nature communications

Article

https://doi.org/10.1038/s41467-025-59780-5

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

Received: 10 September 2024

Accepted: 2 May 2025

Published online: 17 May 2025

Check for updates

Xuejian Cui[1], Qijin Yin[1], Zijing Gao[1], Zhen Li[1], Xiaoyang Chen[1], Hairong Lv[1], Shengquan Chen[2], Qiao Liu[3], Wanwen Zeng[3] ✉ & Rui Jiang[1] ✉

Cis-regulatory elements (CREs), including enhancers, silencers, promoters and insulators, play pivotal roles in orchestrating gene regulatory mechanisms that drive complex biological traits. However, current approaches for CRE identification are predominantly sequence-based and typically focus on individual CRE types, limiting insights into their cell-type-specific functions and regulatory dynamics. Here, we present CREATE, a multimodal deep learning framework based on Vector Quantized Variational AutoEncoder, tailored for comprehensive CRE identification and characterization. CREATE integrates genomic sequences, chromatin accessibility, and chromatin interaction data to generate discrete CRE embeddings, enabling accurate multi-class classification and robust characterization of CREs. CREATE excels in identifying cell-type-specific CREs, and provides quantitative and interpretable insights into CRE-specific features, uncovering the underlying regulatory codes. By facilitating large-scale prediction of CREs in specific cell types, CREATE enhances the recognition of disease- or phenotype-associated biological variabilities of CREs, thus advancing our understanding of gene regulatory landscapes and their roles in health and disease.

Gene regulation is a fundamental biological process that orchestrates gene expression through intricate networks 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 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 features are essential for establishing cell-type-specific gene expression patterns, driving cellular diversity, maintaining tissue homeostasis, and enabling 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.

Each type of CRE serves a distinct function in gene regulation: silencers suppress transcription, and enhancers amplify transcription, and promoters initiate transcription, and insulators act as boundary elements to regulate gene expression[3,4]. However, identifying CREs experimentally remains challenging, as it is resource-intensive, time-consuming, and constrained by limited knowledge of CRE-specific genetic signatures[9,12]. Massive genomic and epigenomic data derived from the rapid advancement of high-throughput sequencing technologies[13–16] has provided a valuable opportunity to identify cell-type-specific CREs using computational methods. For example,

[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, China. [2] School of Mathematical Sciences and LPMC, Nankai University, Tianjin, China. [3] Department of Statistics, Stanford University, Stanford, CA, USA. ✉ e-mail: wanwen@stanford.edu; ruijiang@tsinghua.edu.cn

DeepSEA uses convolutional neural networks (CNNs) to predict chromatin-profiling data such as transcriptional factors (TFs) binding sites, histone modification sites, and chromatin accessibility from DNA sequences[17]. DanQ combines convolutional and recurrent neural network architectures to predict non-coding DNA function from DNA sequences[18]. Enhancer-Silencer transition (ES-transition) is a sequence-based CNNs model for identifying cell-type-specific enhancers and silencers in the human genome, and has been utilized to uncover enhancer-silencer transitions depending on cellular context[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, current computational methods encounter many limitations and challenges. First, most existing approaches are designed to identify a single type of CRE[20–23], particularly enhancers, which have been extensively studied[24–30]. In contrast, silencers, 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 of CRE identification predominantly rely on DNA sequences[17–19], overlooking the potentially useful cell-type-specific features of CREs. Incorporating multi-omics data, including chromatin accessibility and chromatin interactions, can provide critical insights into the regulatory mechanisms and cell-type-specific functionalities of CREs, enabling comprehensive understanding of gene regulation. Third, conventional deep learning models often face challenges in deriving interpretable biological implications[32,33], hindering the meaningful large-scale identification of CREs and the understanding of the functional variability of CREs across diverse biological contexts.

To address these gaps, we propose CREATE (Cis-Regulatory Elements identificAtion via discreTe Embedding), a CNN-based supervised learning model that leverages the Vector Quantized Variational AutoEncoder (VQ-VAE) framework[34–36]. CREATE integrates genomic sequences with epigenetic features to enable a holistic approach for the identification and classification of multiple CRE types. The VQ-VAE framework is particularly well-suited for this task, as it can distill genomic and epigenomic data into discrete CRE embeddings, capturing the nuanced differences between various CRE types. By leveraging the discrete embeddings, CREATE facilitates the generation of a CRE-specific feature spectrum, offering both quantitative and interpretable insights into the specificity and functions of different CRE types. CREATE addresses several limitations of existing methods by integrating multi-omics data to better capture cell-type-specific CRE roles and regulatory mechanisms. Additionally, CREATE demonstrates superior performance in accurately identifying CREs and exhibits consistent performance across diverse input data and hyperparameters. By enabling large-scale predictions of CREs in specific cell types, CREATE offers valuable insights into disease- or phenotype-associated biological variabilities in CREs and serves as a useful tool for constructing a comprehensive CRE atlas. In summary, CREATE represents a promising step forward in computational CRE identification, and provides a strong foundation for advancing research into gene regulation and its implications for human health and disease.

Results

Overview of CREATE

CREATE is a CNN-based model built upon the VQ-VAE framework[34,35] to predict and classify multi-class CREs by integrating multi-omics data. By leveraging one-hot encoded genomic sequences, chromatin accessibility scores, and chromatin interaction scores, CREATE generates discrete CRE embeddings that enable a comprehensive and interpretable characterization of CREs (Fig. 1a and section "Methods").

The architecture of CREATE (Fig. 1b and section "Methods") comprises four key modules: (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: This module replaces the output embeddings of encoder module with their nearest neighbors in a discrete embedding space, referred to as the "codebook", adapted from VQ-VAE[34,35]. In brief, the features in the codebook are concatenated to form the final CRE embeddings. Unlike traditional VAE-based models[37], CREATE employs a dynamic codebook that is updated during training. This mechanism complements traditional continuous latent spaces by capturing the discrete patterns inherent to regulatory activities. (3) Decoder Module: The decoder reconstructs the original multi-omics inputs from the discrete embeddings through a two-step process. The integration decoder firstly reconstructs the integrated feature representation from the discrete embeddings, which is then further processed by omics-specific decoders to recover the individual data modalities, ensuring that the reconstructed data faithfully aligns with the original input features. (4) Classifier: To enhance the model's ability to distinguish between different CRE types, CREATE includes a classifier to encourage CREs of the same type are mapped to similar vectors, while those of different types are spread out across different vectors. This approach enhances accurate and interpretable classifications. All components of CREATE are trained jointly.

CREATE introduces several key innovations compared to existing studies: (1) Comprehensive Multi-Omics Integration: By incorporating genomic sequences alongside chromatin accessibility and interaction data, CREATE provides a holistic representation of the genomic context, enabling the identification of cell-type-specific CREs with enhanced performance. (2) Interpretability of Discrete Embeddings: The discrete embeddings offer biologically meaningful insights into CRE specificity, enabling clear interpretations of the regulatory roles of CREs and facilitating their classification across multiple types. Overall, CREATE represents a significant advancement in the computational identification of CREs. By integrating multi-omics data, capturing discrete embeddings, and offering interpretable characterization, CREATE provides 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 (see section "Methods"). We conducted 10-fold cross-validation experiments to compare the performance of CREATE 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 using the macro-averaged way (see section "Methods").

CREATE significantly outperforms the baseline methods across all evaluation metrics in both K562 and HepG2 cell types (one-sided paired Wilcoxon signed-rank tests P-values < 1e-3) (Fig. 2a, b and Supplementary Fig. 2a, b). For K562, 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 10.5% improvement over the second-best method, DeepICSH (0.743 ± 0.003) (Fig. 2d). These trends were mirrored in HepG2, where CREATE demonstrates comparable improvements, including a 9.1% increase in macro-averaged auPRC (Supplementary Fig. 2c, d).

Among the various CRE types, silencers and enhancers posed particular challenges due to their similar epigenetic signatures[31].

