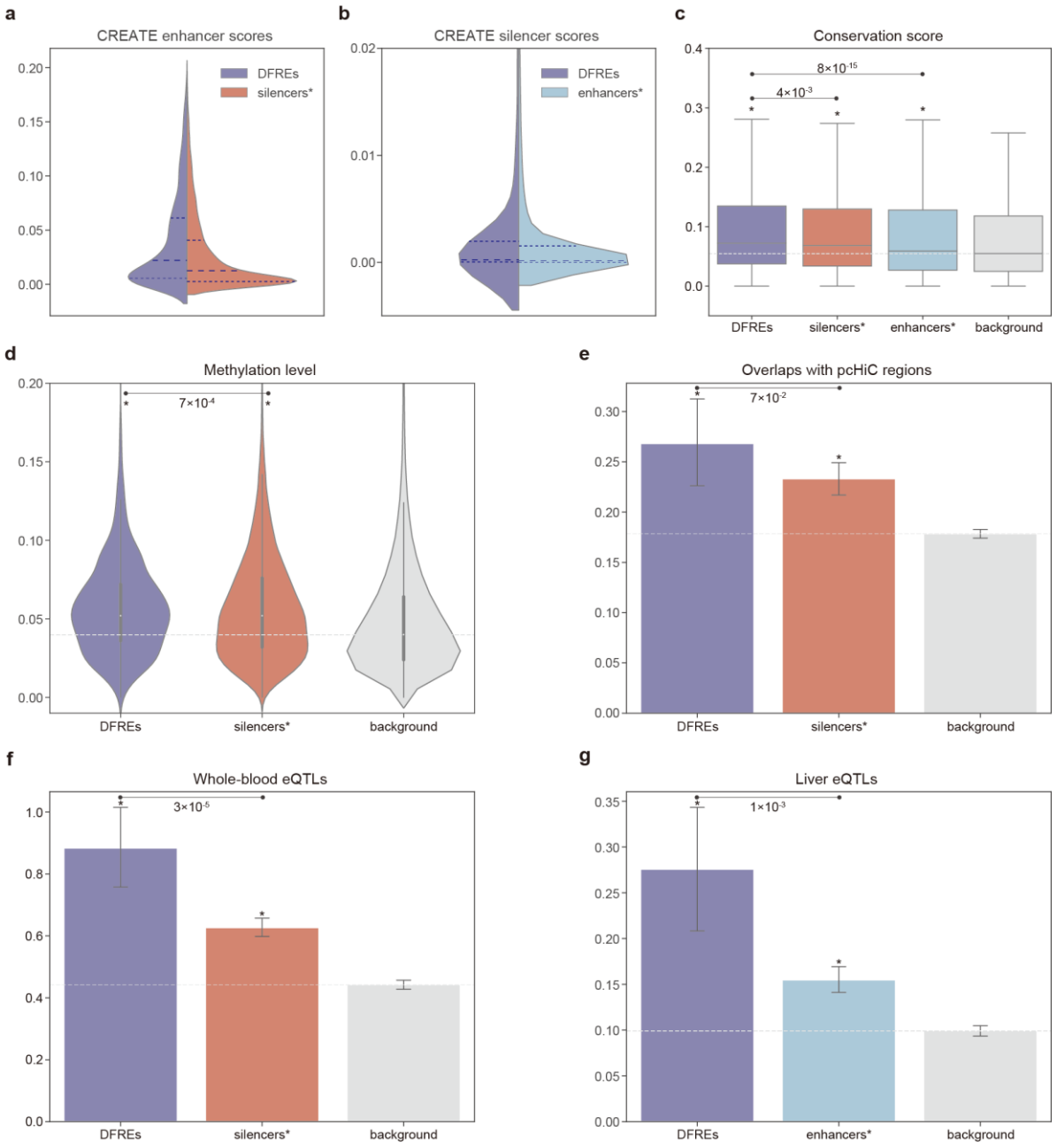

Fig. 6 | Characterization of DFREs identified by CREATE. a, Violin plot of the CREATE enhancer scores for DFREs and normal silencers (silencers*). **b**, Violin plot of the CREATE silencer scores for DFREs and normal enhancers (enhancers*). Each violin plot contains three horizontal dashed lines denoting the median, the upper quartile, and the lower quartile. **c**, Box plot of conservation scores at DFREs, normal silencers (silencers*), normal enhancers (enhancers*) and the background regions. The asterisks above the boxes indicate the significant enrichments compared with the background regions. (∗) *P*-value < 2e-31. Each box plot ranges from the upper to lower quartiles with the median as the horizontal line, whiskers extend to 1.5 times the interquartile range. **d**, Violin plot of methylation levels at DFREs, normal silencers (silencers*) and the background regions. (∗) *P*-value < 2e-6. **e**, Bar plot of overlaps between the pcHiC regions and DFREs, normal silencers (silencers*) or the background regions. (∗) *P*-value < 2e-3. **f**, Bar plot of overlaps between the whole-blood eQTLs and DFREs, normal silencers (silencers*) or the background regions. (∗) *P*-value < 5e-21. **g**, Bar plot of overlaps between the liver eQTLs and DFREs, normal enhancers (enhancers*) and the background regions. (∗) *P*-value < 2e-6. The error bars denote the 95% confidence interval, and the centers of error bars denote the average value.

Ultimately, CREATE's advanced capability to differentiate and interpret DFREs enriches the field's ability to map regulatory landscapes and uncover the underlying mechanisms of gene regulation.

Disease-associated variations analysis and tissue-specific enrichments in CREs. CREs play crucial roles in disease susceptibility and phenotype variations, often harboring single-nucleotide polymorphisms (SNPs) and eQTLs associated with various diseases and traits[78, 79]. To highlight the capacity of CREATE in uncovering disease-relevant variations within CREs, we analyzed overlaps with SNPs from dbSNP[80-82] database and eQTLs from GTEx[76, 77] for both the true CREs and the predicted CREs (Methods). Our results reveal that the gene variation distributions at predicted CREs align closely with those at true CREs (Fig. 7a and Supplementary Fig. 9a-c). In detail, the predicted CREs are significantly enriched with more rare SNPs than the background regions (one-sided Wilcoxon rank-sum tests *P*-values < 2e-9) (Fig. 7b), whereas a similar significant result was not observed for common SNPs (Supplementary Fig. 9d), suggesting the functional importance of identified CREs. The enrichment of rare gene variants in CREs, especially silencers and enhancers, supports the notions that disease-associated variants are more frequently located in gene regulatory regions[9, 12, 83], and rare variants are more impactful in complex diseases compared to common gene variants[84, 85]. Similarly, compared to background regions, significant enrichment on silencers and enhancers also occurred with whole-blood eQTLs (one-sided Wilcoxon rank-sum tests *P*-values < 4e-49) (Supplementary Fig. 9e), but not with all eQTLs (Supplementary Fig. 9f), reinforcing their tissue specificity. Besides, gene variation levels for both true CREs and predicted CREs gradually decrease with an increase in CREATE background scores (Fig. 7c and Supplementary Fig. 10a,d-i) but increase with an increase in CREATE silencer scores (Supplementary Fig. 10b-c), explicating the ability of CREATE for quantifying the impact of gene variations.

