communications biology

Article

https://doi.org/10.1038/s42003-024-06626-3

scHiCyclePred: a deep learning framework for predicting cell cycle phases from single-cell Hi-C data using multi-scale interaction information

Check for updates

Yingfu Wu[1,2,3,5], Zhenqi Shi[1,5], Xiangfei Zhou[1,5], Pengyu Zhang[3,5], Xiuhui Yang[1,5], Jun Ding[4] ✉ & Hao Wu[1,2] ✉

The emergence of single-cell Hi-C (scHi-C) technology has provided unprecedented opportunities for investigating the intricate relationship between cell cycle phases and the three-dimensional (3D) structure of chromatin. However, accurately predicting cell cycle phases based on scHi-C data remains a formidable challenge. Here, we present scHiCyclePred, a prediction model that integrates multiple feature sets to leverage scHi-C data for predicting cell cycle phases. scHiCyclePred extracts 3D chromatin structure features by incorporating multi-scale interaction information. The comparative analysis illustrates that scHiCyclePred surpasses existing methods such as Nagano_method and CIRCLET across various metrics including accuracy (ACC), F1 score, Precision, Recall, and balanced accuracy (BACC). In addition, we evaluate scHiCyclePred against the previously published CIRCLET using the dataset of complex tissues (Liu_dataset). Experimental results reveal significant improvements with scHiCyclePred exhibiting improvements of 0.39, 0.52, 0.52, and 0.39 over the CIRCLET in terms of ACC, F1 score, Precision, and Recall metrics, respectively. Furthermore, we conduct analyses on three-dimensional chromatin dynamics and gene features during the cell cycle, providing a more comprehensive understanding of cell cycle dynamics through chromatin structure. scHiCyclePred not only offers insights into cell biology but also holds promise for catalyzing breakthroughs in disease research. Access scHiCyclePred on GitHub at https:// github.com/ HaoWuLab-Bioinformatics/ scHiCyclePred.

The cell cycle is a highly complex process that involves dynamic changes in various cellular components, including RNA, DNA, and proteins[1–5]. To delve into the dynamics of the cell cycle, it is essential to analyze the relationship between the cell cycle phases and the state of these cellular components, which underpins this fundamental biological process. In this study, we focus on the cell cycle, a multifaceted biological process conventionally categorized into four distinct stages: G1 phase (the stage of cellular growth), Early-S phase (the early stage of DNA synthesis), Mid-S phase (the intermediate stage of DNA synthesis), and Late-S/G2 phase (the concluding stage of DNA synthesis). Notably, the three-dimensional structures of chromatin undergo dynamic variations across these stages, thereby exerting profound effects on gene expression patterns. In a previous study, Ye et al. found that changes in chromatin structure are relatively obvious during the S phase due to DNA replication[6]. Therefore, achieving a precise classification of these four cell cycle stages is paramount for gaining deeper insights into cellular functionality and regulatory mechanisms.

Fluorescence imaging is a powerful tool for understanding the relationship between subcellular processes and cellular behavior. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing (scRNA-seq) data has revealed hidden subpopulations of cells[7–15]. Previous

[1]School of Software, Shandong University, Jinan, Shandong, China. [2]Shenzhen Research Institute of Shandong University, Shenzhen, Guangdong, China. [3]College of Information Engineering, Northwest A&F University, Yangling, Shaanxi, China. [4]Department of Medicine, Meakins-Christie Laboratories, McGill University, Montreal, QC, Canada. [5]These authors contributed equally: Yingfu Wu, Zhenqi Shi, Xiangfei Zhou, Pengyu Zhang, Xiuhui Yang.
✉e-mail: jun.ding@mcgill.ca; haowu@sdu.edu.cn

studies have utilized fluorescence imaging to predict cell cycle phases, such as Ersoy et al.'s fast implementation of multi-phase graph partitioning active contours (fastGPAC)[16], which integrated fluorescently tagged feature distribution with a support vector machine (SVM) to predict cell cycle phases using images of the fluorescently tagged protein GFP-PCNA. Du et al.[17] developed a cell cycle phases classification algorithm based on 3D fluorescent images of the chromatin marker histone-GFP that extracted 3D intensity, shape, and texture information and combined weighted-SVM and neural network algorithms. Schonenberger et al.[18] proposed a workflow for classifying cell cycle phases in PCNA-immunolabeled cells using high-quality single-time-point images based on the unique patterns of PCNA distribution. Despite their effectiveness, these approaches are laborious, expensive, and low throughput, posing a significant challenge to research into the dynamic regulation of the cell cycle.

The advent of high-throughput single-cell technologies has significantly elevated the dimensions of single-cell data. This development has provided a perspective for deducing cell cycle phases. Among single-cell technologies, scRNA-seq data has emerged as an extremely useful tool for examining cellular heterogeneity in gene expression with unprecedented precision. Consequently, it has been widely employed for identifying cell cycle phases. A case in point is Hsiao et al.[19], who utilized scRNA-seq data analysis to characterize and infer quantitative cell cycle phases. Kowalczyk et al.[20] employed scRNA-seq data to uncover changes in cell cycle phases during the aging of hematopoietic stem cells.

The emergence of scHi-C technology has revolutionized the study of chromatin's three-dimensional structure[21]. The ability to determine cell cycle phases from scHi-C data is critical for analyzing and comprehending changes in chromatin's spatial structure during various cell cycle phases. This knowledge is indispensable for revealing cell cycle dynamics. However, accurately predicting cell cycle phases directly from scHi-C data remains a formidable challenge. Consequently, some studies have focused on constructing the pseudo-trajectory sequence of the cell cycle. For example, Nagano et al.[22] obtained scHi-C data from mouse embryonic stem cells (mESCs) at different cell cycle stages and utilized machine learning algorithms to calculate the 'repli-score', ratio of short-range connections and frequency of mitotic interactions for each cell. These indicators enabled the cells to follow the pseudo-trajectory of cell cycles.

Subsequently, we refer to this method and dataset as Nagano_method and Nagano_dataset, respectively. Liu et al.[23] presented a co-assayed scRNA-seq and scHi-C technology, along with datasets derived from complex tissues. Particularly, the dataset originated from developing mouse embryos, and we designate this dataset as Liu_dataset. To achieve precise cell-cycle phasing at the single-cell level, Liu et al. took advantage of the double-modality data of Hi-C and RNA-seq employed simultaneously (HiRES) and devised a strategy to assign single cells to different cell-cycle phases based on both DNA and RNA data. Each phase was characterized by distinct profiles of cell-cycle gene expression, DNA replication, and contact distribution. Ye et al.[6] proposed a method for constructing cell cycle pseudo-trajectories based on the combination of multiple feature sets called CIRCLET. The study generated four distinct feature sets and their combinations as inputs for CIRCLET, and the pseudo-trajectory sequence of single cells was reconstructed by calculating the distance between cells using the dimensionality reduction method (Wishbone), constructing the K-Nearest Neighbor (KNN) graph between cells based on the calculated distance, and dividing the pseudo-trajectory sequence into two semicircle trajectories using the KNN graph.

However, direct prediction of cell cycle phases using only the constructed cell cycle pseudo-trajectory sequence remains challenging, as it is necessary to know the number of cells contained in each cell cycle phase. Suppose that there are M cells in the G1 phase, N cells in the Early-S phase, P cells in the Mid-S phase, and Q cells in the Late-S/G2 phase. The precise implementation is to sort the pseudo-trajectory values corresponding to each cell in ascending order, with the prediction results for the 1 to M cells being the G1 phase, the $M+1$ to $M+N$ cells being the Early-S phase, the $M+N+1$ to $M+N+P$ cells being the Mid-S phase, and the $M+N+P+1$ to $M+N+P+Q$ cells being the Late-S/G2 phase. Despite the potential of constructing the pseudo-trajectory sequence of the cell cycle from scHi-C data, the lack of general datasets containing information on the number of cells in each cell cycle phase renders it impossible to predict the cell cycle phases of individual cells.