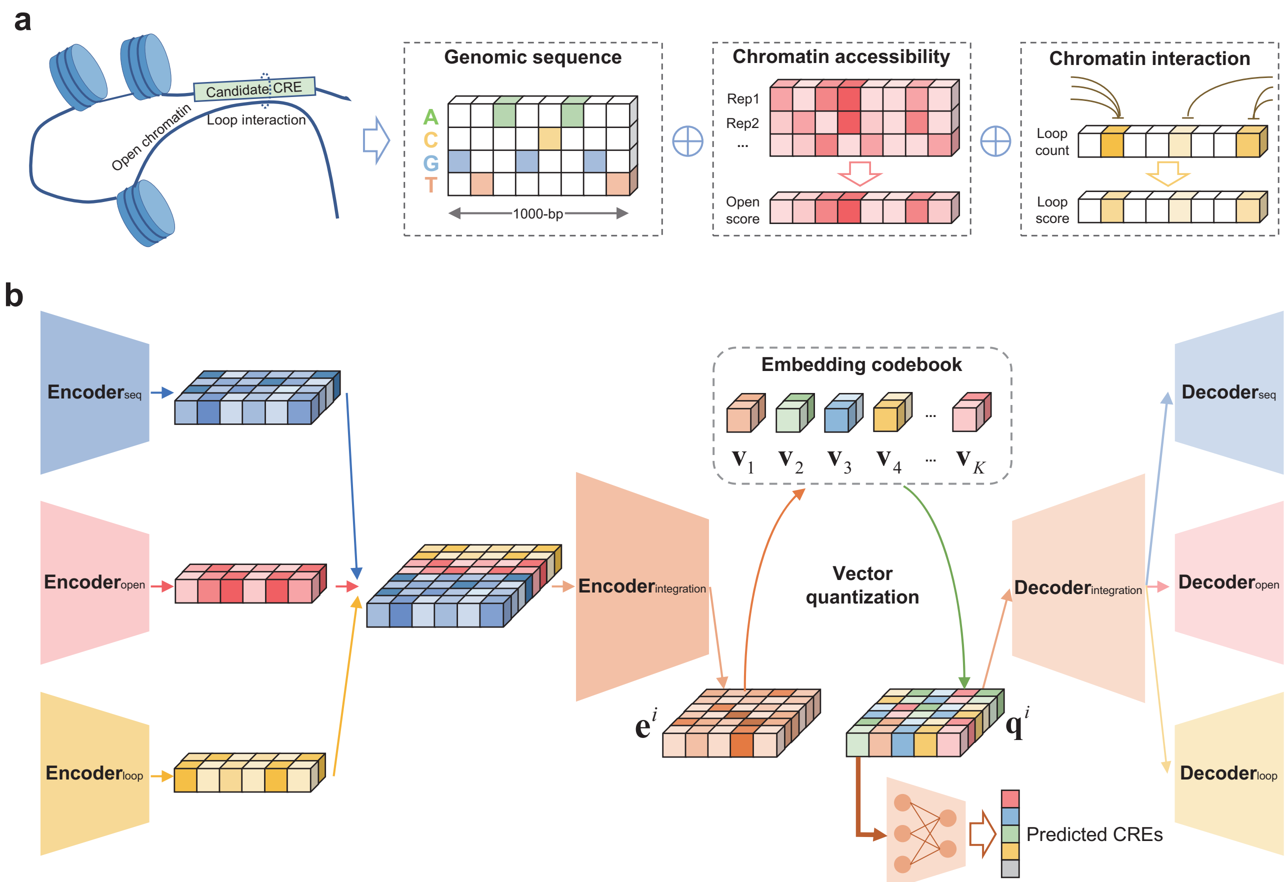

Fig. 1 | Overview of CREATE. a The input of CREATE model. CREATE takes as input the genomic sequence, chromatin accessibility score and chromatin interaction score. **b** The architecture of CREATE model. CREATE consists of encoders, a vector quantization module and decoders. The encoder module of CREATE 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 adapting 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.

Despite this, CREATE demonstrates a clear distinction between these difficult-to-differentiate elements. For K562, CREATE achieves a notable improvement in identifying silencers, with an auPRC that was 13.9% higher than the second-best method, highlighting its ability to accurately identify this underexplored CRE type (Fig. 2f and Supplementary Fig. 2e). Similarly, for enhancers, CREATE delivers a remarkable 22.1% improvement in auPRC compared to the second-best method (Supplementary Fig. 3a, b). This significant enhancement underscores the capability of CREATE capability to disentangle CREs with similar epigenomic patterns. For other CRE types including promoters, insulators and background regions, CREATE demonstrates optimal classification performance compared to baseline methods (Supplementary Fig. 3c–h). The trends observed in HepG2 closely mirrored those in K562, affirming the consistency of CREATE across different cellular contexts (Supplementary Fig. 4). Due to the limited availability of experimentally validated data for all CRE types, these two cell types served as primary benchmarks for comparison. To further evaluate its generalizability, we applied CREATE to GM12878 and HeLa-S3 cell types, where experimentally validated silencers were not available, and observed CREATE achieves accurate identification of CREs, demonstrating its ability to generalize effectively even in data-limited scenarios (Supplementary Note 1 and Supplementary Fig. 5).

CREATE consistently outperforms baseline methods in handling different numbers of background regions (Supplementary Note 2 and Supplementary Fig. 6), and excels in distinguishing overlapped CREs (Supplementary Note 3 and Supplementary Fig. 7), highlighting the advanced capability of CREATE in capturing the complexity of intricate regulatory mechanisms. The comprehensive evaluation of CREATE highlights its advanced capability in accurately identifying and distinguishing various CRE types, particularly those less studied or less abundant, such as silencers and enhancers. The superior performance of CREATE in capturing CRE variability and cell-type-specificity underscores its potential as a powerful tool for advancing the understanding of gene regulation, while its ability to generalize across different cell types and handle complex scenarios demonstrates its versatility and scalability for diverse genomic analyses.

Effectiveness and robustness of CREATE

The effectiveness of CREATE stems from its ability to accurately classify multiple CRE types, achieved through two key innovations: integration of multi-omics data and innovative discrete embedding. To assess the contributions of different input types, 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, the full CREATE model consistently demonstrates the highest classification performance (one-sided paired Wilcoxon signed-rank tests P-values < 1e-2), highlighting the importance of integrating chromatin accessibility and chromatin interactions for superior CRE identification (Fig. 3a and Supplementary Fig. 8a, b). Notably, CREATE shows

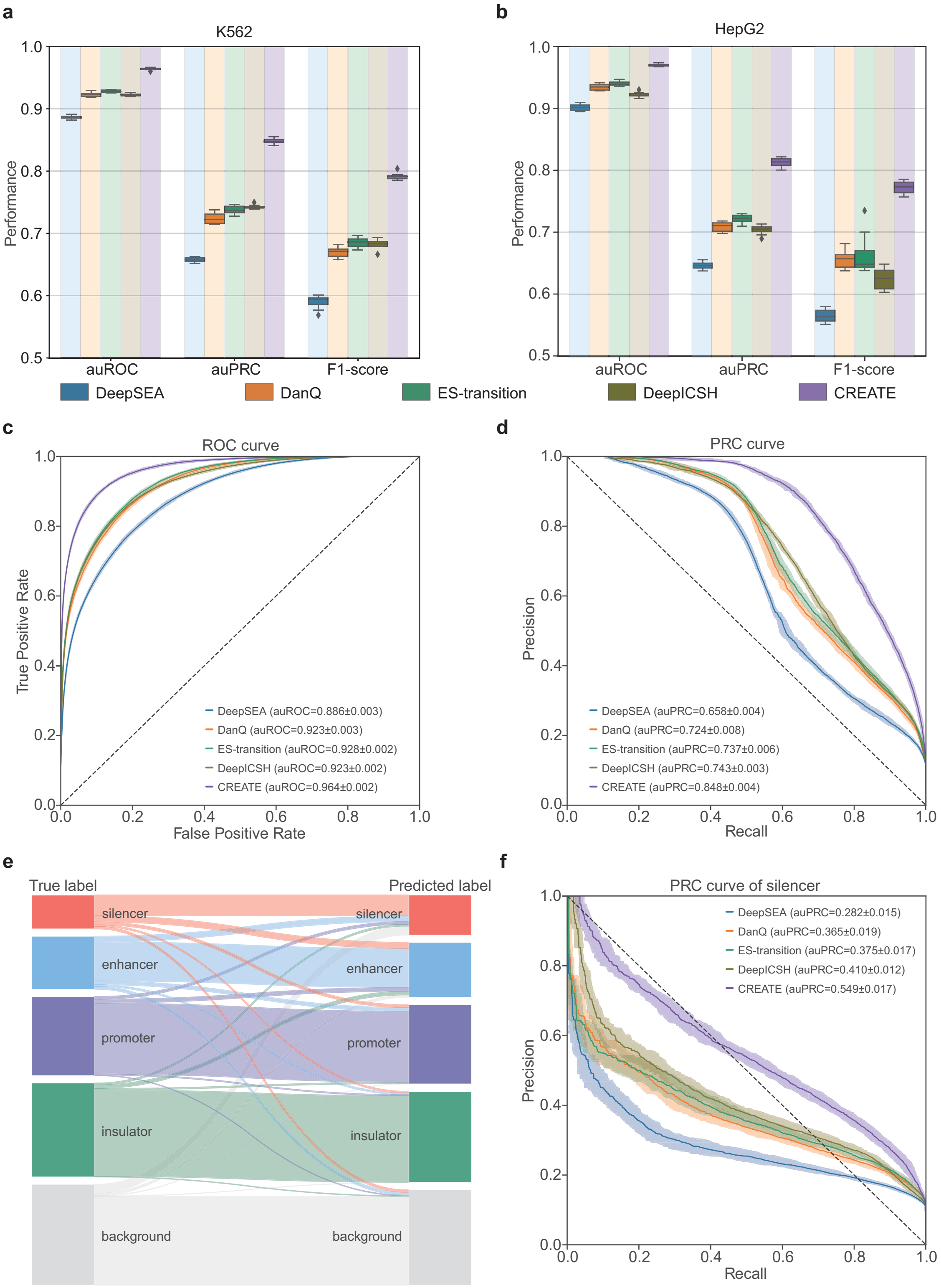

Fig. 2 | Evaluation of CREATE compared with the baseline methods. a, **b** Boxplot of 10-fold cross-validation classification performance ($n = 10$) evaluated by auROC, auPRC and F1-score on K562 cell type (**a**) and HepG2 cell type (**b**). Each box plot ranges from the upper to lower quartiles with the median as the horizontal line, whiskers extend to 1.5 times the interquartile range, and points represent outliers. Receiver Operating Characteristic curve (**c**) and Precision-Recall curve (**d**) comparing CREATE and baseline methods on K562 cell type. **e** The mapping between 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 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.