To further quantitatively assess the power of CREATE in unlocking the CRE-specific sequence characteristics, we portrayed the correlation between the CREATE scores and the motif enrichment significance at true or predicted CREs (Fig. 7d-e and Supplementary Figs. 11-14). We recognized the strong positive correlation between the CREATE silencer scores and

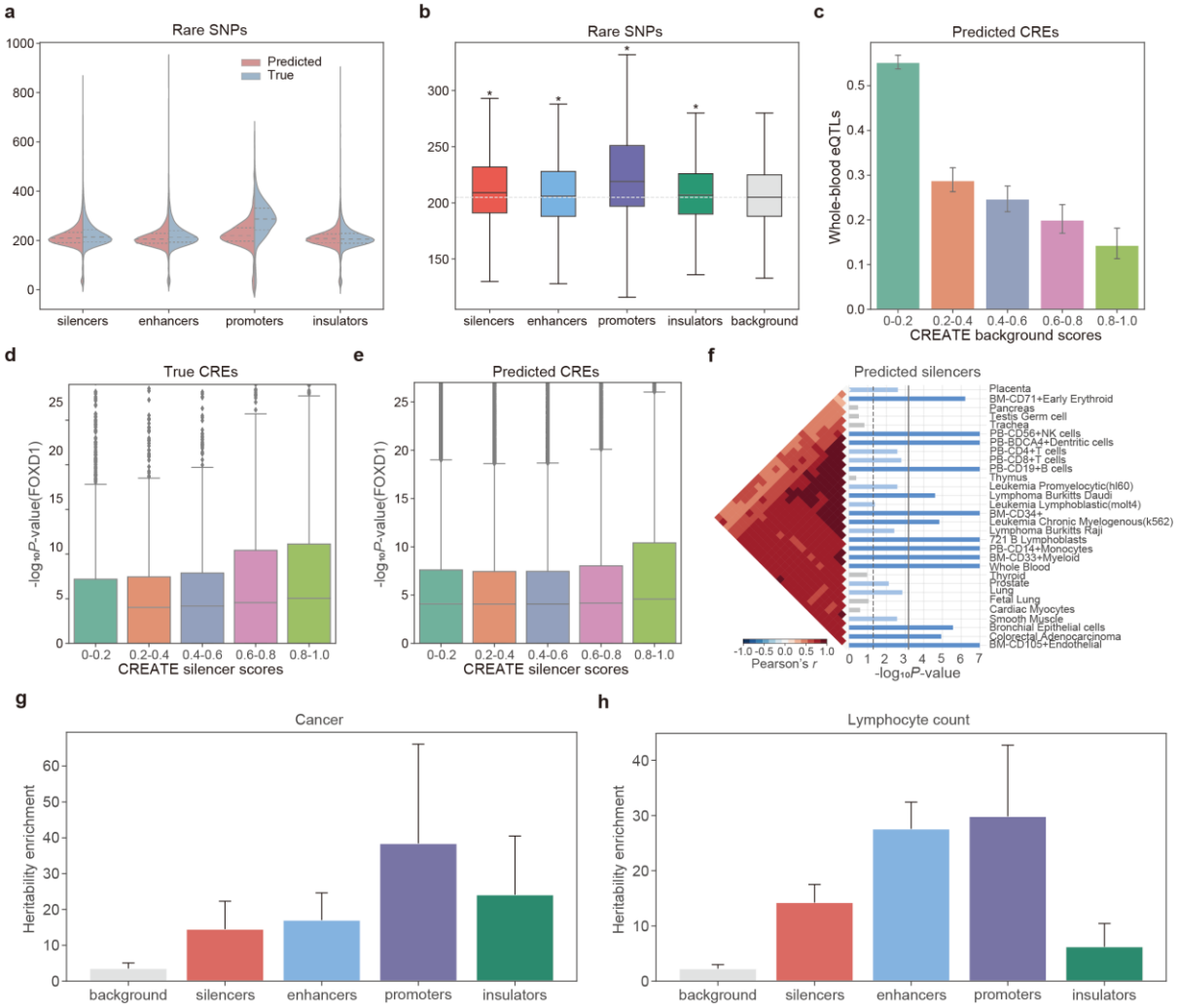

Fig. 7 | Identification of the biological variability of CREs by CREATE. a, Violin plot of overlaps between the rare SNPs and true CREs or predicted CREs on the K562 cell type. Each violin plot contains three horizontal dashed lines denoting the median, the upper quartile, and the lower quartile. **b**, Box plot of overlaps between the rare SNPs and the predicted CREs or background regions on the K562 cell type. The asterisks above the boxes indicate the significant enrichments compared with the background regions. (∗) *P*-value < 2e-9. Each box plot ranges from the upper to lower quartiles with the median as the horizontal line, whiskers extend to 1.5 times the interquartile range. **c**, Correlation between the CREATE background scores and overlaps between the whole-blood eQTLs and predicted CREs on the K562 cell type. **d-e**, Correlation between the CREATE silencer scores and the motif enrichment significance ($-\log_{10}$*P*-value) of FOXD1 at true CREs (**d**) and predicted CREs (**e**) on the K562 cell type. Each box plot ranges from the upper to lower quartiles with the median as the horizontal line, whiskers extend to 1.5 times the interquartile range, and points represent outliers. **f**, Top 30 significantly enriched tissues in SNPsea analysis on the predicted silencers of K562 cell type. The vertical dashed line represents the one-sided *P*-value cutoff at the 0.05 level, while the solid lines denotes the cutoff at 0.05 level for the one-sided *P*-value with Bonferroni correction. Each plot also contains the ordered expression profiles using hierarchical clustering with unweighted pair-group method with arithmetic means, and the Pearson correlation coefficients indicating the correlation between profiles. **g-h**, Heritability enrichments estimated by LDSC within predicted CREs and background regions identified by CREATE for blood-related traits including cancer (**g**) and lymphocyte count (**h**). The error bars denote jackknife standard errors over 200 equally sized blocks of adjacent SNPs about the estimates of enrichment, and the centers of error bars represent the average value.

the enrichment significance of silencer-related TFs (Fig. 5b-c), as well as between the CREATE enhancer scores and the enrichment significance of enhancer-related TFs (Supplementary Fig. 8b-c).