Furthermore, CIRCLET can solely derive pseudo-trajectory sequences for cells and cannot directly predict the cell cycle of individual cells, especially those at cycle boundaries. At the same time, the implementation of CIRCLET is not entirely automated, necessitating to set the starting point of the cell pseudo-trajectory sequences, which often requires some level of experience and trial-and-error to adjust. Secondly, the accuracy of this method in identifying cell cycle stages is relatively modest, indicating considerable potential for enhancement. Therefore, accurate and user-friendly computational methods for predicting cell cycle phases based solely on scHi-C data are urgently needed.

To overcome the hurdles of predicting cell cycle phases from scHi-C data, we present a computational framework, scHiCyclePred. This framework integrates three feature sets extracted from scHi-C data and employs a CNN model based on multi-feature fusion utilizing deep learning methods to predict cell cycle phases. In addition to the existing contact probability distribution versus genomic distance (CDD) feature set, we propose two additional feature sets: the bin contact probability feature set (BCP) and a small intra-domain contact probability (SICP) feature set, aimed at enhancing the accuracy of cell cycle phase prediction. Furthermore, we benchmark the performance of scHiCyclePred against existing methods and demonstrate that it outperforms them in predicting cell cycle phases. Finally, we analyze the changing patterns of chromatin's three-dimensional structure during the four cell cycle phases using a model interpretation approach. Overall, our proposed framework provides an accurate and user-friendly computational method for predicting cell cycle phases based solely on scHi-C data and sheds a light on understanding the dynamics of chromatin during the cell cycle.

Results

Overview of scHiCyclePred

The deep learning-based framework of scHiCyclePred consists of two crucial steps: the extraction of multiple feature sets and a CNN model based on multi-feature fusion (Fig. 1). The former extracts features of chromatin's three-dimensional structure from the scHi-C data based on multi-scale interaction information. This step involves extracting the following feature sets: (1) CDD feature set from the overall cellular interaction information, (2) BCP feature set from the overall chromatin interaction information, and (3) SICP feature set from the intra-domain interaction information on chromatin. To integrate the knowledge of multi-scale interactions in cells and intuitively predict the cell cycle stage, we develop a CNN model based on multi-feature fusion that integrates the three feature sets generated by the convolution module.

In the CNN model based on multi-feature fusion, three feature vectors for each cell are input into the model, which generates three vectors in parallel after passing through two convolution modules composed of a Conv1d layer, BatchNorm layer, Maxpool layer, and Dropout layer, followed by a flattening process. These three generated vectors are then merged into a single vector. The scores from different categories are mapped using a linear layer and "log_softmax", and the classification outcome is determined by the index with the highest score. In the following sections, we provide a detailed description of the workflows of the two steps in the scHiCyclePred framework.

Effectiveness evaluation of single feature set and multiple feature sets

To demonstrate the effectiveness of the multi-scale contact probability feature sets, we validate and analyze the performance of our extracted feature sets in this section. To accomplish this, we input each of the three feature sets into the network model independently and validate the classification performance obtained by using each feature set separately. To

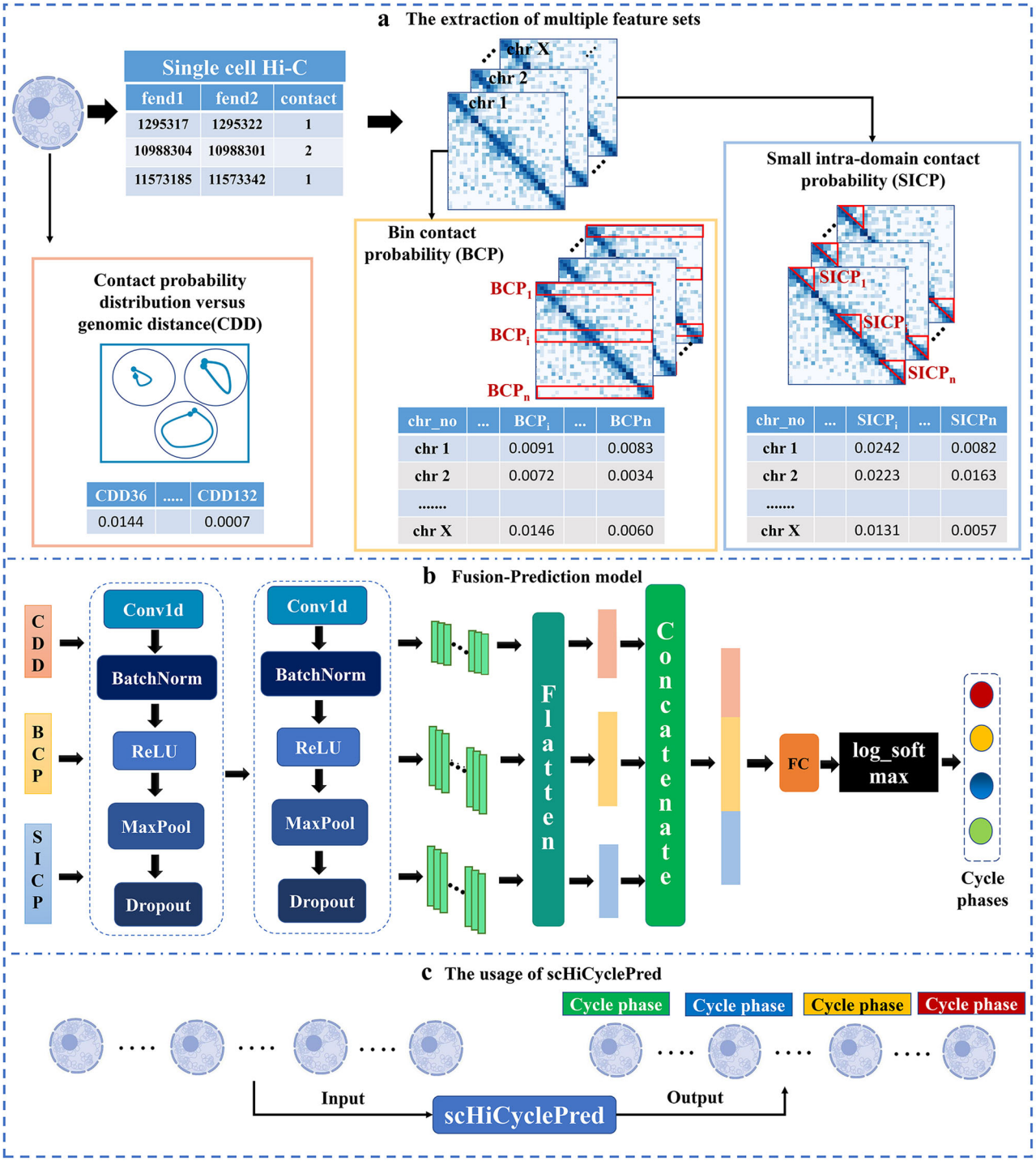

Fig. 1 | The framework of scHiCyclePred. a The extraction of multiple feature sets. scHiCyclePred combines read pair locus mapping file and chromatin interaction pair file to generate a unique chromosome contact matrices for each chromosome in every cell. To enhance cell cycle prediction accuracy and reveal variations in three-dimensional structure across different cell cycles, we extract features representing chromosome three-dimensional structure from diverse perspectives. Specifically, we extract three feature sets: contact probability distribution versus genomic distance (CDD), bin contact probability (BCP), and small intra-domain contact probability (SICP). **b** CNN model based on multi-feature fusion. We develop a deep learning model that combines convolution and feature fusion modules to accurately predict cell cycle phases. **c** The usage of scHiCyclePred. Directly apply the trained model to predict the cell data with unknown cell cycles.

ensure the robustness of our results, we partition the Nagano_dataset into fifty independent training and testing sets by altering the random seeds, with the Nagano_dataset being the total of each training and testing set. Subsequently, we evaluate the prediction performance of the three feature sets using four evaluation metrics: accuracy (ACC), F1 Score (F1), area under the receiver operating characteristic curve (AUC), and area under the precision-recall curve (Average Precision, AP). Specifically, the AUC value and AP value represent the area under the ROC curve and the area under the Precision-recall (PR) curve, respectively (Fig. 2a).