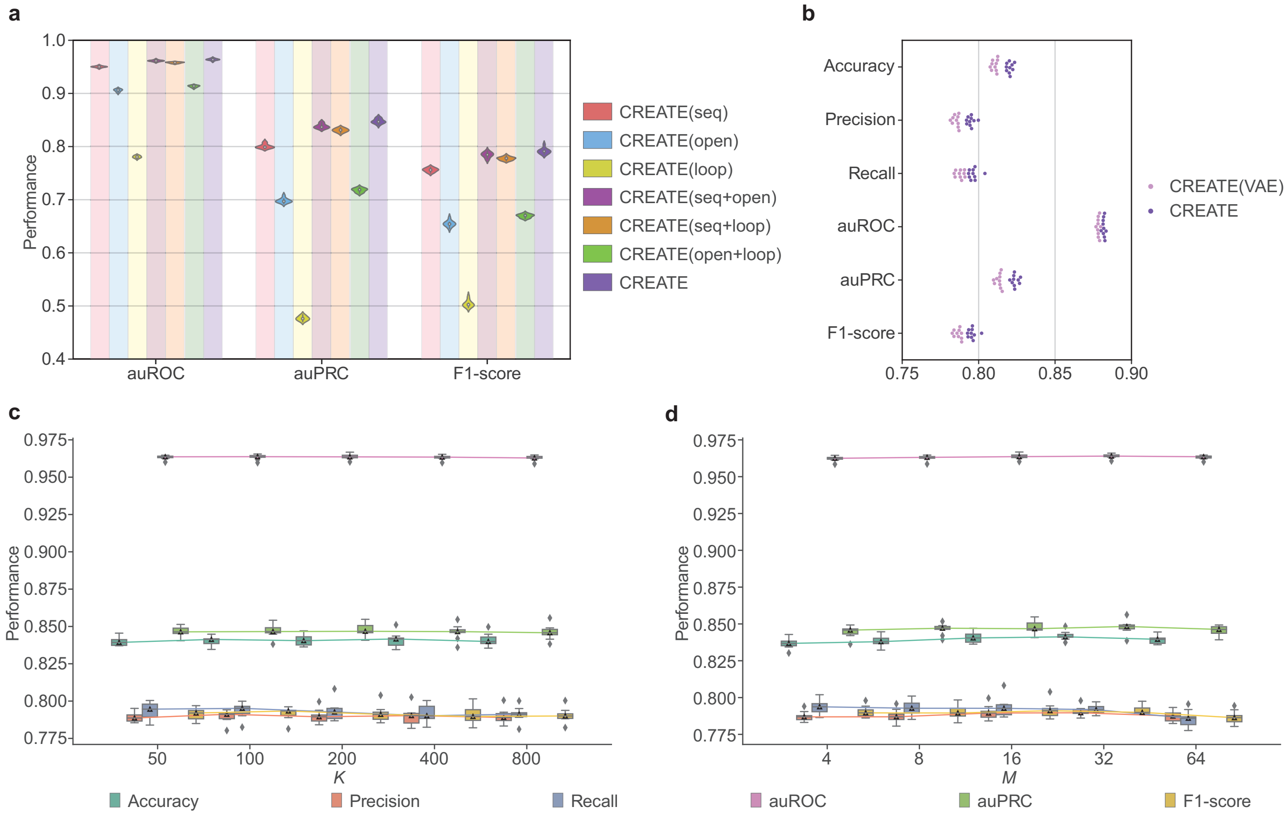

Fig. 3 | **Effectiveness and robustness of CREATE. a** Violin plot of 10-fold cross-validation classification performance ($n = 10$) evaluated by auROC, auPRC and F1-score for model ablation of CREATE on K562 cell type. **b** Swarm plot of classification performance evaluated by accuracy, precision, recall, auROC, auPRC and F1-score for CREATE compared with CREATE (VAE) on K562 cell type. **c** Classification performance of CREATE under different values of K (size of codebook) on K562 cell type ($n = 10$). **d** Classification performance of CREATE under different values of M (time of split quantization) on K562 cell type ($n = 10$). 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.

substantial improvements in identifying challenging CRE types such as silencers and enhancers (Supplementary Fig. 8c, d). Even when limited to genomic sequences alone, CREATE (seq) significantly outperforms baseline methods (one-sided paired Wilcoxon signed-rank tests P-values < 1e-3), achieving a tenfold macro-averaged auPRC of 0.800 ± 0.004—surpassing baseline methods by 5.7% in auPRC (Supplementary Fig. 9). This underscores the superior performance of CREATE with genomic sequences alone. Adding chromatin accessibility or chromatin interaction data further enhances performance, although using these additional inputs without genomic sequences results in relatively poorer outcomes (Fig. 3a and Supplementary Fig. 8a). This finding confirms the indispensable role of genomic sequences in CRE identification while emphasizing the complementary value of chromatin accessibility and interactions, particularly for the accurate classification of silencers and enhancers.

To verify the effectiveness of discrete embedding in CREATE, we compared CREATE with CREATE (VAE), a variant using VAE latent space, while keeping other modules and training strategies unchanged. CREATE significantly outperforms CREATE (VAE) across all evaluation metrics (one-sided paired Wilcoxon signed-rank tests P-values < 1e-3) (Fig. 3b). This demonstrates that discrete embeddings effectively capture complex CRE features, enabling CREATE to achieve superior classification performance.

The robustness of CREATE is reflected in its ability to maintain stable classification performance across diverse experimental conditions, including varying hyperparameters, input configurations, and challenges such as data imbalance. To validate the stability and robustness 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[33,36], 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 K562 cell type. CREATE exhibits 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 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[34,35], 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 K562 cell type. The results demonstrate that CREATE consistently obtains stable classification performance under different values of α (Supplementary Fig. 8e). 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 K562 cell type. The results demonstrate the stability of the classification performance under different values of μ (Supplementary Fig. 8f).

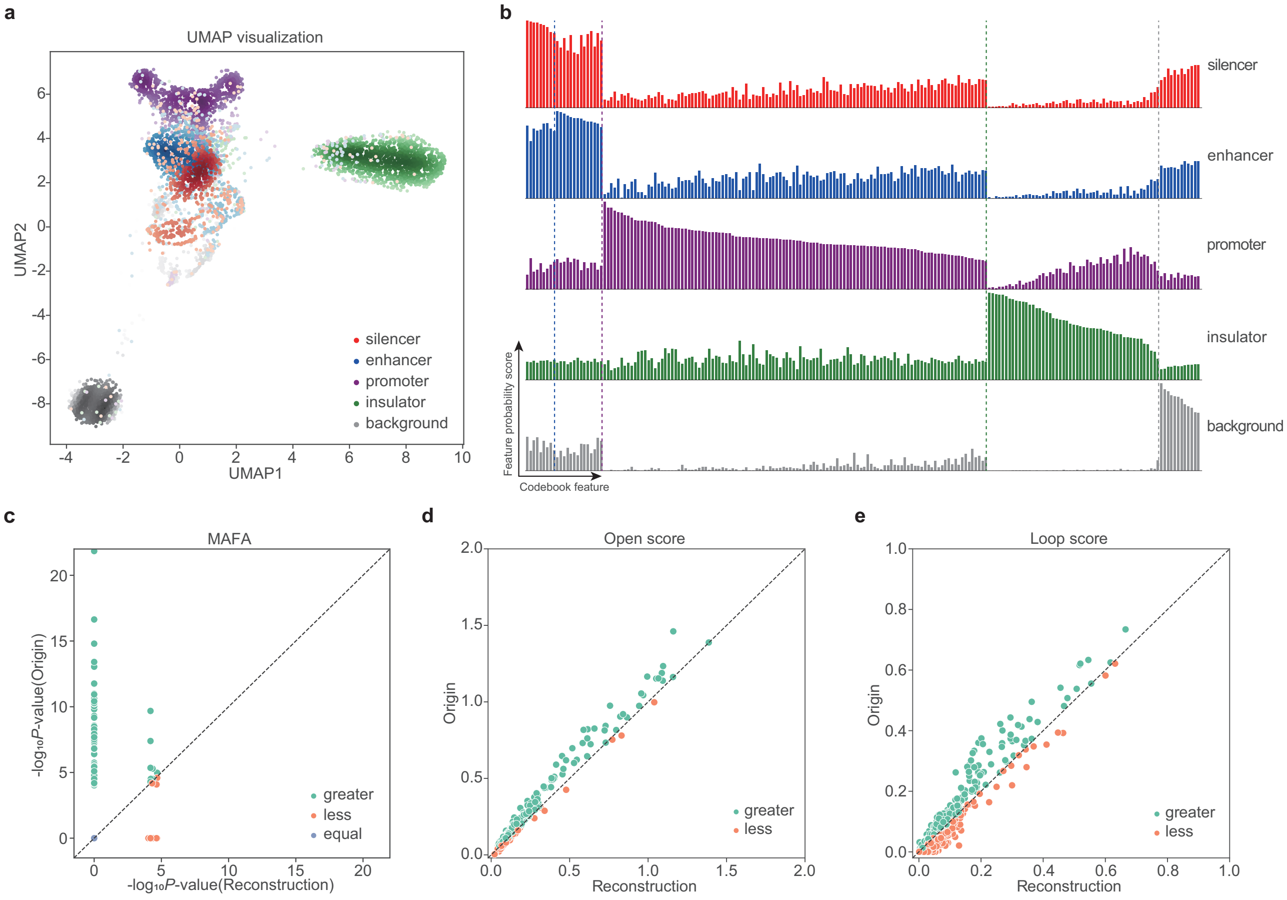

Fig. 4 | Generation and interpretation of CRE-specific feature spectrum. a UMAP visualization of the CRE embeddings from CREATE on the testing data in one of the tenfold cross-validation experiments of K562 cell type. **b** CRE-specific feature spectrum. There is a distinct set of specific features that are enriched or depleted in the feature spectrum of each CRE on 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. *P*-value is from the tool FIMO (see section "Methods"). **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.

To mitigate data imbalance, CREATE employs a data augmentation strategy after data partitioning and obtains significantly higher classification performance than model using class weights in the loss function, demonstrating the effectiveness of data augmentation in enhancing performance of CREATE (Supplementary Note 4 and Supplementary Figs. 10, 11). Reducing the number of silencers in K562 leads to decreases in both macro-averaged auPRC and silencer-specific auPRC, confirming the impact of insufficient training CREs, especially silencers (Supplementary Note 4 and Supplementary Fig. 12).

Additionally, CREATE demonstrates robustness to various input configurations (Supplementary Note 5 and Supplementary Figs. 14, 15). When incorporating different chromatin interaction datasets, CREATE consistently maintains high classification performance, with macro-averaged auPRCs significantly surpassing those achieved by CREATE (seq+open) (one-sided paired Wilcoxon signed-rank tests *P*-values < 5 e-2). These findings highlight the adaptability of CREATE to diverse input scenarios and training data.

To summarize, the combination of multi-omics integration and effective discrete embeddings underpins the high effectiveness of CREATE in CRE classification. By maintaining stable classification performance across a range of hyperparameters and offering flexibility in handling various data configurations, CREATE proves to be a robust tool for studying gene regulatory mechanisms and advancing our understanding of transcriptional regulation and disease processes.

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)[38] plot (Fig. 4a). Clearly, promoters, insulators and background regions are effectively separated. We also noted that silencers and enhancers are relatively hard to separate. To further validate the capability of CREATE in quantitatively articulating CRE specificity, we obtained specific feature spectrum for each type of CRE (Supplementary Fig. 16a and section "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 K562, there are specific features uniquely enriched in the feature spectrum of each CRE (Fig. 4b). For example, we 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. The feature spectrum of silencers contains a set of features (to the left of the blue dashed line), whose probability scores are notably higher than those 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 HepG2 cell type (Supplementary Fig. 16b, 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[39–41], 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. 16d, e), whereas similar comparison results were not observed for known activators, such as POU6F1[42] and MYC[43] (Supplementary Fig. 16f, 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. 17a–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[44]. ZNF582 has been implicated in DNA methylation processes, which are crucial for maintaining silencer function[45]. SCRT2 is less characterized, but its differential binding indicates a possible regulatory role in silencing mechanisms[46].