To illustrate the ability of CREATE for revealing the tissues influenced by the identified risk loci within the true CREs and the predicted CREs, we applied SNPsea[86] for tissue enrichment analysis (Methods). For the silencers and enhancers predicted by CREATE, we discovered more tissues related to blood than background regions (Fig. 7f and Supplementary Fig. 15), which aligns with the outcomes of true CREs (Supplementary Fig. 16). Concretely, Leukemia_Chronic_Myelogenous(k562) was identified as a significantly enriched tissue for the predicted silencers and enhancers in the K562 cell type (*P*-values < 2e-5), confirming the tissue specificity of CREs identified by CREATE.

To validate the competence of CREATE for studying the variations in phenotypes based on the true CREs and the predicted CREs, we utilized partitioned linkage disequilibrium score regression (LDSC)[87] for heritability enrichment analysis (Methods). Specifically, the enrichment of heritability in the predicted CREs is higher than that in the predicted background regions for the blood-related phenotypes, such as cancer, lymphocyte count and so on (Fig. 7g-h and Supplementary Fig. 17). Along with the enrichment results for the true CREs and background regions (Supplementary Fig. 18), CREs predicted by CREATE have inherited the pattern of heritability contribution for complex traits and diseases.

Altogether, CREATE excels in identifying CREs with significant disease-related variations and tissue-specific enrichments, providing critical insights into regulatory dynamics during development and disease progression. This capability underscores the power of CREATE in advancing our understanding of gene regulation and its implications for complex traits and diseases.

Discussion

In this study, we introduce CREATE, a groundbreaking multimodal architecture that integrates DNA sequences, cell-type-specific chromatin accessibility, and chromatin interaction features for the multi-class prediction of CREs. Utilizing discrete CRE embeddings, we have verified

the superior performance of CREATE in accurate CRE identification compared to state-of-the-art methods, as well as its effectiveness and stability across various input combinations, hyperparameters and forms of latent space. One of the key strengths of CREATE lies in its ability to offer improved interpretability as the CRE-specific feature spectrum, which quantitatively elucidates the CRE specificity and captures CRE-specific epigenomic characteristics. Moreover, CREATE has been validated the substantial potential in identifying cell-type-specific CREs on a large scale and uncovering biological variabilities of CREs, illustrating the ability of CREATE for unveiling the underlying regulatory dynamics that drive transcriptional regulation and disease development.

However, despite its successes, there are several areas where CREATE could be further improved. Our study identifies a few limitations and suggests several future directions for enhancing the method: 1) Data imbalance and insufficient research on certain CREs. The current data imbalance, particularly for silencers and certain cell types, impairs the overall performance of CRE identification. The obscure understanding of general silencer characteristics, the limited number of experimentally validated silencers and the restricted number of cell types studied, pose challenges in both model training and the selection of background regions. To address this, we plan to update our predicted silencers to the SilencerDB database[88] and expand our identification to include more cell types. We also anticipate that advances in biological technologies, such as HiChIP with a broader range of ChIPs, will enhance multi-class CRE identification and aid in constructing a more comprehensive regulatory atlas. 2) Per-base-paired input features and input-specific encoder-decoder structure. While the per-base-paired input features and the input-specific encoder-decoder structure are effective for extracting detailed and comprehensive CRE embeddings, some epigenetic features, such as TF binding, are not well-represented in this format. To improve scalability and representation, we propose integrating prior biological knowledge as additional constraints directly applied to the codebook. This approach is expected to enhance the model's ability of capturing complex epigenetic features and improve overall performance. 3) Development of a unified foundation model. Gene regulatory analysis methods typically focus on specialized models for specific problems. This paradigm limits the generalizability and integration of findings across different contexts. We aspire to develop a foundation model

for the unified characterization of key gene regulatory factors, leveraging the shareability, scalability and interpretability of the discrete embedding in CREATE. We anticipate that such a foundation model will facilitate a deeper understanding of gene regulation mechanisms and their implications for disease development, ultimately enabling biological discoveries and applications in developmental biology and precision medicine.

In conclusion, CREATE represents a significant advancement in the prediction and interpretation of CREs, offering superior performance and insights compared to existing methods. Its ability to integrate diverse data types and deliver interpretable results positions it as a valuable tool for exploring gene regulation and disease mechanisms. Future improvements and expansions of CREATE will continue to refine its capabilities and extend its applicability, driving forward our understanding of the complex interplay between gene regulation and disease.

Methods

Data collection and preprocessing. All datasets used in this study were publicly available and collected from different sources. We downloaded experimentally validated silencers for K562 and HepG2 cell types from the SilencerDB database[88]. We downloaded experimentally validated enhancers for K562 and HepG2 cell types from the FANTOM5 project[24, 26]. We obtained transcription start sites (TSSs) from the EPD database[89] and defined 1kb regions surrounding TSSs (500 bp upstream and 500 bp downstream) as promoters. Since CTCF characterized as an insulator by blocking chromatin interactions[90, 91], we took as insulators the CTCF Chromatin immunoprecipitation sequencing (ChIP-seq) peaks for K562 and HepG2 cell types collected from the ENCODE project[92, 93].

In addition, we collected multiple histone modification ChIP-seq peaks and chromatin accessibility peaks for K562 and HepG2 cell types from the Roadmap project[94] and ENCODE project (Supplementary Table 1). After filtering the regions overlapped with the experimentally validated CREs, known genes and consensus black list, we obtained 270,259 and 232,456 candidate CREs for large-scale prediction on the K562 and HepG2 cell types, respectively.

We randomly sampled DNA sequences from the entire human reference genome, excluding the experimentally validated and candidate CREs, known genes, consensus black list. After filtering overlapping regions between CREs, we obtained 6754 silencers, 10,528 enhancers, 15,699 promoters, 18,631 insulators and 20,000 background regions for the K562 cell type, and 1456 silencers, 11,407 enhancers, 14,535 promoters, 15,650 insulators and 20,000 background regions for the HepG2 cell type. The input for each CRE comprises three components: a one-hot encoded 1000-bp sequence from the human GRCh37/hg19 reference genome, a vector containing chromatin open scores per base pair, and another vector containing chromatin loop scores per base pair.

Chromatin open score. Chromatin accessibility is pivotal for identifying CREs, given that active regulatory DNA elements are typically situated in accessible chromatin regions[5, 6]. To incorporate the information of chromatin accessibility, we adopted OpenAnnotate[95] to efficiently calculate the raw read open scores of CREs and background regions per base pair.

We derived the chromatin open score per base pair by averaging the raw read open scores across replicates for each respective cell type.

Chromatin loop score. Chromatin looping interactions exert a substantial influence on gene regulation by establishing connections between regulatory elements and target genes[7, 8]. We incorporated cell-type-specific chromatin interaction data from HiChIP, which precisely profiles both regulatory and structural interactions[16, 96], to enhance the identification of CREs. We first calculated the number of chromatin loops per base pair for each CRE, and then obtained the chromatin loop score after logarithmic transformation.