The results indicate that the five feature sets are effective in predicting the four cell cycle phases (Fig. 2b). Overall, the BCP and SICP features demonstrate superior accuracy and stability compared to the other three features. In contrast, the stability of the Insulation Score of Each Bin (INS) feature is relatively lower. Although the Pairs' Contact Coverage (PCC)

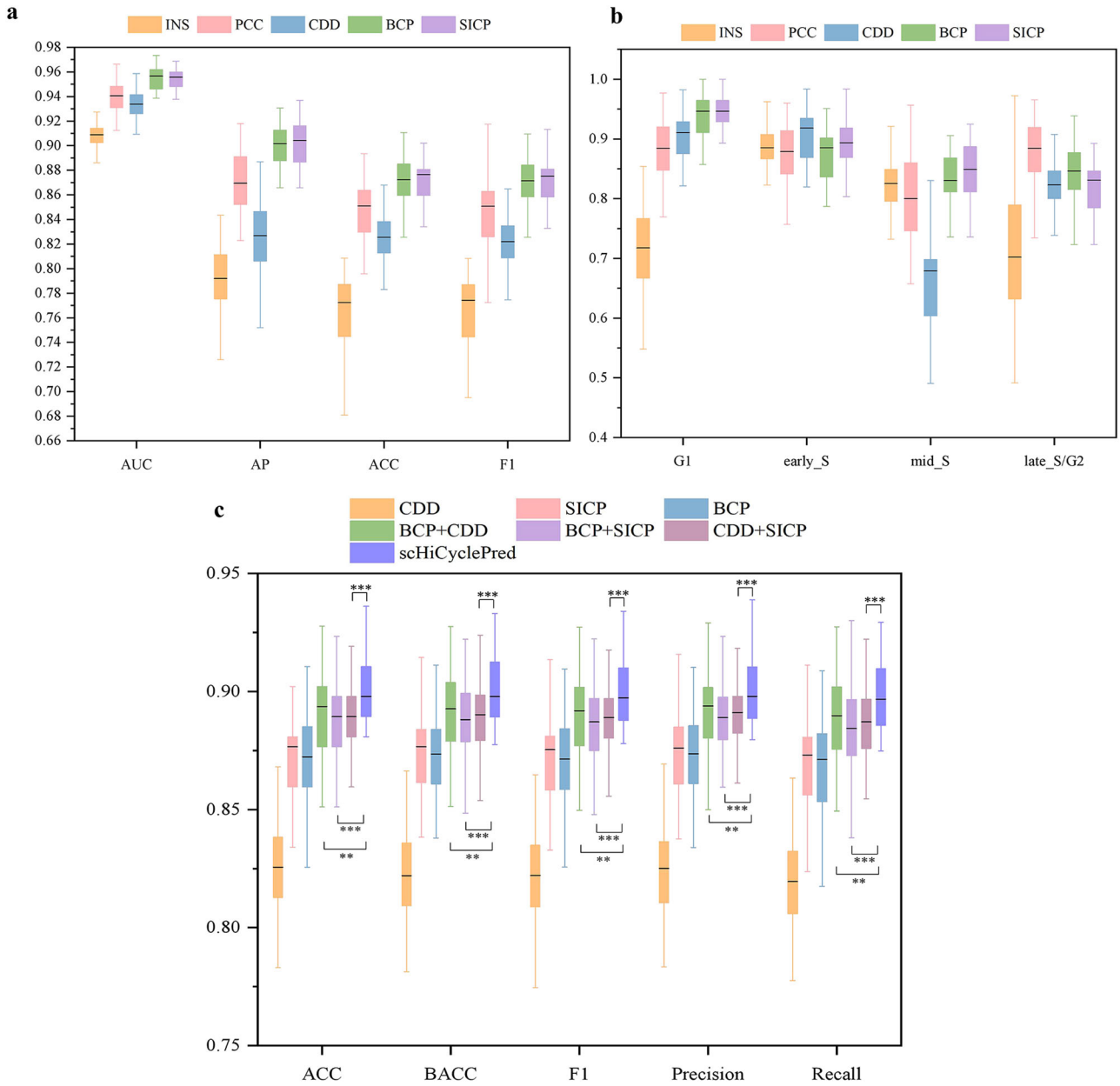

Fig. 2 | The effectiveness evaluation of the single feature set and the multiple feature sets. a Performance evaluation of five single feature sets. **b** The accuracy of five individual feature sets across cell cycle phases. **c** Performance comparison between single feature set and multiple feature sets. The lower and upper edges of the boxplot represent the minimum and the maximum values of the results, respectively. The bottom edge of the box represents the first quartile (Q1), while the top edge represents the third quartile (Q3). The median value is depicted by a black line. (*** indicates $P < 1 \times 10^{-3}$, ** indicates $P < 1 \times 10^{-2}$, * indicates $P < 5 \times 10^{-2}$).

feature occasionally exhibits higher accuracy than CDD, its concatenation of chromosome segment information at three resolutions (100 kb, 500 kb, 1 Mb) results in computationally expensive operations, especially in deep models, potentially significantly increasing the model's runtime. Additionally, the accuracy of the PCC feature is closely tied to the selection of hyperparameters, requiring different hyperparameters for different datasets, which demands a certain level of experience and methodology. Therefore, our model ultimately selects the relatively simpler CDD, BCP, and SICP features, which do not require manual parameter tuning, for further analysis in the subsequent sections. The detailed prediction results of the three feature sets for different phases show that: (1) In the G1, Early-S, and Late-S/G2 phases, the results of the three feature sets are similar; (2) In the Mid-S phase, the results of the SICP and BCP feature sets are significantly larger than those of the CDD feature set (Fig. 2b). The prediction performance of the Mid-S phase is not adequately captured based on the CDD feature set, indicating that the change in contact probability distribution is relatively minor. This observation aligns with the findings reported in CIRCLET.

To evaluate the effectiveness of our CNN model based on multi-feature fusion, we compare the performance of the three-feature-set fusion model with that of six other models constructed by retaining the corresponding feature extraction modules: three single-feature-set models (CDD, BCP, and SICP) and three two-feature-set fusion models (BCP-CDD, CDD-SICP, and BCP-SICP). We utilize the fifty independent training and testing sets mentioned in the previous section for this experiment. We use four evaluation metrics, namely ACC, F1, Precision, and balanced accuracy (BACC)[24–26], to assess the prediction performance of each model (Fig. 2c).

The effectiveness of our extracted features is demonstrated by the fact that the three single-feature-set models yield higher performance (Fig. 2c). Furthermore, feature fusion enhances the performance of the three two-feature-set models, highlighting the importance of our feature fusion. The

final results indicate that the CNN model based on multi-feature fusion exhibited the best performance in terms of ACC, F1, Precision, and BACC, with the least difference between the maximum and minimum values, i.e., the most stable result. Based on these comparison results, our CNN model based on multi-feature fusion can effectively fuse feature sets at all scales and deliver superior cell cycle phase prediction ability.

Performance evaluation of scHiCyclePred in predicting cell cycle phases

To demonstrate the superiority of our proposed scHiCyclePred method, we compare it to two cell cycle trajectory construction methods (Nagano_method and CIRCLET) using four evaluation metrics in this section. Although cell cycle trajectory construction is not primarily designed for predicting cell cycle phases, in order to evaluate Nagano_method's and CIRCLET's results efficiently, we generate predictive labels based on cell number and cycle order. As Nagano_method and CIRCLET methods do not require the division of training and testing sets, both methods performed predictions directly on the original Nagano_dataset. To compare scHiCyclePred with these two methods, we use them to calculate the average of the ACC, F1, and Precision metrics obtained from fifty testing sets. Additionally, we evaluate the effectiveness of our constructed deep learning model by employing three conventional machine learning methods, namely SVM[27], Logistic Regression[28], and Random Forest[29,30]. The same fifty independent training and testing sets mentioned earlier are utilized for this experiment (Fig. 3).