Similarly, the CRE-specific motif information is also harbored in the major feature of enhancers, promoters and insulators (Supplementary Fig. 17d–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

Despite the critical roles of CREs in gene regulation, the number of experimentally validated CREs remains limited. However, many cell types have been extensively profiled with epigenomic experiments, offering a rich source of candidate CREs that can serve as inputs for computational predictions. CREATE leverages this wealth of data to provide a powerful and interpretable tool for comprehensive CRE characterization, enabling accurate identification and classification of different CRE types. In our study, we collected 270,259 candidate CREs on K562 cell type and 232,456 candidate CREs on HepG2 cell type in 14 different epigenomic experiments for large-scale prediction (Supplementary Table 1 and Methods). Using trained CREATE models, we determined cutoff scores for each CRE type based on validation sets with false positive rate (FPR) not exceeding 0.01. We labeled candidate regions as predicted silencers, enhancers, promoters, or insulators when their scores exceeded the corresponding cutoff scores, while the remaining sequences classified as background regions. This approach identified 26,012 predicted silencers, 29,423 predicted enhancers, 2057 predicted promoters, and 10,558 predicted insulators in K562 cell type. Similarly, in HepG2 cell type, we identified 16,000 predicted silencers, 49,145 predicted enhancers, 4422 predicted promoters and 13,122 predicted insulators. Regarding the relatively low number of predicted promoters, we feel the main reason is that the candidate CREs contain a limited number of promoter-related regions, as experimentally validated promoters have been excluded from these candidate CREs (Supplementary Note 8).

The proportions of predicted CREs derived from different epigenomic tracks align with their expected biological roles, supporting the biological relevance of the predictions. For example, H3K27me3, a histone modification associated with silencers[47–49], was observed at higher proportions in predicted silencers (9.0% in K562 and 16.0% in HepG2) compared to other CRE types (average 2.2% in K562 and average 4.1% in HepG2) (highlighted with red dashed box; Fig. 5a and Supplementary Fig. 18a). Similarly, histone modifications associated with enhancers, such as H3K9ac[50,51], H3K27ac[51–53], H3K4me1[51,54,55], H3K4me2[55] and H3K4me3[55,56], 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. 18a). Moreover, the predicted CREs exhibited cell-type-specific patterns similar to those of true (experimentally validated) CREs (Supplementary Notes 6 and 7 and Supplementary Figs. 19–21).

To further validate the epigenomic characteristics of predicted CREs on a large scale, we conducted comprehensive comparative analyses. First, TFs play a crucial regulatory role in regulating gene transcription by binding to CREs, and sequence-specific TF motifs are key indicators of CRE functionality[57–59]. To assess whether predicted CREs exhibit similar regulatory characteristics as true CREs (experimentally validated CREs), we performed motif enrichment analysis on both true 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[60], TFAP2A[61], MAFA[39], MAFB[62], LHX6[40], PAX8[41], NFIA[63] and PRDM6[64], which are also enriched in predicted silencers (Fig. 5c). Motifs belonging to active TFs, including POU6F1[42], MYC[43], ZFHX3[65] and SOX8[66], are enriched consistently across true and predicted enhancers (Supplementary Fig. 18b, 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[43,65,67,68], which aligns with the potential conversion between silencers and enhancers under different conditions[19,69]. Motif enrichment analysis also reveals similar patterns for promoters, where the motifs enriched in gene promoters (experimentally validated promoters of coding genes) were consistently found in predicted promoters (Supplementary Fig. 18d, e). These results highlight the ability of CREATE in accurately capturing CRE-specific sequence characteristics, reinforcing the biological relevance of its predictions.

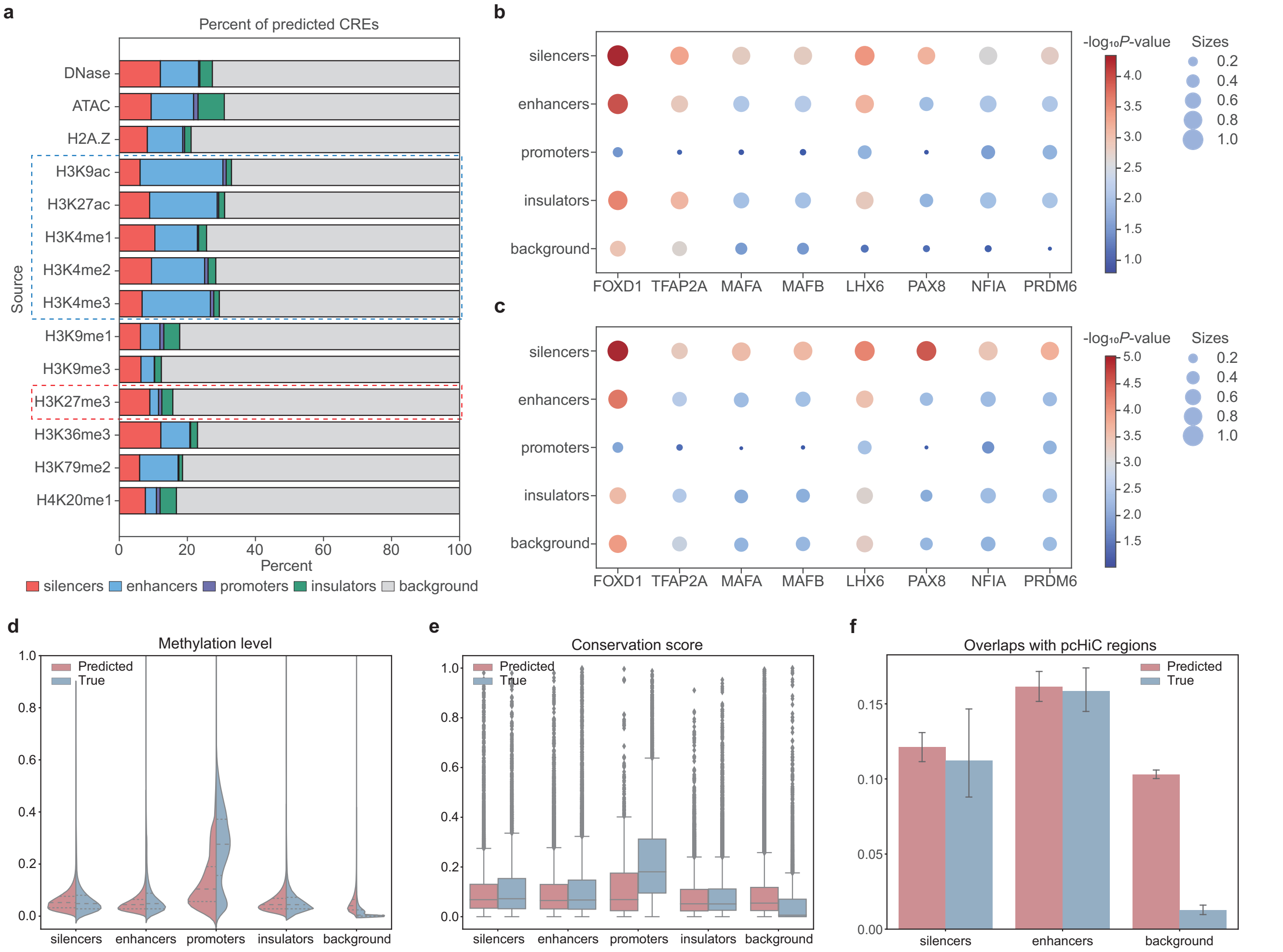

Fig. 5 | Characteristics of predicted CREs by CREATE. a Percentage of predicted CREs and background regions from different candidate sources in K562 cell type. Candidate source indicates which type of chromatin accessible or histone modification peaks the candidate regions originates from. Bubble plot of motif enrichment significance (-log$_{10}$*P*-value) of repressive TFs (silencer-related TFs) at true CREs (**b**) and predicted CREs (**c**) on K562 cell type. The legend title "Sizes" represents the proportion of CREs significantly enriched with motifs of the tested TF (*P*-value < 0.01). *P*-value is from the tool FIMO (see e section "Methods"). **d** Violin plot of methylation levels at true CREs and predicted CREs on 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 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 the number of pcHiC regions overlapping with true and predicted silencers, enhancers and background regions on K562 cell type. The error bars denote the 95% confidence interval, and the centers of error bars denote the average value. About predicted CREs, there are 26,012 silencers, 29,423 enhancers, 2057 promoters, 10,558 insulators, and 202,209 background regions (**d**–**f**). About true CREs, there are 6754 silencers, 10,528 enhancers, 15,699 promoters, 18,631 insulators, 20,000 background regions (**d**–**f**).

Second, DNA methylation is an important epigenetic modification involved in gene regulation, particularly gene silencing[70,71]. 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 because predicted promoters are more likely to be non-coding RNA (ncRNA) promoters (Supplementary Note 8 and Supplementary Figs. 22 and 23). 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) (Fig. 5d and Supplementary Fig. 18f), which is also observed for true CREs compared to true background regions (one-sided Wilcoxon rank-sum tests *P*-values < 3e-308) (Fig. 5d).

Third, conservation is a key feature of CREs across vertebrates, as conserved regions often play critical roles in gene regulation[72–74]. We computed the phastCons scores for predicted CREs as their conservation scores (Methods). We noticed that the conservation scores of predicted CREs exhibit strong consistency with those of true CREs, with predicted promoters being the only exception (Fig. 5e). This exception observed in predicted promoters may be because they are more likely to be ncRNA promoters (Supplementary Note 8 and Supplementary Figs. 22 and 23). Predicted CREs except insulators are significantly more conserved than predicted background regions (one-sided Wilcoxon rank-sum tests *P*-values < 1e-13) (Fig. 5e and Supplementary Fig. 18g), and true CREs are significantly more conserved than true background regions (one-sided Wilcoxon rank-sum tests *P*-values < 3e-308) (Fig. 5e).