Data augmentation. To ensure enough training samples for our model, we applied a data augmentation strategy to CREs[21, 22, 97]. As illustrated in Supplementary Fig. 1, for each CRE with length of 1000 base pairs, we shifted a window along the reference genome with a stride of 10 from the midpoint towards both ends. To mitigate the impact of data imbalance, we optionally incorporated data augmentation with varying augmentation ratios (5:5:3:3:1) for silencers, enhancers, promoters, insulators and background regions in the training data. Additionally, we augmented CREs by including the reverse complement of each original sequence. To prevent information leakage, the augmentation ratios for CREs in the validation and testing data are kept consistent at 5. Take the average of the predicted probabilities for all augmented sequences of the input sequence as the predicted probability for that input sequence.

The CREATE framework. We fed CREATE with a concatenated vector $\mathbf{X}^i \in \mathbb{R}^{6\times L}$ for the *i*-th input sample including a one-hot encoded genomic sequence $\mathbf{S}^i \in \mathbb{R}^{4\times L}$, a chromatin open score vector $\mathbf{O}^i \in \mathbb{R}^{1\times L}$ and a chromatin loop score vector $\mathbf{L}^i \in \mathbb{R}^{1\times L}$, where L is the length of sequence (L=1000). CREATE comprises encoders, a vector quantization module, and decoders. The encoder module of CREATE includes encoders for multiple input-specific learning and an encoder for integrating multiple inputs. Each encoder consists of a convolutional layer, a max-pooling layer, a ReLU non-linear activation function and a dropout layer. Correspondingly, each decoder consists of a deconvolutional layer, an upsample layer

and a Sigmoid or ReLU non-linear activation function. In addition, we introduced a classifier with three fully connected layers to predict CREs based on their embeddings. Specifically, the output of the encoder module is denoted as $\mathbf{e}^i \in \mathbb{R}^{L' \times D'}$ for the i-th CRE, where L' and D' are the length and dimensionality of the latent embedding respectively, and after split quantization[32, 36], it will be split into $L' \times M$ vectors $\mathbf{e}^i_{l,j} \in \mathbb{R}^D, l \in \{1, \dots, L'\}, j \in \{1, \dots, M\}$, where M is the time of split quantization. Utilizing a shared codebook $\mathbf{v}_k \in \mathbb{R}^D, k \in \{1, \dots, K\}$ with the size of K, we obtained the quantized latent embedding $\mathbf{q}^i \in \mathbb{R}^{L' \times D'}$ for the i-th CRE by substituting the vector $\mathbf{e}^i_{l,j}$ with the nearest counterpart in the codebook as follows:

$$\mathbf{q}^i_{l,j} = \mathbf{v}_{\underset{k \in \{1,\dots,K\}}{\operatorname{argmin}} \left\| \mathbf{e}^i_{l,j} - \mathbf{v}_k \right\|_2^2}, l \in \{1, \dots, L'\}, j \in \{1, \dots, M\}$$

Model training. We employed multiple update methods for different components of CREATE, mirroring the approach taken in the original studies of VQ-VAE[34, 35]. Let $\mathcal{B}_0$ be a mini-batch of data for training.

First, to optimize the decoder and encoder by reducing the distance between the original input and the reconstructed output, we integrated a hybrid reconstruction loss comprising multiple components corresponding to different inputs:

$$L_{recon1}(\mathcal{B}_0) = -\frac{1}{P \cdot L \cdot |\mathcal{B}_0|} \sum_{i=1}^{|\mathcal{B}_0|} \sum_{l=1}^{L} \sum_{p=1}^{P} \left[\mathbf{S}^i_{lp} \log\left(\hat{\mathbf{S}}^i_{lp}\right) + \left(1 - \mathbf{S}^i_{lp}\right) \log\left(1 - \hat{\mathbf{S}}^i_{lp}\right) \right]$$

$$L_{recon2}(\mathcal{B}_0) = \frac{1}{L \cdot |\mathcal{B}_0|} \sum_{i=1}^{|\mathcal{B}_0|} \sum_{l=1}^{L} \left\| \mathbf{O}^i_l - \hat{\mathbf{O}}^i_l \right\|_2^2$$

$$L_{recon3}(\mathcal{B}_0) = \frac{1}{L \cdot |\mathcal{B}_0|} \sum_{i=1}^{|\mathcal{B}_0|} \sum_{l=1}^{L} \left\| \mathbf{L}^i_l - \hat{\mathbf{L}}^i_l \right\|_2^2$$

$$L_{recon}(\mathcal{B}_0) = L_{recon1}(\mathcal{B}_0) + \beta L_{recon2}(\mathcal{B}_0) + \gamma L_{recon3}(\mathcal{B}_0)$$

where P represents the number of different types of bases in the DNA sequence (P=4), β and γ are the weights of L_{recon2} and L_{recon3} respectively ($\beta = 0.01, \gamma = 0.1$).

Second, to promote the encoder output to closely align with the selected codebook features and avoid excessive fluctuation, we introduced the encoder loss to aid in updating the encoder:

$$L_{encoder}(\mathcal{B}_0) = \frac{1}{M \cdot L' \cdot |\mathcal{B}_0|} \sum_{i=1}^{|\mathcal{B}_0|} \sum_{l=1}^{L'} \sum_{j=1}^{M} \left\| \mathbf{e}^i_{l,j} - \mathrm{sg}\left(\mathbf{v}_{\underset{k \in \{1,\dots,K\}}{\operatorname{argmin}} \left\| \mathbf{e}^i_{l,j} - \mathbf{v}_k \right\|_2^2} \right) \right\|_2^2$$

where sg(·) denotes the stop-gradient operator with zero partial derivatives.

Third, we followed the recommendation from both the original studies of VQ-VAE and recent related researches[98-101] to utilize exponential moving average (EMA) for updating the codebook. Considering n_k is the number of vectors matched to $\mathbf{v}_k$ and $\mathbf{e}^*_{k,m}$ is the m-th vector, we directly took the mean of the vectors in the set $\{\mathbf{e}^*_{k,m} | m = 1, \dots, n_k\}$ to optimize the code $\mathbf{v}_k$ as follows:

$$N_k^{(t)} = (1-\mu)N_k^{(t-1)} + \mu n_k^{(t)}$$

$$\mathbf{u}_k^{(t)} = (1-\mu)\mathbf{u}_k^{(t-1)} + \sum_{m=1}^{n_k^{(t)}} \mu \mathbf{e}_{k,m}^{*,(t)}$$

$$\mathbf{v}_k^{(t)} = \frac{\mathbf{u}_k^{(t)}}{N_k^{(t)}}$$

where μ is the update ratio of codebook. We initialized N_k as a zero vector and $\mathbf{u}_k$ randomly from a normal distribution with a mean of 0 and a standard deviation of 1.