The results indicate that scHiCyclePred outperforms Nagano_method and CIRCLET in four metrics: ACC, F1, Precision, and Recall (Fig. 3a). Furthermore, scHiCyclePred still achieves optimal performance in terms of ACC, F1, Precision, BACC, and Recall metrics compared to using the three conventional machine learning methods (Fig. 3b). These findings demonstrate that the scHiCyclePred method achieves superior performance in predicting cell cycle phases.

In addition to comparing scHiCyclePred with CIRCLET on the Nagano_dataset, we also evaluated its performance on the complex tissue dataset (Liu_dataset) to validate its generalization and stability. It is worth noting that CIRCLET incorporates PCC features, and the choice of threshold significantly impacts its performance. Therefore, we conducted experiments to assess performance across different thresholds and selected the best-performing result for comparison (Supplementary Table 1). Additionally, to evaluate the effectiveness of our deep learning model, we employed three traditional machine learning methods: SVM, Logistic Regression (LR), and Random Forest (RF). These experiments utilized the same 50 independent training and testing sets as previously mentioned (Fig. 3).

The results indicate that scHiCyclePred achieved an accuracy of 0.76, surpassing CIRCLET's accuracy of 0.37 by 0.39. Additionally, scHiCyclePred outperformed the best-performing method on the other three metrics by a significant margin (Fig. 3c). It is noteworthy that our model's performance metrics substantially exceeded those of the three machine learning methods (Fig. 3d). In conclusion, this suggests that our model maintains optimal and stable performance across different datasets.

Robustness validation of scHiCyclePred

We evaluate the scHiCyclePred model and three machine learning methods using two distinct approaches to validate scHiCyclePred's robustness on datasets for the following purposes: (1) Validating the effectiveness of scHiCyclePred on imbalanced datasets by downsampling the original Nagano_dataset. (2) Testing the effectiveness of scHiCyclePred on drop-processed datasets. For the drop experiment, the Nagano_dataset's chromatin interaction pair file (raw_data file) is initially split into a training set ($train_1$) and a testing set ($test_1$) at a ratio of 80% to 20%, ensuring that the sum of $train_1$ and $test_1$ is equal to the number of raw_data_file. The model is then trained with $train_1$ to produce $model_1$, which is subsequently tested with $test_1$ The random seed is then altered, and four additional sets of training and testing data are generated in the same manner. Next, for each set of testing set, 5%y to 50%y (with an increment of 5%) of rows of data are randomly selected, where y represents the total number of rows in the testing set. Taking into account that interaction pairs with contact numbers of 1 or 2 account for 98.8% of the total number of interaction pairs (details regarding the number of interaction pairs with varying contact numbers, as well as their corresponding proportions are provided in Supplementary Fig. 1), we adjust the number of contacts between two segments (counting columns) for the selected rows using the following procedure: randomly adding or subtracting one from the number of contacts. Subsequently, each training set is matched to 10 testing sets, resulting in a total of 50 testing sets across the five training sets. Each group of experiments is trained using its corresponding training set, and only the 10 testing sets in the same group as the current training set are utilized to test the trained model to prevent data leakage. Taking all aspects into account, the comparison results of ACC, F1 score, Precision, and BACC metrics demonstrate the effectiveness of scHiCyclePred on the drop-processed datasets (Fig. 4a).

For the downsampling experiment, we partition the Nagano_dataset into nine imbalanced datasets (labeled A to I) composed of different proportions (Fig. 4b). Using a ratio of 8:2, each of the nine datasets is split into two parts: a training set and a testing set. The training set is used to train the model, whereas the testing set is used to evaluate its performance. The comparison results of the four evaluation metrics (i.e., ACC, F1, Precision, and BACC) demonstrate the effectiveness of scHiCyclePred on imbalanced datasets (Fig. 4c, d).

Overall, these results underscore that scHiCyclePred outperforms other methods and exhibits robustness on drop-processed datasets (Fig. 4a). Moreover, scHiCyclePred demonstrates robust stability and generalization across various imbalanced datasets. In contrast, the three machine learning methods failed to achieve comparable results. (Fig. 4c, d and Supplementary Fig. 2). In summary, the results from both experiments demonstrate the effectiveness and robustness of scHiCyclePred in predicting cell cycle phases on various datasets.

Analysis of chromatin change patterns across various cell cycle phases

We further investigate the pattern of changes in the 3D structure of chromatin during different cell cycle phases. Specifically, we utilize the SHapley Additive exPlanations (SHAP) method[27,29] to analyze the feature importance of the three feature sets during the four cycles. The SHAP graph displays the vertical axis representing features and the dots representing samples, with redder colors indicating higher feature values and bluer colors indicating lower feature values. The horizontal axis demonstrates the SHAP value, where positive values indicate a positive effect on the prediction, and negative values indicate a negative effect.

Regarding the CDD feature set, our analysis is centered on the top 20 significant features. The evaluation of the importance of the CDD feature set reveals that 12 features are prominently represented within the top 20 features across the four cell cycle phases: $CDD_{61\text{-}70}$, CDD_{72}, and CDD_{74} (Fig. 5). In the G1, Early-S, and Mid-S phases, there is a tendency for bluer dots to cluster in the positive semi-axis, indicating a positive effect, whereas during the Late-S/G2 phase, redder dots predominantly cluster in the positive semi-axis. This observation suggests that the proportion of short-distance chromatin interactions gradually increases with cell cycle progression, leading to a progressive tightening of the three-dimensional chromosome structure. Notably, this finding is consistent with the results reported by Ye et al[6]. In addition, CDD_{111}, CDD_{117}, and CDD_{119} appear in the 20 most important features of Early-S and Late-S/G2 cycles, indicating that there may be more significant changes in long-distance contacts from Early-S to Late-S/G2 cycles.

For the BCP feature set and SICP feature set, we primarily analyze the top 50 important features. The results of the BCP feature set's importance evaluation reveal that for fragments on different chromatin, their three-dimensional structure change patterns vary during the cell cycle phases. Based on the varying contact numbers, the evaluation of the BCP feature set's importance indicates the following: (1) BCP_{41} of chr15 and BCP_{64} of