Fourth, CREs frequently regulate gene expression by connecting promoters through chromatin loops, which can be identified by promoter-capture HiC (pcHiC)[75]. To evaluate the functional relevance of predicted CREs, we quantified the number of pcHiC regions overlapping with true CREs and predicted CREs (see section "Methods"), and perceived predicted silencers and enhancers harbor significantly more overlaps with pcHiC regions than predicted background regions (one-sided Wilcoxon rank-sum tests *P*-values < 2e-3), similar to true CREs compared to true background regions (one-sided Wilcoxon rank-sum tests *P*-values < 5e-11) (Fig. 5f). These findings suggest that the predicted CREs likely play functional roles in modulating target gene expression through chromatin interactions. Furthermore, the expression levels of genes associated with predicted CREs align with their known regulatory functions. Specifically, predicted enhancers are linked to increased gene expression, while predicted silencers are associated with reduced gene expression (Supplementary Note 10 and Supplementary Fig. 24). This consistency further validates the effectiveness of CREATE in large-scale CRE prediction.

Collectively, CREATE enables large-scale prediction of CREs with high accuracy and biological relevance, bridging the gap between experimental limitations and extensive epigenomic data. Predicted CREs align closely with true CREs across multiple metrics, including epigenomic markers, methylation levels, conservation scores, and chromatin interactions, validating their functional roles. CREATE facilitates the construction of a comprehensive CRE atlas, providing critical insights into gene regulation.

Characterization of dual-function regulatory elements

Dual-function regulatory elements (DFREs) are regions that act as silencers in one cell type and enhancers in another, depending on the cellular context[9,19,69,76]. These multifunctional elements are critical for understanding the complexity of gene regulation, as their ability to switch functions enables dynamic control of gene expression. Following the original study of DFRE[19], within a total of 26,012 silencers predicted in K562, we identified 2409 (9.3%) that overlapped with predicted enhancers in HepG2 by at least 600 bps as DFREs and the rest 23,603 (90.7%) as normal ones. These DFREs show similar silencer scores predicted in K562 to normal silencers (two-sided Wilcoxon rank-sum test *P*-value > 0.79) (Supplementary Fig. 25a), while these DFREs display significantly higher enhancer scores predicted in K562 than normal silencers (one-sided Wilcoxon rank-sum test *P*-value < 6e-47) (Fig. 6a). Furthermore, we identified 36,448 predicted enhancers in HepG2 that are non-overlapping with any silencers or enhancers in K562 and treated them as normal enhancers. The above DFREs show similar enhancer scores predicted in HepG2 to normal enhancers (two-sided Wilcoxon rank-sum test *P*-value > 0.43) (Supplementary Fig. 25b), while these DFREs display significantly higher silencer scores predicted in HepG2 than normal enhancers (one-sided Wilcoxon rank-sum test *P*-value < 2e-5) (Fig. 6b). These results clearly demonstrate the ability of CREATE in identifying and characterizing multifunctional regulatory elements.

To further explore the biological significance of the above 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 in K562 have the lowest conservation scores (one-sided Wilcoxon rank-sum tests *P*-values < 2e-31) (Fig. 6c), suggesting that these elements are evolutionarily conserved 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 compared to normal silencers (one-sided Wilcoxon rank-sum test *P*-value < 7e-4) (Fig. 6d) and normal enhancers (one-sided Wilcoxon rank-sum test *P*-value < 5e-79) (Supplementary Note 11 and Supplementary Fig. 26a). This observation highlights the unique epigenomic signatures of DFREs, suggesting that their dual functionality may be 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, Supplementary Note 11 and Supplementary Fig. 26b). This indicates that DFREs may be actively involved in chromatin looping interactions, which are critical for mediating gene expression and regulatory network organization. Fourth, we computed the number of expression quantitative trait loci (eQTLs) from whole-blood or liver tissues from GTEx[77,78] located in DFREs, normal silencers, normal enhancers and background regions in K562 (see section "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.

In summary, the ability of CREATE to differentiate DFREs from normal regulatory elements and characterize their epigenomic, evolutionary, and functional properties (Supplementary Note 11 and Supplementary Figs. 27–30) highlights its potential of understanding complex mechanisms behind dual-function regulatory functions.

Disease-associated variants analysis and tissue-specific enrichments in CREs

CREs play critical roles in disease susceptibility and phenotype variants, often harboring single-nucleotide polymorphisms (SNPs) and eQTLs associated with various traits and diseases[79,80]. To evaluate the capacity of CREATE in identifying disease-relevant variants within CREs, we analyzed the number of SNPs from dbSNP[81–83] database and eQTLs from GTEx[77,78] located in true CREs and predicted CREs (see section "Methods"). The results show that the distributions of genetic variants at predicted CREs align closely with those at true CREs (Fig. 7a and Supplementary Fig. 32a–c). Notably, predicted CREs are significantly enriched with more rare SNPs than predicted background regions (one-sided Wilcoxon rank-sum tests *P*-values < 2e-9) (Fig. 7b), whereas no significant enrichment was observed for common SNPs (Supplementary Fig. 32d). Rare variants are more impactful in complex diseases compared to common variants[84,85]. The enrichment results suggest that the identified CREs likely capture cell-type-specific regulatory functionality, reflecting the contributions of disease-associated variants[9,12,86]. Further supporting this finding, whole-blood eQTLs are significantly enriched in silencers and enhancers compared to background regions (one-sided Wilcoxon rank-sum tests *P*-values < 4e-49) (Supplementary Fig. 32e). However, no significant enrichment was observed when considering all eQTLs (Supplementary Fig. 32f), reinforcing the tissue-specific regulatory role of CREs identified by CREATE. Moreover, genetic variants levels in both true and predicted CREs gradually decrease with increasing CREATE background scores (Fig. 7c and Supplementary Fig. 33a, d–i), while increase with increasing CREATE silencer scores (Supplementary Fig. 33b, c), explicating the ability of CREATE in quantifying the regulatory impact of genetic variants.

To quantitatively assess the ability of CREATE in capturing CRE-specific sequence characteristics, we portrayed the correlation between the CREATE scores and motif enrichment significance in true and predicted CREs. We recognized a strong positive correlation between CREATE silencer scores and the enrichment significance of silencer-related TFs (Fig. 7d, e and Supplementary Figs. 34–37), as well as between CREATE enhancer scores and the enrichment significance of enhancer-related TFs (Supplementary Fig. 18b, c). These results demonstrate the capacity of CREATE in revealing CRE-specific sequence features and their association with TF-binding motifs.

To explore the tissue-specific regulatory roles of variants in true and predicted CREs, we performed tissue enrichment analysis using SNPsea[87] (see section "Methods"). Silencers and enhancers predicted by CREATE show significant tissue enrichment for blood-related tissues compared to background regions (Fig. 7f and Supplementary Fig. 38), consistent with the enrichment patterns of true CREs (Supplementary Fig. 39). Concretely, K562 was identified as a significantly enriched tissue for predicted silencers and enhancers in K562 cell type (*P*-values < 2e-5), affirming the tissue specificity of CREs identified by CREATE.

To further validate the ability of CREATE in studying the impact of variants in phenotypes, we conducted heritability enrichment analysis utilizing partitioned linkage disequilibrium score regression (LDSC)[88]