Fourth, we further incorporated a classifier based on the CRE embeddings, with the cross-entropy loss function given by:

$$L_{class}(\mathcal{B}_0) = -\frac{1}{C \cdot |\mathcal{B}_0|} \sum_{i=1}^{|\mathcal{B}_0|} \sum_{c=1}^{C} y_c^i \log(\hat{y}_c^i)$$

where C represents the number of types of CREs (C=5). To sum up, we trained CREATE using EMA and the total loss function as follows:

$$L_{CREATE}(\mathcal{B}_0) = L_{recon}(\mathcal{B}_0) + \alpha L_{encoder}(\mathcal{B}_0) + L_{class}(\mathcal{B}_0)$$

where α is the weights of $L_{encoder}$.

In this study, we implemented CREATE with "Pytorch" package[102]. In details, there are three one-dimensional convolutional layers (filters=256,128,128; size=8,8,8) with layer normalization in the input-specific encoder module, followed by three one-dimensional convolutional layers (filters=512,384,128; size=1,8,8). In all cases, we set the mini-batch size to 1024 and employed the Adam stochastic optimization algorithm[103] with a learning rate of 5e-5. We trained CREATE with a maximum of 300 epochs and implemented early stopping if there were no reductions in validation auPRC for 20 consecutive epochs. We set the dimension of the latent embedding to 128 and trained CREATE with M of 16, K of 200, α of 0.25, and μ

of 0.01.

Model evaluation. To comprehensively evaluate the performance of CREATE for CRE identification, we conducted 10-fold cross-validation experiments by dividing all CREs into 8:1:1 ratios for training, validation and testing data, respectively. We evenly distributed each type of CRE into 10 folds. We compared the classification performance with four baseline methods including DeepSEA[17], DanQ[18], ES-transition[19] and DeepICSH[20], with the area under the Receiver Operating Characteristic Curve (auROC), the area under the Precision-Recall Curve (auPRC), F1-score, accuracy, precision and recall as evaluation metrics.

Feature spectrum. Supplementary Fig. 6a illustrates the process of generating the feature spectrum. For the *j*-th codebook feature, we counted its occurrence frequency in the latent embeddings of input regions, and we summed over these frequencies across all regions of the *i*-th CRE to gain the frequency c_{ij}. We next derived a probability matrix ($\mathbf{P} \in \mathbb{R}^{C\times K}$) by the following formula:

$$t_{ij} = \frac{c_{ij}}{\sum_{k\in\{1,\dots,K\}} c_{ik}}$$

$$p_{ij} = \frac{t_{ij}}{\sum_{c\in\{1,\dots,C\}} t_{cj}}$$

where *C* is the number of types of CREs and *K* is the number of codebook features. In this matrix, a row corresponds to a type of CRE and a column to a codebook feature, and an element p_{ij} indicates a feature probability score, representing the likelihood of the *j*-th codebook feature appearing in the latent embeddings of the *i*-th CRE. For the *j*-th codebook feature, we identified the element p_j with the highest feature probability score and the corresponding CRE C_j yielding the score, as follows.

$$p_j = \max_{i\in\{1,\dots,C\}} p_{ij}$$

$$C_j = \operatorname*{argmax}_{i\in\{1,\dots,C\}} p_{ij}$$

We then grouped the features corresponding to the same CRE together based on their CRE indices (C_j), and further sorted these features in descending order according to their feature

probability scores (p_j). Finally, we attained the rearranged matrix $\mathbf{F} \in \mathbb{R}^{C \times K}$ as the interpretable feature spectrum.

Downstream analyses.

Motif enrichment analysis. To discover enriched TF motifs for the true CREs and the predicted CREs by CREATE, we applied the tool FIMO[104] with default settings to scan a set of input sequences for searching known human TFs in the HOCOMOCO[105] database. For each input sequence, we used Fisher's method to combine the *P*-values of reported binding sites for each TF, and we obtained a *P*-value vector representing the significance that 678 human TFs matched in the input sequence.

Methylation levels computation. The methylation state data at CpG in the K562 cell type was obtained from ENCODE[106] (https://www.encodeproject.org/files/ENCFF867JRG/; https://www.encodeproject.org/files/ENCFF721JMB/). Using BEDTools[107], we computed the methylation levels for the true CREs and the predicted CREs by CREATE.

Conservation scores computation. We downloaded the phastCons[71, 108] scores for multiple alignments of 45 vertebrate genomes to the human genome from UCSC[109, 110] (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phastCons46way/vertebrate.phastCons46way.bw). The phastCons scores for the true CREs and the predicted CREs were calculated via UCSC tool *bigWigAverageOverBed*[111].

Overlaps with promoter-capture HiC regions. The pcHiC data of K562 cell type was downloaded from NCBI[112] under accession number "GSE236305". Using BEDTools[107], we computed the number of overlaps between the pcHiC regions and the true CREs or the predicted CREs by CREATE.

Gene variation analysis. We downloaded human all SNPs and common SNPs from dbSNP[80-82] database (https://ftp.ncbi.nih.gov/snp/organisms/human_9606_b151_GRCh37p13/VCF/), all GTEx eQTLs, whole-blood eQTLs and liver eQTLs from the Genotype-Tissue Expression Project[76, 77], GTEx database version 7 (https://www.gtexportal.org/home/downloads/adult-gtex). By excluding the common SNPs from all human SNPs, we obtained rare SNPs. Using BEDTools[107], we considered the number of overlaps with SNPs or eQTLs as the corresponding

gene variation levels for the true CREs or the predicted CREs by CREATE.

Tissue enrichment analysis. To identify the tissues influenced by the identified risk loci within the true CREs and the predicted CREs by CREATE, we performed SNPsea analysis[86] with default settings. Based on the tissue-specific expression profiles of 17,581 genes across 79 human tissues (Gene Atlas[113]), we quantified the enrichments of these profiles on the true CREs and the predicted CREs, and displayed the top 30 significantly enriched tissues in the heatmaps.

Heritability enrichment analysis. To quantify the enrichment of heritability for blood-related phenotypes within the true CREs and the predicted CREs by CREATE, we conducted heritability enrichment analysis using partitioned LDSC[87] with default settings. LDSC took European samples from the 1000 Genomes Project as the LD reference panel. We downloaded the HapMap3 SNPs and GWAS summary statistics from the Broad LD Hub (https://doi.org/10.5281/zenodo.7768714), and then quantified the enrichment of heritability for blood-related phenotypes, and displayed the results for the true CREs and the predicted CREs.

Baseline methods. In this study, we compared CREATE to multiple baseline methods by expanding them into multi-class models, including DeepSEA[17], DanQ[18], ES-transition[19] and DeepICSH[20]. DeepSEA was implemented from their original source code (https://deepsea.princeton.edu/). DanQ was implemented from their original source code repository (https://github.com/uci-cbcl/DanQ). ES-transition was implemented from their original source code repository (https://github.com/ncbi/SilencerEnhancerPredict). DeepICSH was implemented from their original source code repository (https://github.com/lyli1013/DeepICSH).