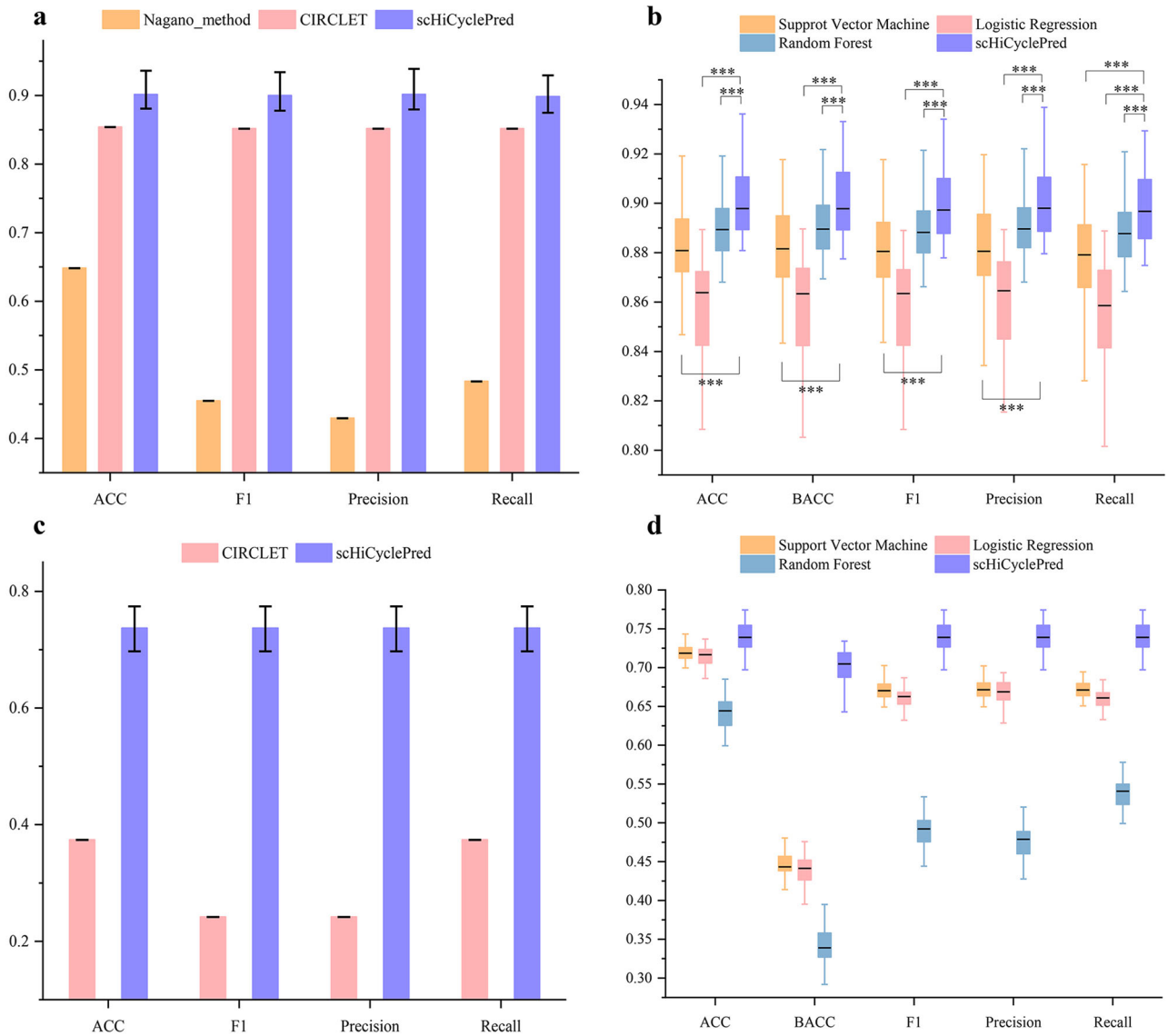

Fig. 3 | Performance comparison of scHiCyclePred with other methods on different datasets. a Performance comparison of scHiCyclePred with Nagano_method and CIRCLET on Nagano_dataset. Each bar in the graph represents the average value, while the bottom of the vertical line indicates the minimum value, and the top represents the maximum value. **b** Performance comparison of scHiCyclePred with other machine learning methods on Nagano_dataset. **c** Performance comparison of scHiCyclePred with CIRCLET on Liu_dataset. The bar in the graph represents the average value, while the bottom of the vertical line indicates the minimum value, and the top represents the maximum value. **d** Performance comparison of scHiCyclePred with other machine learning methods on Liu_dataset. The lower and upper edges represent the minimum value and the maximum value of the results, respectively. The bottom edge of the box represents the first quartile (Q1), the top edge of the box represents the third quartile (Q3), and the black line represents the median value. (*** indicates $P < 1 \times 10^{-3}$).

chr4: In the G1 phase, the majority of bluer colors are concentrated in the negative range, indicating a negative effect. In contrast, during the Early-S phase, most bluer colors are concentrated in the positive range, signifying a positive effect. Overall, this suggests that during the Early-S phase, the chromatin contact numbers decrease compared to the G1 phase, indicating a gradual loosening of its three-dimensional structure; (2) BCP_{98} of chr6: In the G1 phase, the majority of bluer colors are concentrated in the positive range, indicating a positive effect. In contrast, during the Early-S phase, most bluer colors are concentrated in the negative range, signifying a negative effect. Overall, this suggests that during the Early-S phase, the chromatin contacts increase compared to the G1 phase, indicating a gradual tightening of its three-dimensional structure; (3) BCP_{142} of chr5: In the G1 phase, the majority of bluer colors are concentrated in the negative range, indicating a negative effect. In contrast, during the Early-S, Mid-S, and Late-S/G2 phases, most bluer colors are concentrated in the positive range, signifying a positive effect. In summary, we find that the chromatin contact numbers in the remaining three phases gradually decrease compared to the G1 phase, indicating a gradual loosening of their three-dimensional structure (Fig. 6a–d).

The results of the importance evaluation for the SICP feature set reveal the following insights: (1) During the transition from the G1 to the Late-S/G2 phases, the majority of bluer color clusters gradually transform into redder color clusters and eventually transition back to bluer color clusters. This suggests that their three-dimensional structure undergoes a transition from loosening to tightening, subsequently followed by a trend towards loosening again. This phenomenon is particularly evident in the analysis of $SICP_{127}$ and $SICP_{128}$ of chr10; (2) The three-dimensional structure change patterns of adjacent bins exhibit similarities. For instance, $SICP_{3\text{-}5}$ of chr11 and $SICP_{87}$ and $SICP_{88}$ of chr12, etc demonstrate similar patterns of change; (3) From the perspective of the three-dimensional structural changes in the bin neighborhood, the prominent contact patterns of the Mid-S cycle and the Late-S/G2 cycle are relatively similar (Fig. 7a–d). One possible reason for