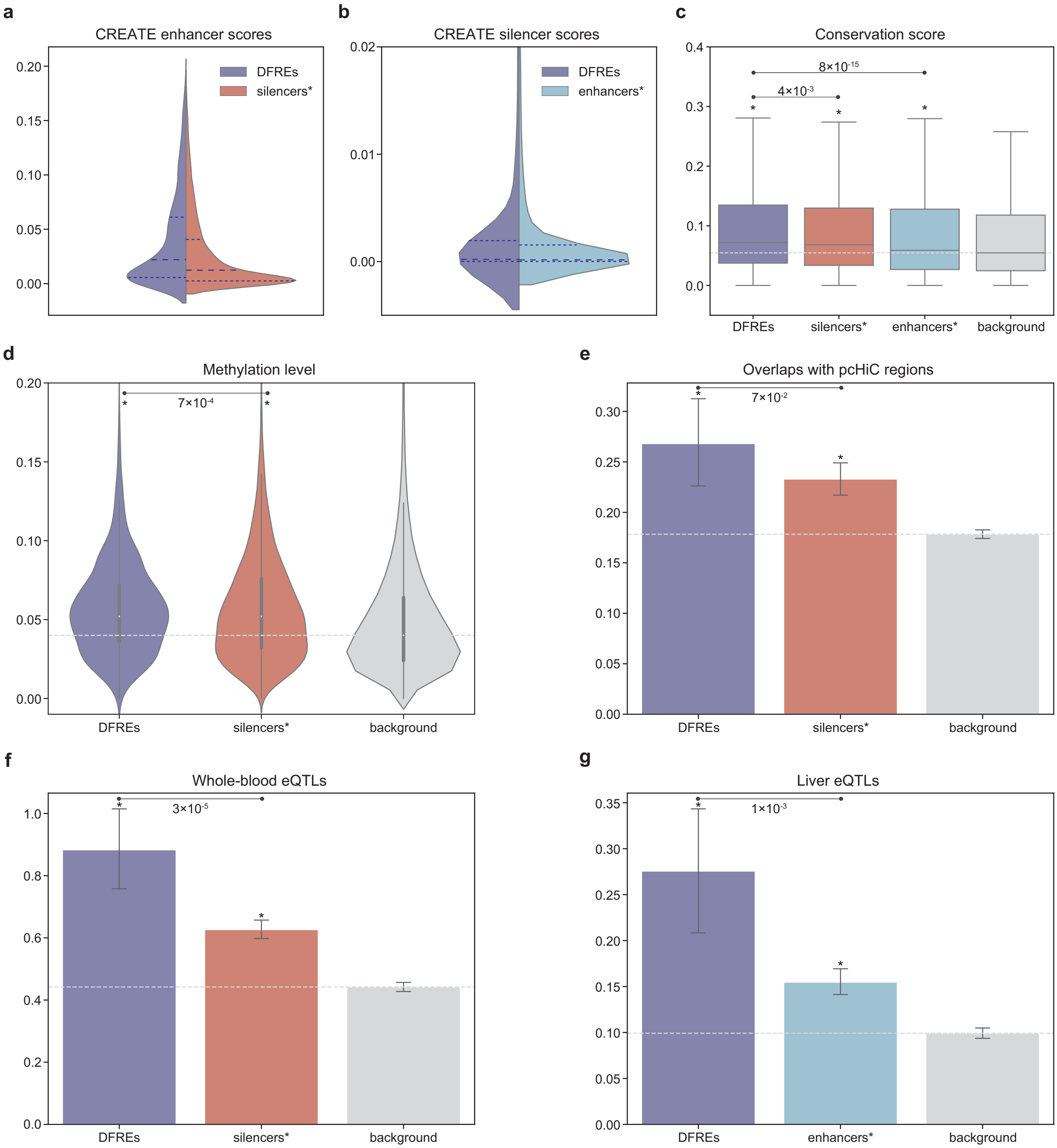

Fig. 6 | Characterization of DFREs functioning as silencers in K562 and as enhancers in HepG2. a Violin plot of the enhancer scores predicted by CREATE in K562 for DFREs and normal silencers (silencers*). **b** Violin plot of the silencer scores predicted by CREATE in HepG2 for DFREs and normal enhancers (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 ($n = 2409$), normal silencers (silencers*) ($n = 23{,}603$), normal enhancers (enhancers*) ($n = 36{,}448$) and background regions ($n = 202{,}209$) in K562. The asterisks above the boxes indicate the significant enrichments compared with background regions. (*) One-sided Wilcoxon rank-sum test *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 background regions in K562. (*) One-sided Wilcoxon rank-sum test *P*-value < 2e-6. Each box plot in violin ranges from the upper to lower quartiles with the median as the horizontal line, and whiskers extend to 1.5 times the interquartile range. **e** Bar plot of the number of pcHiC regions overlapping with DFREs, normal silencers (silencers*) and background regions in K562. (*) One-sided Wilcoxon rank-sum test *P*-value < 2e-3. **f** Bar plot of the number of whole-blood eQTLs located in DFREs, normal silencers (silencers*) and background regions in K562. (*) One-sided Wilcoxon rank-sum test *P*-value < 5e-21. **g** Bar plot of the number of liver eQTLs located in DFREs, normal enhancers (enhancers*) and background regions in HepG2. (*) One-sided Wilcoxon rank-sum test *P*-value < 2e-6. The error bars denote the 95% confidence interval, and the centers of error bars denote the average value.

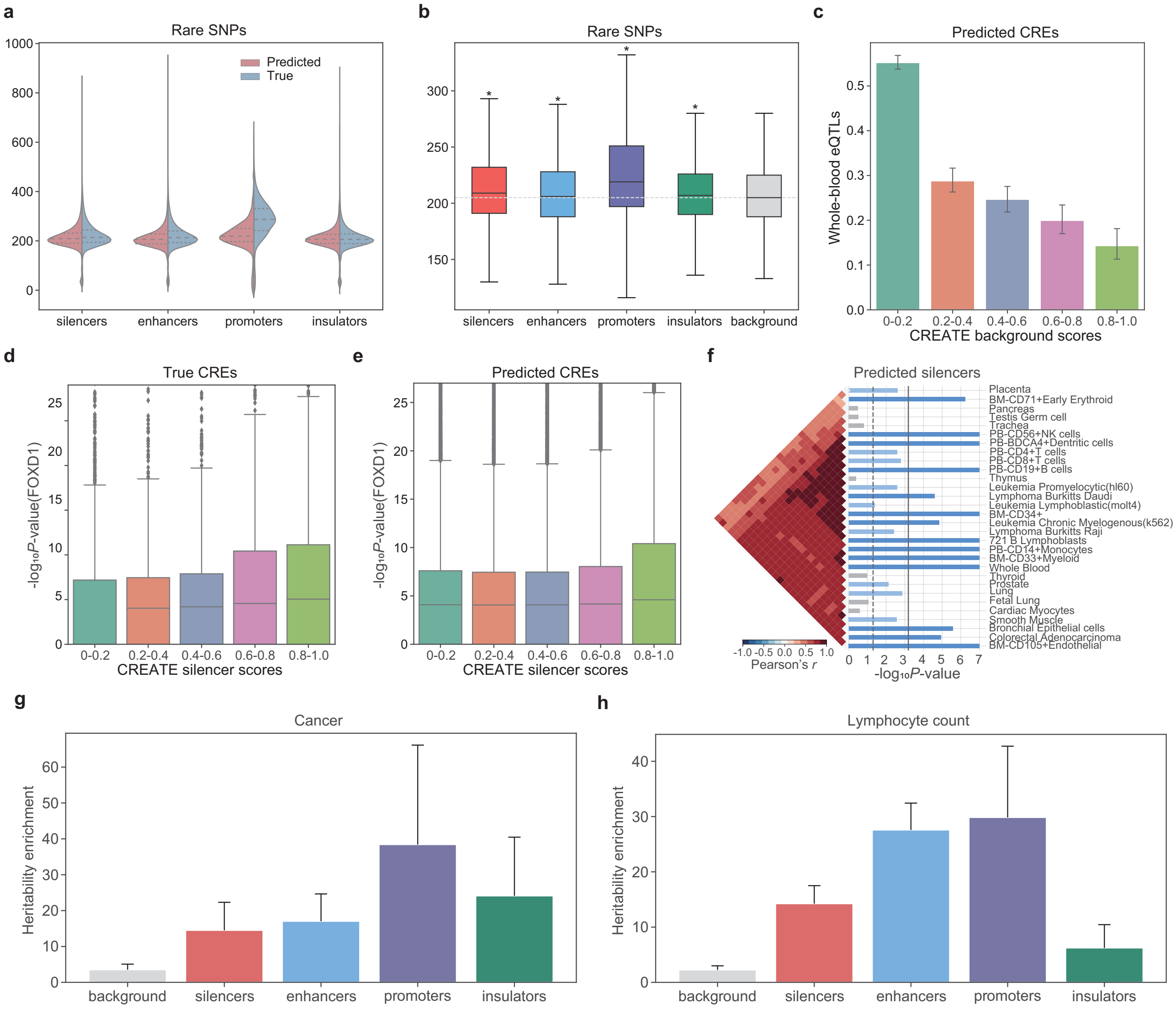

Fig. 7 | Identification of the biological variability of CREs by CREATE. a Violin plot of the number of rare SNPs within true CREs and predicted CREs on K562 cell type. Each violin plot contains three horizontal dashed lines denoting the median, the upper quartile, and the lower quartile. **b** Box plot the number of rare SNPs within predicted CREs and background regions on K562 cell type. The asterisks above the boxes indicate the significant enrichments compared with the background regions. (*) One-sided Wilcoxon rank-sum test *P*-value < 2e-9. There are 26,012 predicted silencers, 29,423 predicted enhancers, 2057 predicted promoters, 10,558 predicted insulators, and 202,209 predicted background regions. 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 the number of whole-blood eQTLs within predicted CREs on K562 cell type. 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 K562 cell type. *P*-value is from the tool FIMO (see section "Methods"). 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 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 line 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. 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.

(see section "Methods"). Predicted CREs demonstrated higher heritability enrichment for blood-related phenotypes, including cancer and lymphocyte count, compared to predicted background regions (Fig. 7g, h and Supplementary Fig. 40). This enrichment pattern is consistent with true CREs in both K562 (Supplementary Fig. 41) and HepG2 cell types (Supplementary Note 12 and Supplementary Figs. 42 and 43). These findings indicate that CREs predicted by CREATE inherit the heritability contribution patterns of true CREs, further supporting their biological relevance in complex traits and diseases.

Altogether, CREs identified by CREATE are enriched with disease-associated variants and exhibit tissue-specific regulatory roles, providing critical insights into regulatory dynamics during development and disease progression. These capabilities underscore the potential of CREATE in advancing our understanding of gene regulation and its implications for complex traits and diseases.

Discussion

CREATE represents a substantial step forward in the prediction and interpretation of CREs, offering a versatile multimodal architecture that integrates DNA sequences, cell-type-specific chromatin accessibility, and chromatin interaction data. Unlike existing methods that typically rely on genomic sequence input and focus on a single CRE type, CREATE enables multi-class classification and uncovers diverse CRE types, including silencers, enhancers, promoters, insulators, and background regions. Utilizing discrete CRE embeddings, we have verified the superior performance of CREATE in accurate CRE identification compared to the existing methods, as well as its adaptability and stability across various configurations of input data, hyperparameters. One of the key strengths of CREATE lies in its ability to quantitatively and interpretably elucidate CRE-specific characteristics using the CRE-specific feature spectrum. Moreover, CREATE has been validated the potential in identifying cell-type-specific novel CREs on a large scale and uncovering disease-associated genetic variants of these predicted CREs, illustrating the ability of CREATE in unveiling the underlying regulatory dynamics that drive transcriptional regulation and disease development.

Despite its notable achievements, there are areas where CREATE could be further refined. Current limitations include data imbalance, particularly for silencers and underrepresented cell types, impairing the overall performance of CRE identification. To address this, we plan to update our predicted silencers to the SilencerDB database[89] and expand our identification to include more cell types. Additionally, the per-base-paired input features and input-specific encoder-decoder structure, while effective for extracting detailed and comprehensive CRE embeddings, face challenges in representing complex epigenetic features, such as TF binding. To improve scalability and representation, future iterations of CREATE will focus on expanding the diversity of training data, incorporating prior biological knowledge into the model, and leveraging emerging technologies like HiChIP to enrich the input feature space. Looking ahead, we aim to develop a unified foundation model for characterization of gene regulation. By leveraging the shareability, scalability and interpretability of discrete embedding, we envision a model that supports broad applications across diverse contexts, enabling the integration and generalization of findings in regulatory genomics. This evolution has the potential to uncover new insights into gene regulation mechanisms and their role in disease development, further advancing our understanding of developmental biology and precision medicine.