Statistics and reproducibility

No statistical method was used to predetermine sample size. No data were excluded from the analyses. The experiments were not randomized. Data collection and analysis were not performed blind to the conditions of the experiments.

Data availability

All datasets used in this study were obtained from public sources. We downloaded experimentally validated silencers from the SilencerDB database[88] (http://health.tsinghua.edu.cn/SilencerDB/), enhancers from the FANTOM5 project[24, 26] (https://bioinfo.vanderbilt.edu/AE/HACER/), TSSs from the EPD database[89] (https://epd.expasy.org/epd), insulators from the ENCODE project[92, 93] (https://www.encodeproject.org/files/ENCFF085HTY/; https://www.encodeproject.org/files/ENCFF237OKO/) for the K562 and HepG2 cell types. We downloaded the histone modification ChIP-seq peaks and chromatin accessibility peaks from the Roadmap project[94] (http://egg2.wustl.edu/roadmap/data/byFileType/peaks/consolidated/narrowPeak) and ENCODE project (https://www.encodeproject.org/files/ENCFF055NNT/; https://www.encodeproject.org/files/ENCFF333TAT/; https://www.encodeproject.org/files/ENCFF558BLC/; https://www.encodeproject.org/files/ENCFF842UZU/; https://www.encodeproject.org/files/ENCFF439EIO/; https://www.encodeproject.org/files/ENCFF913MQB/) for the K562 and HepG2 cell types. The non-coding genes were obtained from GENCODE[114] for human (GRCh37.p13/hg19). All regions in this study are either in the genome of GRCh37/hg19 or have been converted to GRCh37/hg19 by UCSC liftOver[115] tool.

Code availability

The CREATE software, including detailed documents and tutorial, is freely available on GitHub (https://github.com/cuixj19/CREATE).

Acknowledgements

This work was supported by the National Key Research and Development Program of China grant nos. 2021YFF1200902 (R.J.), 2023YFF1204802 (R.J.), the National Natural Science Foundation of China grants no. 62273194 (R.J.).

Author contributions

R.J. and W.Z. conceived the study and supervised the project. X.C. designed, implemented and validated CREATE. Q.Y., Z.G., Z.L., X.C., S.C. and Q.L. helped analyze the results. X.C. and W.Z. wrote the manuscript, with input from all the authors.

Ethics declarations

Competing interests

The authors declare no competing interests.

Reference

1. Berger, S.L. The complex language of chromatin regulation during transcription. *Nature* **447**, 407-412 (2007).
2. Lee, T.I. & Young, R.A. Transcriptional regulation and its misregulation in disease. *Cell* **152**, 1237-1251 (2013).
3. Maston, G.A., Evans, S.K. & Green, M.R. Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.* **7**, 29-59 (2006).
4. Chatterjee, S. & Ahituv, N. Gene regulatory elements, major drivers of human disease. *Annu. Rev. Genomics Hum. Genet.* **18**, 45-63 (2017).
5. Klemm, S.L., Shipony, Z. & Greenleaf, W.J. Chromatin accessibility and the regulatory epigenome. *Nat. Rev. Genet.* **20**, 207-220 (2019).
6. Minnoye, L. et al. Chromatin accessibility profiling methods. *Nat. Rev. Methods Primers* **1**, 10 (2021).
7. Kadauke, S. & Blobel, G.A. Chromatin loops in gene regulation. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms* **1789**, 17-25 (2009).
8. Dekker, J., Marti-Renom, M.A. & Mirny, L.A. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat. Rev. Genet.* **14**, 390-403 (2013).
9. Pang, B., van Weerd, J.H., Hamoen, F.L. & Snyder, M.P. Identification of non-coding silencer elements and their regulation of gene expression. *Nat. Rev. Mol. Cell Biol.* **24**, 383-395 (2023).
10. Sealfon, R.S., Wong, A.K. & Troyanskaya, O.G. Machine learning methods to model multicellular complexity and tissue specificity. *Nature Reviews Materials* **6**, 717-729 (2021).
11. Vaishnav, E.D. et al. The evolution, evolvability and engineering of gene regulatory DNA. *Nature* **603**, 455-463 (2022).
12. Huang, D., Petrykowska, H.M., Miller, B.F., Elnitski, L. & Ovcharenko, I. Identification of human silencers by correlating cross-tissue epigenetic profiles and gene expression. *Genome Res.* **29**, 657-667 (2019).
13. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-293 (2009).
14. Song, L. & Crawford, G.E. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb. Protoc.* **2010**, pdb. prot5384 (2010).
15. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y. & Greenleaf, W.J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213-1218 (2013).
16. Mumbach, M.R. et al. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat. Methods* **13**, 919-922 (2016).
17. Zhou, J. & Troyanskaya, O.G. Predicting effects of noncoding variants with deep learning–based sequence model. *Nat. Methods* **12**, 931-934 (2015).
18. Quang, D. & Xie, X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res.* **44**, e107-e107 (2016).
19. Huang, D. & Ovcharenko, I. Enhancer–silencer transitions in the human genome. *Genome Res.* **32**, 437-448 (2022).
20. Zhang, T., Li, L., Sun, H., Xu, D. & Wang, G. DeepICSH: a complex deep learning framework for identifying cell-specific silencers and their strength from the human genome. *Brief. Bioinform.* **24**, bbad316 (2023).
21. Chen, S., Gan, M., Lv, H. & Jiang, R. DeepCAPE: a deep convolutional neural network for the accurate

prediction of enhancers. *Genomics Proteomics Bioinformatics* **19**, 565-577 (2021).