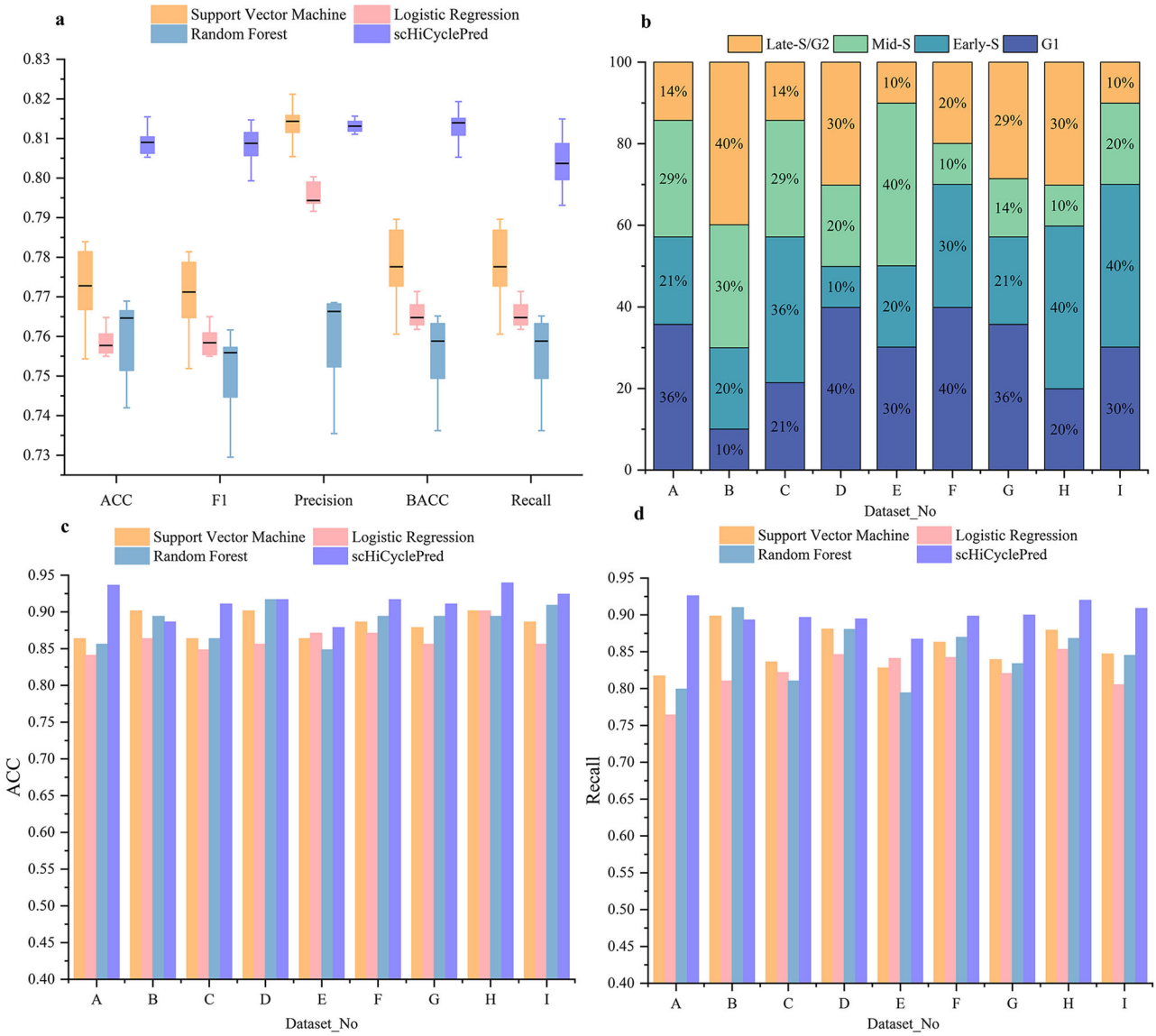

Fig. 4 | Performance of scHiCyclePred on imbalanced datasets and drop-processed datasets. a Performance of scHiCyclePred and machine learning methods on the drop-processed datasets. The lower and upper edges of the boxplot represent the minimum and the maximum values of the results, respectively. The bottom edge of the box represents the first quartile (Q1), the top edge represents the third quartile (Q3), and the black line represents the median value. **b** Distribution of cells across cell cycle phases in nine imbalanced datasets (A-I). Further performance evaluation is conducted on this dataset in Fig. 4c, d. **c**, **d** Evaluation of scHiCyclePred and machine learning methods on an imbalanced dataset using ACC and Recall metrics.

(3) may be that the degree of three-dimensional structure change of chromatin gradually decreases from the Mid-S cycle to the Late-S/G2 cycle. This finding is consistent with previous findings[6].

In this study, we meticulously annotate the patterns associated with each stage of the cell cycle and link them to specific genes based on feature importance ranking. The number of these genes varies depending on the number of genes corresponding to the features. Specifically, as the CDD feature set captures information from all chromosomes and cannot extract corresponding chromosome segments (bin), we proceed to rank and select the top 50 features based on feature importance in the BCP and SICP features. These selected features correspond to chromosome segments (bin) that map to all genes in the reference genome, with each bin having a size of 1 Mb. To validate the accuracy of these patterns, we conduct an extensive analysis of relevant research literature concerning the cell cycle. Our findings are as follows: (1) In the G1 phase, we discover a significant correlation between the gene lists from papers[31–33] and the genes annotated by our patterns, with the same p-value of 0.005858. (2) During the Early-S phase, we observe a strong correlation between the gene lists in papers[34–36] and the genes annotated through our patterns, with p values of 2.388×10^{-5}, 0.0003518, and 0.001689. (3) Similarly, for the Mid-S phase, there is a notable correlation between the gene lists from papers[37–39] and our annotated genes, with p-values of 9.535×10^{-5}, 0.00224, and 0.003723. (4) Finally, in the Late-S/G2 phases, we find a substantial correlation between the gene lists in papers[36,40,41] and our annotated genes, with p values of 4.067×10^{-6}, 0.0001014, and 0.0006912. In summary, our analysis consistently demonstrates a strong association between our annotated patterns and the cell cycle, reaffirming the precision and relevance of our approach.

Discussion

In this study, we introduce a method called scHiCyclePred for predicting cell cycle phases using scHi-C data. Our method involves the extraction of multiple feature sets based on scHi-C data and the development of a CNN model based on multi-feature fusion for cell cycle phase prediction. Three feature sets are extracted, including the existing CDD feature set (overall cellular interaction information) and two feature sets: the BCP feature set (overall chromatin interaction information) and the SICP feature set (intra-

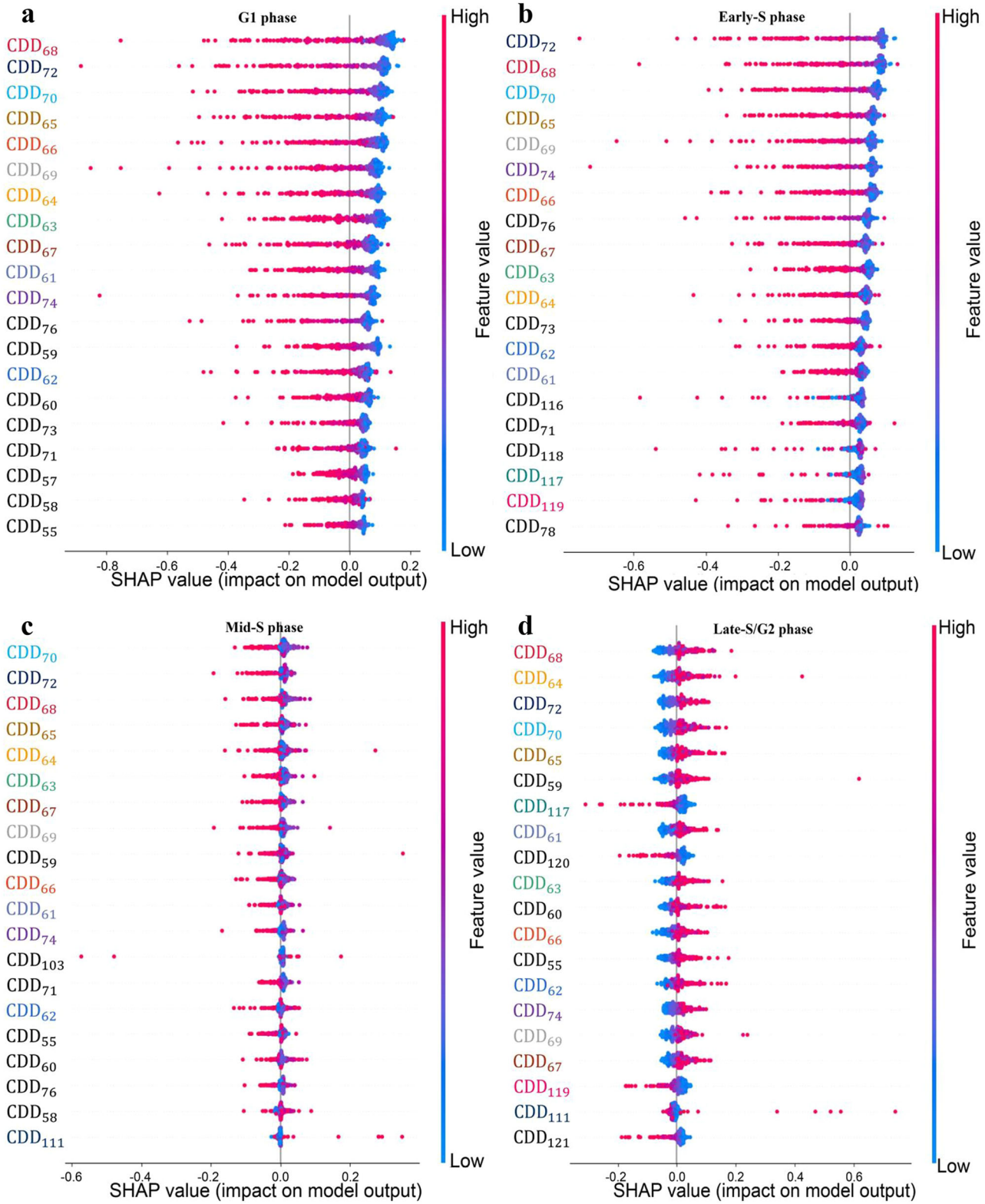

Fig. 5 | Importance evaluation results of the first 20 features in the CDD feature set. a G1 phase. **b** Early-S phase. **c** Mid-S phase. **d** Late-S/G2 phase. The SHAP graph displays the SHAP value on the horizontal axis, representing the assigned importance to each feature in the given sample. The vertical axis represents features, and the dots represent samples, with redder colors indicating higher feature values and bluer colors indicating lower feature values. We analyze the top 20 significant features of the CDD feature set and find that 12 features are present across all four cell cycle phases: $CDD_{61\text{-}70}$, CDD_{72}, and CDD_{74}. Features with the same color in the graph represent identical features.