In conclusion, CREATE offers a powerful and interpretable framework for multi-class CRE prediction and interpretation. By integrating diverse data types and providing biologically meaningful insights, it establishes a solid foundation for future research. With ongoing improvements and expansions, CREATE is poised to contribute significantly to the exploration of gene regulation and its interplay with disease, paving the way for innovative discoveries and practical applications in genomics and medicine.

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[89]. K562 represents human chronic myeloid leukemia cell line and HepG2 represents human hepatocellular carcinoma (liver cancer) cell line. 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[90] and defined 1kb regions surrounding TSSs (500 bp upstream and 500 bp downstream) as promoters. We took as insulators the CTCF[91,92] Chromatin immunoprecipitation sequencing (ChIP-seq) peaks for K562 and HepG2 cell types collected from the ENCODE project[93,94].

About the candidate CREs for large-scale prediction, we collected multiple histone modification ChIP-seq peaks and chromatin accessibility peaks for K562 and HepG2 cell types from the Roadmap project[95] and ENCODE project (Supplementary Table 1). After filtering the regions overlapping with the experimentally validated CREs, known genes, and consensus black list, a curated list of regions within the genome that are systematically excluded from analysis in genomics studies[93,94], we obtained 270,259 and 232,456 candidate CREs for large-scale prediction on K562 and HepG2 cell types, respectively.

We generated background regions by randomly sampling DNA sequences from the entire human reference genome that excludes 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 K562 cell type, and 1456 silencers, 11,407 enhancers, 14,535 promoters, 15,650 insulators and 20,000 background regions for 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[96] 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[16,97], to enhance the identification of CREs. We first calculated the number of chromatin loops per base pair for each CRE or background region, 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[21,22,98]. As illustrated in Supplementary Fig. 1, for each CRE with length of 1000 bp, we shifted a window along the reference genome with a stride of 10 from the midpoint towards both ends and ensured that no overlap occurred between augmented sequences derived from different original sequences to avoid information leakage during training. 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 an encoder module, a vector quantization module, and a decoder module. The encoder module includes multiple omics-specific encoders and an integration encoder for synthesizing information from omics-specific encoders. 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 up-sample layer, and a ReLU non-linear activation function or a Sigmoid non-linear activation function for the sequence-specific reconstruction decoder. In addition, we introduced a classifier with three fully connected layers to predict CREs based on their embeddings. Specifically, the output of 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[33,36], it will be split into $L' \times M$ vectors $\mathbf{e}^i_{l,j} \in \mathbb{R}^D, l \in \{1, \ldots, L'\}, j \in \{1, \ldots, M\}$, where M is the time of split quantization. Utilizing a shared codebook $\mathbf{v}_k \in \mathbb{R}^D, k \in \{1, \ldots, K\}$ 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:

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

Model training

We employed multiple updating 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:

$$L_{recon1}(\mathcal{B}_0) = -\frac{1}{P \cdot L \cdot |\mathcal{B}_0|} \sum_{i=1}^{|\mathcal{B}_0|} \sum_{l=1}^{L} \sum_{p=1}^{P} \left[\mathrm{S}^i_{lp} \log\left(\hat{\mathrm{S}}^i_{lp}\right) + \left(1 - \mathrm{S}^i_{lp}\right) \log\left(1 - \hat{\mathrm{S}}^i_{lp}\right) \right] \tag{2}$$

$$L_{recon2}(\mathcal{B}_0) = \frac{1}{L \cdot |\mathcal{B}_0|} \sum_{i=1}^{|\mathcal{B}_0|} \sum_{l=1}^{L} ||\mathrm{O}^i_l - \hat{\mathrm{O}}^i_l||_2^2 \tag{3}$$

$$L_{recon3}(\mathcal{B}_0) = \frac{1}{L \cdot |\mathcal{B}_0|} \sum_{i=1}^{|\mathcal{B}_0|} \sum_{l=1}^{L} ||\mathrm{L}^i_l - \hat{\mathrm{L}}^i_l||_2^2 \tag{4}$$

$$L_{recon}(\mathcal{B}_0) = L_{recon1}(\mathcal{B}_0) + \beta L_{recon2}(\mathcal{B}_0) + \gamma L_{recon3}(\mathcal{B}_0) \tag{5}$$

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:

$$L_{encoder}(\mathcal{B}_0) = \frac{1}{M \cdot L' \cdot |\mathcal{B}_0|} \sum_{i=1}^{|\mathcal{B}_0|} \sum_{l=1}^{L'} \sum_{j=1}^{M} \left|\left| \mathbf{e}^i_{l,j} - \mathrm{sg}\left(\mathbf{v}_{\underset{k \in \{1,\ldots,K\}}{\operatorname{argmin}} ||\mathbf{e}^i_{l,j} - \mathbf{v}_k||_2^2} \right) \right|\right|_2^2 \tag{6}$$

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

Third, we followed the recommendation from both the original studies of VQ-VAE[34,35] and recent related researches[99–102] 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 $\left\{ \mathbf{e}^*_{k,m} \middle| m = 1, \ldots, n_k \right\}$ to optimize the code $\mathbf{v}_k$ as follows:

$$N_k^{(t)} = (1 - \mu) N_k^{(t-1)} + \mu n_k^{(t)} \tag{7}$$

$$\mathbf{u}_k^{(t)} = (1 - \mu) \mathbf{u}_k^{(t-1)} + \sum_{m=1}^{n_k^{(t)}} \mu \mathbf{e}^{*,(t)}_{k,m} \tag{8}$$

$$\mathbf{v}_k^{(t)} = \frac{\mathbf{u}_k^{(t)}}{N_k^{(t)}} \tag{9}$$

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:

$$L_{class}(\mathcal{B}_0) = -\frac{1}{C \cdot |\mathcal{B}_0|} \sum_{i=1}^{|\mathcal{B}_0|} \sum_{c=1}^{C} y^i_c \log\left(\hat{y}^i_c\right) \tag{10}$$

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:

$$L_{CREATE}(\mathcal{B}_0) = L_{recon}(\mathcal{B}_0) + \alpha L_{encoder}(\mathcal{B}_0) + L_{class}(r_0) \tag{11}$$

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

In this study, we implemented CREATE with "Pytorch" package[103]. 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[104] 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 divided all CREs into training, validation, and testing sets in an 8:1:1 ratio before data augmentation. Subsequently, data augmentation was performed independently within each set. This approach ensures that the augmented CREs from the same original CRE appear in only one set, and the augmented CREs across different sets are completely non-overlapping. 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. We obtained classification metrics, including auROC, auPRC, precision, recall and F1-score, by first calculating metrics for each class of CRE, and then finding the macro average across all classes. These metrics were calculated using "one-over-rest" mode, meaning that when calculating the metrics for a particular type of CRE, all other types of CREs and background samples are considered as the control group.

Feature spectrum

Supplementary Fig. 16a 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:

$$t_{ij} = \frac{c_{ij}}{\sum_{k \in \{1,\ldots,K\}} c_{ik}} \tag{12}$$

$$p_{ij} = \frac{t_{ij}}{\sum_{c \in \{1,\ldots,C\}} t_{cj}} \tag{13}$$

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.

$$p_j = \max_{i \in \{1,\ldots,C\}} p_{ij} \tag{14}$$

$$C_j = \operatorname*{argmax}_{i \in \{1,\ldots,C\}} p_{ij} \tag{15}$$

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 true CREs and predicted CREs by CREATE, we applied the tool FIMO[105] with default settings to scan a set of input sequences for searching known human TFs in the HOCOMOCO[106] 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 K562 and HepG2 cell types were obtained from ENCODE[107] (https://www.encodeproject.org/files/ENCFF867JRG/, https://www encodeproject.org/files/ENCFF721JMB/ and https://www.encodeproject.org/files/ENCFF064GJQ/, https://www.encodeproject.org/files/ENCFF369YQW/). Using BEDTools[108], we computed the methylation levels for true CREs and predicted CREs by CREATE.

Conservation scores computation. We downloaded the phastCons[72,109] scores for multiple alignments of 45 vertebrate genomes to the human genome from UCSC[110,111] (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phastCons46way/vertebrate.phastCons46way.bw). The phastCons scores for true CREs and predicted CREs were calculated via UCSC tool *bigWigAverageOverBed*[112].

Overlaps with promoter-capture HiC regions. The pcHiC data of K562 and HepG2 cell types were downloaded from NCBI[113] under accession number "GSE236305" and "GSE262496". Using BEDTools[108], we computed the numbers of pcHiC regions overlapping with true CREs and predicted CREs by CREATE.

Genetic variants analysis. We downloaded human all SNPs and 37,906,831 common SNPs from dbSNP[81–83] 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[77,78], GTEx database version 7 (https://www.gtexportal.org/home/downloads/adult-gtex). The common SNPs are based on germline origin and a minor allele frequency (MAF) of ≥0.01 in at least one major population. By excluding the common SNPs from all human SNPs, we obtained 596,252,807 rare SNPs. Using BEDTools[108], we considered the number of SNPs or eQTLs located in true CREs and predicted CREs by CREATE as the corresponding genetic variants levels.

Tissue enrichment analysis. To identify the tissues influenced by the identified risk loci within true CREs and predicted CREs by CREATE, we performed SNPsea analysis[87] with default settings. Based on the tissue-specific expression profiles of 17,581 genes across 79 human tissues (Gene Atlas[114]), we quantified the enrichments of these profiles on true CREs and predicted CREs, and displayed the top 30 significantly enriched tissues in the heatmaps.

Heritability enrichment analysis. To quantify the enrichment of heritability for blood- or liver-related phenotypes within true CREs and predicted CREs by CREATE, we conducted heritability enrichment analysis using partitioned LDSC[88] 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 true CREs and predicted CREs. Referring to the related studies[33,115,116], the GWAS summary statistics of cancer was downloaded from the Broad LD Hub and consisted of multiple types of cancer, including leukemia.

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 its original source code (https://deepsea.princeton.edu/). DanQ was implemented from its original source code repository (https://github.com/uci-cbcl/DanQ). ES-transition was implemented from its original source code repository (https://github.com/ncbi/SilencerEnhancerPredict). DeepICSH was implemented from its 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.