22. Min, X. et al. Predicting enhancers with deep convolutional neural networks. *BMC Bioinformatics* **18**, 35-46 (2017).
23. Oubounyt, M., Louadi, Z., Tayara, H. & Chong, K.T. DeePromoter: robust promoter predictor using deep learning. *Front. Genet.* **10**, 453150 (2019).
24. Andersson, R. et al. An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455-461 (2014).
25. Khan, A. & Zhang, X. dbSUPER: a database of super-enhancers in mouse and human genome. *Nucleic Acids Res.* **44**, D164-D171 (2016).
26. Wang, J. et al. HACER: an atlas of human active enhancers to interpret regulatory variants. *Nucleic Acids Res.* **47**, D106-D112 (2019).
27. Zeng, W., Min, X. & Jiang, R. EnDisease: a manually curated database for enhancer-disease associations. *Database* **2019**, baz020 (2019).
28. Jiang, Y. et al. SEdb: a comprehensive human super-enhancer database. *Nucleic Acids Res.* **47**, D235-D243 (2019).
29. Bai, X. et al. ENdb: a manually curated database of experimentally supported enhancers for human and mouse. *Nucleic Acids Res.* **48**, D51-D57 (2020).
30. Gao, T. & Qian, J. EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. *Nucleic Acids Res.* **48**, D58-D64 (2020).
31. Ogbourne, S. & Antalis, T.M. Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes. *Biochem. J.* **331**, 1-14 (1998).
32. Cui, X. et al. Discrete latent embedding of single-cell chromatin accessibility sequencing data for uncovering cell heterogeneity. *Nat. Comput. Sci.*, 1-14 (2024).
33. Zheng, A. et al. Deep neural networks identify sequence context features predictive of transcription factor binding. *Nat. Mach. Intell.* **3**, 172-180 (2021).
34. Van Den Oord, A. & Vinyals, O. Neural discrete representation learning. In *Proc. Advances in Neural Information Processing Systems* **30** (2017).
35. Razavi, A., Van den Oord, A. & Vinyals, O. Generating diverse high-fidelity images with vq-vae-2. In *Adv. Neural Inf. Process. Syst.* **32** (2019).
36. Kobayashi, H., Cheveralls, K.C., Leonetti, M.D. & Royer, L.A. Self-supervised deep learning encodes high-resolution features of protein subcellular localization. *Nat. Methods* **19**, 995-1003 (2022).
37. McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).
38. Igarashi, K. et al. Multivalent DNA binding complex generated by small Maf and Bach1 as a possible biochemical basis for β-globin locus control region complex. *J. Biol. Chem.* **273**, 11783-11790 (1998).
39. Zhang, Z. et al. The LIM homeodomain transcription factor LHX6: a transcriptional repressor that interacts with pituitary homeobox 2 (PITX2) to regulate odontogenesis. *J. Biol. Chem.* **288**, 2485-2500 (2013).
40. Gu, X. et al. PBRM1 loss in kidney cancer unbalances the proximal tubule master transcription factor hub to repress proximal tubule differentiation. *Cell Rep.* **36** (2021).
41. Wang, J., Jia, Q., Jiang, S., Lu, W. & Ning, H. POU6F1 promotes ferroptosis by increasing lncRNA-CASC2 transcription to regulate SOCS2/SLC7A11 signaling in gastric cancer. *Cell Biol. Toxicol.* **40**, 1-17 (2024).
42. Cowling, V.H. & Cole, M.D. Mechanism of transcriptional activation by the Myc oncoproteins. In *Semin. Cancer Biol.* **16**, 242-252 (2006).
43. Chittka, A., Nitarska, J., Grazini, U. & Richardson, W.D. Transcription factor positive regulatory domain

4 (PRDM4) recruits protein arginine methyltransferase 5 (PRMT5) to mediate histone arginine methylation and control neural stem cell proliferation and differentiation. *J. Biol. Chem.* **287**, 42995-43006 (2012).

44. Li, N., He, Y., Mi, P. & Hu, Y. ZNF582 methylation as a potential biomarker to predict cervical intraepithelial neoplasia type III/worse: A meta-analysis of related studies in Chinese population. *Medicine* **98**, e14297 (2019).
45. Ha, T.J. et al. Identification of novel cerebellar developmental transcriptional regulators with motif activity analysis. *BMC Genomics* **20**, 1-17 (2019).
46. Della Rosa, M. & Spivakov, M. Silencers in the spotlight. *Nat. Genet.* **52**, 244-245 (2020).
47. Ngan, C.Y. et al. Chromatin interaction analyses elucidate the roles of PRC2-bound silencers in mouse development. *Nat. Genet.* **52**, 264-272 (2020).
48. Cai, Y. et al. H3K27me3-rich genomic regions can function as silencers to repress gene expression via chromatin interactions. *Nat. Commun.* **12**, 719 (2021).
49. Karmodiya, K., Krebs, A.R., Oulad-Abdelghani, M., Kimura, H. & Tora, L. H3K9 and H3K14 acetylation co-occur at many gene regulatory elements, while H3K14ac marks a subset of inactive inducible promoters in mouse embryonic stem cells. *BMC Genomics* **13**, 1-18 (2012).
50. Oka, R. et al. Genome-wide mapping of transcriptional enhancer candidates using DNA and chromatin features in maize. *Genome Biol.* **18**, 1-24 (2017).
51. Zhu, Y. et al. Predicting enhancer transcription and activity from chromatin modifications. *Nucleic Acids Res.* **41**, 10032-10043 (2013).
52. Bogdanović, O. et al. Dynamics of enhancer chromatin signatures mark the transition from pluripotency to cell specification during embryogenesis. *Genome Res.* **22**, 2043-2053 (2012).
53. Local, A. et al. Identification of H3K4me1-associated proteins at mammalian enhancers. *Nat. Genet.* **50**, 73-82 (2018).
54. Pekowska, A. et al. H3K4 tri‐methylation provides an epigenetic signature of active enhancers. *EMBO J.* **30**, 4198-4210 (2011).
55. Chen, K. et al. Broad H3K4me3 is associated with increased transcription elongation and enhancer activity at tumor-suppressor genes. *Nat. Genet.* **47**, 1149-1157 (2015).
56. Whitfield, T.W. et al. Functional analysis of transcription factor binding sites in human promoters. *Genome Biol.* **13**, 1-16 (2012).
57. Boeva, V. Analysis of genomic sequence motifs for deciphering transcription factor binding and transcriptional regulation in eukaryotic cells. *Front. Genet.* **7**, 174397 (2016).
58. Inukai, S., Kock, K.H. & Bulyk, M.L. Transcription factor–DNA binding: beyond binding site motifs. *Curr. Opin. Genet. Dev.* **43**, 110-119 (2017).
59. Polevoy, H., Malyarova, A., Fonar, Y., Elias, S. & Frank, D. FoxD1 protein interacts with Wnt and BMP signaling to differentially pattern mesoderm and neural tissue. *Int. J. Dev. Biol.* **61**, 293-302 (2017).
60. Scibetta, A.G., Wong, P.-P., Chan, K.V., Canosa, M. & Hurst, H.C. Dual association by TFAP2A during activation of the p21cip/CDKN1A promoter. *Cell Cycle* **9**, 4525-4532 (2010).
61. Sieweke, M.H., Tekotte, H., Frampton, J. & Graf, T. MafB is an interaction partner and repressor of Ets-1 that inhibits erythroid differentiation. *Cell* **85**, 49-60 (1996).
62. Piper, M. et al. NFIA controls telencephalic progenitor cell differentiation through repression of the Notch effector Hes1. *J. Neurosci.* **30**, 9127-9139 (2010).
63. Davis, C.A. et al. PRISM/PRDM6, a transcriptional repressor that promotes the proliferative gene program in smooth muscle cells. *Mol. Cell. Biol.* **26**, 2626-2636 (2006).
64. Parsons, M.J. et al. The regulatory factor ZFHX3 modifies circadian function in SCN via an AT motif-driven axis. *Cell* **162**, 607-621 (2015).