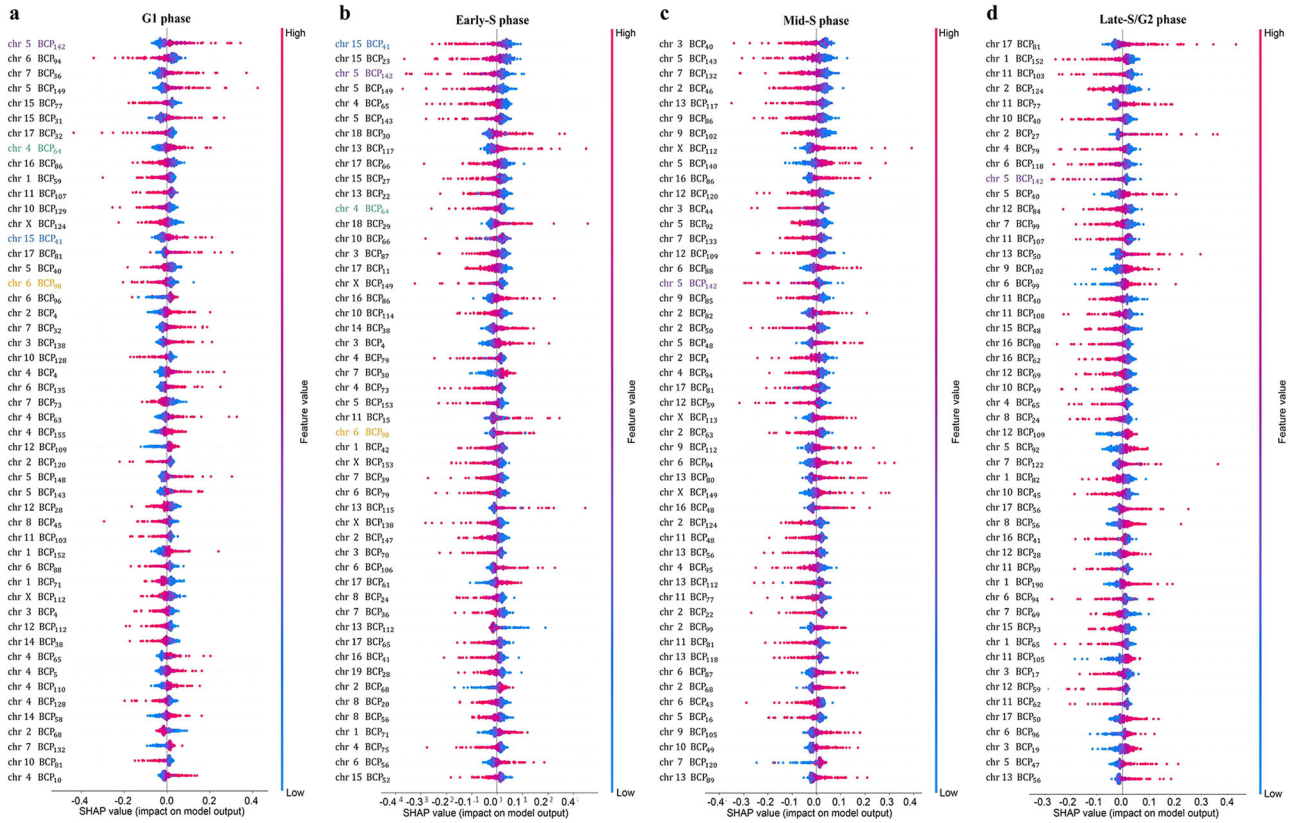

Fig. 6 | Importance evaluation results of the first 50 features in the BCP feature set. a BCP feature set in G1 phase. **b** BCP feature set in Early-S phase. **c** BCP feature set in Mid-S phase. **d** BCP feature set in Late-S/G2 phase. The SHAP graph displays the SHAP value on the horizontal axis, representing the assigned importance to each feature in the given sample. The vertical axis represents features, and the dots represent samples, with redder colors indicating higher feature values and bluer colors indicating lower feature values. Features with the same color in the graph represent identical features.

domain interaction information). Our CNN model based on multi-feature fusion successfully combines the three feature sets and improves prediction performance. Furthermore, we evaluate the CNN model based on multi-feature fusion using ablation experiments and compare the performance of scHiCyclePred with other popular methods, including Nagano_method, CIRCLET, LR, SVM, and RF, in predicting cell cycle phases. Our results demonstrate that the scHiCyclePred method has good performance in predicting cell cycle phases and is more robust than existing scHi-C data-based cell cycle phase prediction methods.

In addition, by evaluating the impact of different features and considering their characteristics, we analyze the patterns of chromosome three-dimensional structure changes across various cell cycle phases. Furthermore, our research results are consistent with prior studies. Importantly, our analysis of different features reveals variations or trends in the three-dimensional chromosome structure between different cell cycle phases, thereby offering a perspective for understanding chromatin dynamics during the cell cycle. However, we note that current scHi-C data is biased due to the coverage consistency of current scHi-C techniques, which hinders the unraveling of the relationship between cell cycle dynamics and three-dimensional structural patterns of chromatin. Therefore, in the future, it will be necessary to address these biases present in scHi-C data and further incorporate our method for cell cycle phase prediction. Overall, our study provides a promising approach for predicting cell cycle phases using scHi-C data, and further research in this field is needed to fully realize the potential of this method. The scHiCyclePred method offers an accurate and user-friendly computational approach to predict cell cycle phases based solely on scHi-C data, and it provides insights into understanding the dynamics of chromatin during the cell cycle.

Materials and methods

Data preparation

The scHi-C data used in this study are obtained from the study by Nagano et al.[22], which includes scHi-C data from 1171 mESCs labeled according to their cell cycle phase using fluorescence-activated cell sorting (FACS). The cells are classified into four phases: 280 cells in the G1 phase, 303 cells in the Early-S phase, 262 cells in the Mid-S phase, and 326 cells in the Late-S or G2 phase. The additional scHi-C data utilized in this study are obtained from the study by Liu et al.[24], encompassing scHi-C data from 6288 mESCs that are also labeled based on their cell cycle phase using FACS. The cells are categorized into four phases: 1606 cells in the G1 phase, 766 cells in the Early-S phase, 1688 cells in the Mid-S phase, and 2228 cells in the Late-S or G2 phase.

Our study utilizes performance metrics such as ACC, BACC, AUC, etc., to assess the performance of various methods (Supplementary Methods). The evaluation of all methods' performance is conducted using the known cell cycle phases (G1, Early-S, Mid-S, and Late-S/G2) labeled by FACS. For the Nagano_dataset, raw data files are downloaded from https://github.com/tanaylab/schic2?tab=readme-ov-file, containing chromatin interaction pair files and read pair locus mapping files. The chromatin interaction pair file contains information such as the sequence number of chromatin fragment pairs and the count of interactions, while the read pair locus mapping file contains the relationship between chromatin fragment sequence number, chromatin information, and precise position information. Within the Liu_dataset, raw data files are downloaded from NCBI GEO (accession number GSE223917), which also includes files detailing chromatin interaction pairs.