Reporting summary

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

Data availability

All datasets used in this study were obtained from public sources. We downloaded experimentally validated silencers from the SilencerDB database[89] (http://health.tsinghua.edu.cn/SilencerDB/), enhancers from the FANTOM5 project[24,26] (https://bioinfo.vanderbilt.edu/AE/HACER/), TSSs from the EPD database[90] (https://epd.expasy.org/epd), insulators from the ENCODE project[93,94] (https://www.encodeproject.org/files/ENCFF085HTY/; https://www.encodeproject.org/files/ENCFF237OKO/) for K562 and HepG2 cell types. We downloaded the histone modification ChIP-seq peaks and chromatin accessibility peaks from the Roadmap project[95] (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 K562 and HepG2 cell types. All regions in this study are either in the genome of GRCh37/hg19 or have been converted to GRCh37/hg19 by UCSC liftOver[117] tool. Source data are provided with this paper.

Code availability

The CREATE software, including detailed documents and tutorial, is freely available on GitHub (https://github.com/cuixj19/CREATE). Codes for the version of CREATE used in this paper are also deposited at Zenodo[118] (https://doi.org/10.5281/zenodo.15245829).

References

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 Prim.* **1**, 10 (2021).
7. Kadauke, S. & Blobel, G. A. Chromatin loops in gene regulation. *Biochim. Biophys. Acta* **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. *Nat. Rev. Mater.* **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 Proteom. Bioinform.* **19**, 565–577 (2021).
22. Min, X. et al. Predicting enhancers with deep convolutional neural networks. *BMC Bioinform.* **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. Zheng, A. et al. Deep neural networks identify sequence context features predictive of transcription factor binding. *Nat. Mach. Intell.* **3**, 172–180 (2021).
33. Cui, X. et al. Discrete latent embedding of single-cell chromatin accessibility sequencing data for uncovering cell heterogeneity. *Nat. Comput. Sci.* **4**, 346–359 (2024).
34. Van Den Oord, A. & Vinyals, O. Neural discrete representation learning. In *Proc. 31st Conference on Neural Information Processing Systems*. 6309–6318 (Curran Associates Inc., 2017).
35. Razavi, A., Van den Oord, A. & Vinyals, O. Generating diverse high-fidelity images with vq-vae-2. In *Proc. 33rd Conference on Neural Information Processing Systems*. 1331 (Curran Associates Inc., 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. Kingma, D. P. & Welling, M. Auto-encoding variational Bayes. In *Proc. 2nd International Conference on Learning Representations* (eds Bengio, Y. & LeCun, Y.) http://arxiv.org/abs/1312.6114 (ICLR, 2014).
38. McInnes, L., Healy, J. & Melville, J. Umap: uniform manifold approximation and projection for dimension reduction. *arXiv Prepr. arXiv* **1802**, 03426 (2018).
39. 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).
40. 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).

41. 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**, 109747 (2021).
42. 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).
43. Cowling, V. H. & Cole, M. D. Mechanism of transcriptional activation by the Myc oncoproteins. *Semin. Cancer Biol.* **16**, 242–252 (2006).
44. 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).
45. 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).
46. Ha, T. J. et al. Identification of novel cerebellar developmental transcriptional regulators with motif activity analysis. *BMC Genomics* **20**, 1–17 (2019).
47. Della Rosa, M. & Spivakov, M. Silencers in the spotlight. *Nat. Genet.* **52**, 244–245 (2020).
48. Ngan, C. Y. et al. Chromatin interaction analyses elucidate the roles of PRC2-bound silencers in mouse development. *Nat. Genet.* **52**, 264–272 (2020).
49. Cai, Y. et al. H3K27me3-rich genomic regions can function as silencers to repress gene expression via chromatin interactions. *Nat. Commun.* **12**, 719 (2021).
50. 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).
51. Oka, R. et al. Genome-wide mapping of transcriptional enhancer candidates using DNA and chromatin features in maize. *Genome Biol.* **18**, 1–24 (2017).
52. Zhu, Y. et al. Predicting enhancer transcription and activity from chromatin modifications. *Nucleic Acids Res.* **41**, 10032–10043 (2013).
53. 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).
54. Local, A. et al. Identification of H3K4me1-associated proteins at mammalian enhancers. *Nat. Genet.* **50**, 73–82 (2018).
55. Pekowska, A. et al. H3K4 tri-methylation provides an epigenetic signature of active enhancers. *EMBO J.* **30**, 4198–4210 (2011).
56. 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).
57. Whitfield, T. W. et al. Functional analysis of transcription factor binding sites in human promoters. *Genome Biol.* **13**, 1–16 (2012).
58. Boeva, V. Analysis of genomic sequence motifs for deciphering transcription factor binding and transcriptional regulation in eukaryotic cells. *Front. Genet.* **7**, 174397 (2016).
59. Inukai, S., Kock, K. H. & Bulyk, M. L. Transcription factor–DNA binding: beyond binding site motifs. *Curr. Opin. Genet. Dev.* **43**, 110–119 (2017).
60. 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).
61. 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).
62. 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).
63. Piper, M. et al. NFIA controls telencephalic progenitor cell differentiation through repression of the Notch effector Hes1. *J. Neurosci.* **30**, 9127–9139 (2010).
64. 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).
65. 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).
66. 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).
67. Baluapuri, A., Wolf, E. & Eilers, M. Target gene-independent functions of MYC oncoproteins. *Nat. Rev. Mol. Cell Biol.* **21**, 255–267 (2020).
68. 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).
69. Doni Jayavelu, N., Jajodia, A., Mishra, A. & Hawkins, R. D. Candidate silencer elements for the human and mouse genomes. *Nat. Commun.* **11**, 1061 (2020).
70. Smith, Z. D. & Meissner, A. DNA methylation: roles in mammalian development. *Nat. Rev. Genet.* **14**, 204–220 (2013).
71. Moore, L. D., Le, T. & Fan, G. DNA methylation and its basic function. *Neuropsychopharmacology* **38**, 23–38 (2013).
72. Siepel, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
73. Villar, D. et al. Enhancer evolution across 20 mammalian species. *Cell* **160**, 554–566 (2015).
74. Ovcharenko, I. et al. Evolution and functional classification of vertebrate gene deserts. *Genome Res* **15**, 137–145 (2005).
75. 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).
76. 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).
77. Lonsdale, J. et al. The genotype-tissue expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
78. Consortium, G. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
79. Maurano, M. T. et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
80. 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).
81. 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).
82. Sherry, S. T. et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
83. Smigielski, E. M., Sirotkin, K., Ward, M. & Sherry, S. T. dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res* **28**, 352–355 (2000).
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. 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).
87. 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).
88. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
89. Zeng, W. et al. SilencerDB: a comprehensive database of silencers. *Nucleic Acids Res.* **49**, D221–D228 (2021).
90. 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).
91. 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).
92. Nora, E. P. et al. Targeted degradation of CTCF decouples local insulation of chromosome domains from genomic compartmentalization. *Cell* **169**, 930–944. e922 (2017).
93. Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57 (2012).
94. Luo, Y. et al. New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Res.* **48**, D882–D889 (2020).
95. Consortium, R. E. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
96. Chen, S. et al. OpenAnnotate: a web server to annotate the chromatin accessibility of genomic regions. *Nucleic Acids Res* **49**, W483–W490 (2021).
97. 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).
98. Li, W., Wong, W. H. & Jiang, R. DeepTACT: predicting 3D chromatin contacts via bootstrapping deep learning. *Nucleic Acids Res* **47**, e60–e60 (2019).
99. Kaiser, L. et al. Fast decoding in sequence models using discrete latent variables. In *Proc. 35th International Conference on Machine Learning*. 2390–2399 (PMLR, 2018).
100. 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 (IEEE, 2021).
101. Williams, W. et al. Hierarchical quantized autoencoders. In *Proc. Adv. Neural Inf. Process. Syst.* **33**, 4524–4535 (2020).
102. Takida, Y. et al. Sq-vae: Variational bayes on discrete representation with self-annealed stochastic quantization. *arXiv Prepr. arXiv* **2205**, 07547 (2022).
103. Paszke, A. et al. Pytorch: An imperative style, high-performance deep learning library. In *33rd Conference on Neural Information Processing Systems* 721 (Curran Associates Inc., 2019).
104. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. Preprint at https://arxiv.org/abs/1412.6980 (2014).
105. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).
106. 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).
107. Zhang, J. et al. An integrative ENCODE resource for cancer genomics. *Nat. Commun.* **11**, 3696 (2020).
108. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
109. 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).
110. Kent, W. J. et al. The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
111. Nassar, L. R. et al. The UCSC genome browser database: 2023 update. *Nucleic Acids Res.* **51**, D1188–D1195 (2023).
112. 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).
113. Wheeler, D. L. et al. Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **36**, D13–D21 (2007).
114. Su, A. I. et al. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA* **101**, 6062–6067 (2004).
115. Tang, S. et al. scCASE: accurate and interpretable enhancement for single-cell chromatin accessibility sequencing data. *Nat. Commun.* **15**, 1629 (2024).
116. Chen, X. et al. Cell type annotation of single-cell chromatin accessibility data via supervised Bayesian embedding. *Nat. Mach. Intell.* **4**, 116–126 (2022).
117. Hinrichs, A. S. et al. The UCSC genome browser database: update 2006. *Nucleic Acids Res.* **34**, D590–D598 (2006).
118. Cui, X. CREATE: cell-type-specific cis-regulatory element identification via discrete embedding. *Zenodo* https://doi.org/10.5281/zenodo.15245829 (2025).

Acknowledgements

This work was supported by the National Key Research and Development Program of China grant nos. 2022YFF1202400 (H.L.) and 2023YFF1204802 (R.J.), the National Natural Science Foundation of China grants no. 62273194 (R.J.), the Beijing Natural Science Foundation grant no. L242026 (R.J.).

Author contributions

R.J. and W.Z. conceived the study and supervised the project. X.J.C. designed, implemented and validated CREATE. Q.Y., Z.G., Z.L., X.Y.C., S.C., Q.L. and H.L. helped analyze the results. X.J.C. and W.Z. wrote the manuscript, with input 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/s41467-025-59780-5.

Correspondence and requests for materials should be addressed to Wanwen Zeng or Rui Jiang.

Peer review information *Nature Communications* thanks Adam Naj 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.