65. Schepers, G.E., Bullejos, M., Hosking, B.M. & Koopman, P. Cloning and characterisation of the Sry-related transcription factor gene Sox8. *Nucleic Acids Res.* **28**, 1473-1480 (2000).
66. Baluapuri, A., Wolf, E. & Eilers, M. Target gene-independent functions of MYC oncoproteins. *Nat. Rev. Mol. Cell Biol.* **21**, 255-267 (2020).
67. Rubio-Alarcón, M. et al. Zfhx3 transcription factor represses the expression of SCN5A gene and decreases sodium current density (INa). *Int. J. Mol. Sci.* **22**, 13031 (2021).
68. Doni Jayavelu, N., Jajodia, A., Mishra, A. & Hawkins, R.D. Candidate silencer elements for the human and mouse genomes. *Nat. Commun.* **11**, 1061 (2020).
69. Smith, Z.D. & Meissner, A. DNA methylation: roles in mammalian development. *Nat. Rev. Genet.* **14**, 204-220 (2013).
70. Moore, L.D., Le, T. & Fan, G. DNA methylation and its basic function. *Neuropsychopharmacology* **38**, 23-38 (2013).
71. Siepel, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034-1050 (2005).
72. Villar, D. et al. Enhancer evolution across 20 mammalian species. *Cell* **160**, 554-566 (2015).
73. Ovcharenko, I. et al. Evolution and functional classification of vertebrate gene deserts. *Genome Res.* **15**, 137-145 (2005).
74. Schoenfelder, S., Javierre, B.-M., Furlan-Magaril, M., Wingett, S.W. & Fraser, P. Promoter capture Hi-C: high-resolution, genome-wide profiling of promoter interactions. *J. Vis. Exp.*, e57320 (2018).
75. Gisselbrecht, S.S. et al. Transcriptional silencers in Drosophila serve a dual role as transcriptional enhancers in alternate cellular contexts. *Mol. Cell* **77**, 324-337. e328 (2020).
76. Lonsdale, J. et al. The genotype-tissue expression (GTEx) project. *Nat. Genet.* **45**, 580-585 (2013).
77. Consortium, G. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318-1330 (2020).
78. Maurano, M.T. et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190-1195 (2012).
79. Schaub, M.A., Boyle, A.P., Kundaje, A., Batzoglou, S. & Snyder, M. Linking disease associations with regulatory information in the human genome. *Genome Res.* **22**, 1748-1759 (2012).
80. Sherry, S.T., Ward, M. & Sirotkin, K. dbSNP—database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Res.* **9**, 677-679 (1999).
81. Sherry, S.T. et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308-311 (2001).
82. Smigielski, E.M., Sirotkin, K., Ward, M. & Sherry, S.T. dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res.* **28**, 352-355 (2000).
83. Hawkins, R.D. et al. Global chromatin state analysis reveals lineage-specific enhancers during the initiation of human T helper 1 and T helper 2 cell polarization. *Immunity* **38**, 1271-1284 (2013).
84. Raychaudhuri, S. et al. Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet.* **5**, e1000534 (2009).
85. Schork, N.J., Murray, S.S., Frazer, K.A. & Topol, E.J. Common vs. rare allele hypotheses for complex diseases. *Curr. Opin. Genet. Dev.* **19**, 212-219 (2009).
86. Slowikowski, K., Hu, X. & Raychaudhuri, S. SNPsea: an algorithm to identify cell types, tissues and pathways affected by risk loci. *Bioinformatics* **30**, 2496-2497 (2014).
87. Finucane, H.K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228-1235 (2015).
88. Zeng, W. et al. SilencerDB: a comprehensive database of silencers. *Nucleic Acids Res.* **49**, D221-D228 (2021).
89. Meylan, P., Dreos, R., Ambrosini, G., Groux, R. & Bucher, P. EPD in 2020: enhanced data visualization

and extension to ncRNA promoters. *Nucleic Acids Res.* **48**, D65-D69 (2020).

90. Bell, A.C., West, A.G. & Felsenfeld, G. The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell* **98**, 387-396 (1999).

91. Nora, E.P. et al. Targeted degradation of CTCF decouples local insulation of chromosome domains from genomic compartmentalization. *Cell* **169**, 930-944. e922 (2017).

92. Consortium, E.P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57 (2012).

93. Luo, Y. et al. New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Res.* **48**, D882-D889 (2020).

94. Consortium, R.E. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330 (2015).

95. Chen, S. et al. OpenAnnotate: a web server to annotate the chromatin accessibility of genomic regions. *Nucleic Acids Res.* **49**, W483-W490 (2021).

96. Zeng, W., Liu, Q., Yin, Q., Jiang, R. & Wong, W.H. HiChIPdb: a comprehensive database of HiChIP regulatory interactions. *Nucleic Acids Res.* **51**, D159-D166 (2023).

97. Li, W., Wong, W.H. & Jiang, R. DeepTACT: predicting 3D chromatin contacts via bootstrapping deep learning. *Nucleic Acids Res.* **47**, e60-e60 (2019).

98. Kaiser, L. et al. Fast decoding in sequence models using discrete latent variables. In *Proc. International Conference on Machine Learning*, 2390-2399 (2018).

99. Peng, J., Liu, D., Xu, S. & Li, H. Generating diverse structure for image inpainting with hierarchical VQ-VAE. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10775-10784 (2021).

100. Williams, W. et al. Hierarchical quantized autoencoders. In *Proc. Advances in Neural Information Processing Systems* **33**, 4524-4535 (2020).

101. Takida, Y. et al. Sq-vae: Variational bayes on discrete representation with self-annealed stochastic quantization. *arXiv preprint arXiv:2205.07547* (2022).

102. Paszke, A. et al. Pytorch: An imperative style, high-performance deep learning library. In *Proc. Advances in Neural Information Processing Systems* **32** (2019).

103. Kingma, D.P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

104. Grant, C.E., Bailey, T.L. & Noble, W.S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017-1018 (2011).

105. Kulakovskiy, I.V. et al. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.* **46**, D252-D259 (2018).

106. Zhang, J. et al. An integrative ENCODE resource for cancer genomics. *Nat. Commun.* **11**, 3696 (2020).

107. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).

108. Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110-121 (2010).

109. Kent, W.J. et al. The human genome browser at UCSC. *Genome Res.* **12**, 996-1006 (2002).

110. Nassar, L.R. et al. The UCSC genome browser database: 2023 update. *Nucleic Acids Res.* **51**, D1188-D1195 (2023).

111. Kent, W.J., Zweig, A.S., Barber, G., Hinrichs, A.S. & Karolchik, D. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* **26**, 2204-2207 (2010).

112. Wheeler, D.L. et al. Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **36**, D13-D21 (2007).

113. Su, A.I. et al. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad.*

Sci. U. S. A. **101**, 6062-6067 (2004).

114. Frankish, A. et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766-D773 (2019).

115. Hinrichs, A.S. et al. The UCSC genome browser database: update 2006. *Nucleic Acids Res.* **34**, D590-D598 (2006).