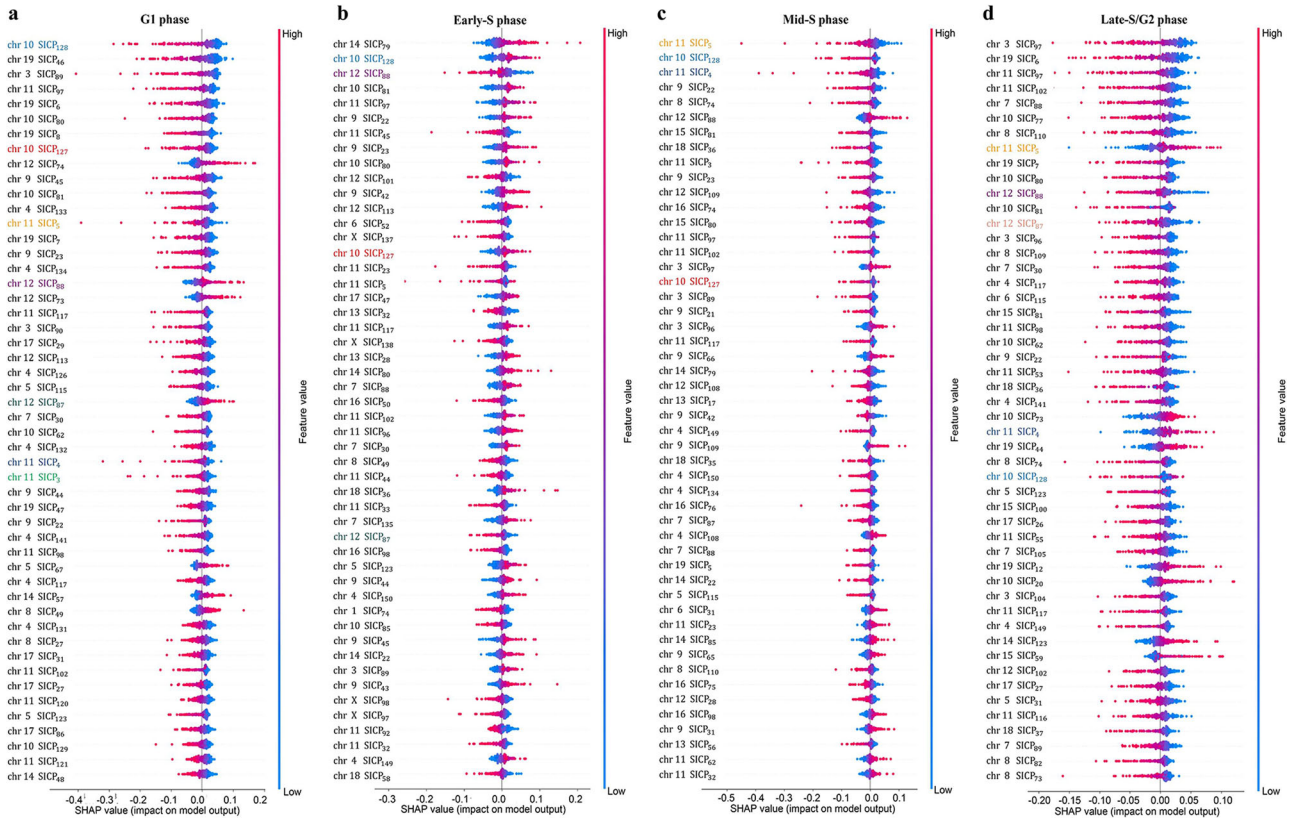

Fig. 7 | Importance evaluation results of the first 50 features in the SICP feature set. a SICP feature set in G1 phase. **b** SICP feature set in Early-S phase. **c** SICP feature set in Mid-S phase. **d** SICP BCP feature set in Late-S/G2 phase. The SHAP graph displays the SHAP value on the horizontal axis, representing the assigned importance to each feature in the given sample. The vertical axis represents features, and the dots represent samples, with redder colors indicating higher feature values and bluer colors indicating lower feature values. Features with the same color in the graph represent identical features.

Most current computational methods for scHi-C data represent the data as multiple chromatin contact matrices with a given resolution[42–50]. However, generating a chromatin contact matrix requires specific information about interacting chromatin fragments in each cell, which is not contained in the raw data files. To address this limitation, we combine the read pair locus mapping file and chromatin interaction pair file to generate a unique chromatin interaction information file for each chromatin in each cell. Next, we divide each chromatin in the cell into multiple segments called bins based on a specified resolution R. This allows us to represent the scHi-C data as a matrix of bin-pair interactions, which can be used as input for our proposed scHiCyclePred framework. Finally, we generate the chromatin contact matrix C by mapping the interaction information from the chromatin interaction information file to the corresponding bin. However, since Nagano et al. could not capture interaction information for chromatin Y, we generate 20 chromatin contact matrices (C_1,C_2,…,C_{19},C_X) per cell to cover all possible combinations of chromatin interactions. This allows us to effectively capture the chromatin interaction patterns and generate accurate representations of the scHi-C data for use in our proposed scHiCyclePred framework.

Contact probability distribution versus genomic distance feature set

Nagano et al. discovered that the contact probability based on linear distance division showed different states in various cell cycle phases, and Ye et al. proposed the contact probability distribution versus genomic distance feature set to determine the cell cycle trajectory[6,23]. Therefore, we employ the contact probability distribution versus genomic distance (CDD) as a feature set in our framework. It is important to note that the CDD feature set is extracted from the cell-wide interaction information rather than the chromatin contact matrix. In this feature set, the interaction pairs are allocated into intervals based on linear distance, with the linear distance range represented by each interval gradually increasing. The mapping formula for the interaction pairs is as follows:

$$interval_{loc} = floor\left(\frac{log_2(loc_1 - loc_2) + s}{s}\right) \quad (1)$$

where loc_1 and loc_2 represent the positions of two interacting fragments on the chromatin, and s denotes the exponential step length of each interval. Based on the study by Nagano et al. and Ye et al.[6,22], we set $s = 0.125$ and the distance range for genes to be [2 K, 9.3 M]. After assigning all contacts, the contact probability distribution versus genomic distance feature set is extracted by calculating the probability of the contact count in each interval as shown in Eq. 2.

$$CDD_{loc} = \frac{Total(interval_{loc})}{Total(Cell)} \quad (2)$$

where $Total(interval_{loc})$ represents the contact count in the corresponding bin, and $Total(Cell)$ represents the total number of chromatin contacts in a cell.

Bin contact probability feature set

The contact frequency distribution of the same site on the same chromatin varies in different cell cycle phases[51–53]. Therefore, we use the bin contact probability (BCP) set of the chromatin as the contact feature set for chromatin in our framework. Specifically, we use the BCP values in the chromatin contact matrix, rather than the CDD feature set, to extract this set of features. We then use the contact feature set of all chromatin in the cell as one of the extracted feature sets. The chromatin contact matrix is C_{n*n},

where n is the number of bins on the chromatin. The bin contact probability of the chromatin is calculated as follows:

$$BCP_i = \frac{\sum_{j=1}^{n} C_{ij}}{Total(chr)} \quad (3)$$

where $Total(chr)$ represents the total contact count of chromatin at the location of bins and the value range of i is [1, n]. The feature set of bin contact probability for each cell is generated by splicing the set of bin contact probability for all chromatin.

Small intra-domain contact probability feature set

Nagano et al. found that INS is affected by cell cycle dynamics[22]. INS shows the depletion of chromatin interaction information in a domain centered on the target bin. Motivated by this finding, we extract the contact probability of small intra-domain (SICP) on the chromatin as one of the feature sets in our framework. The small domain is defined as the region centered on the target bin and bounded by the adjacent first-order linear bins. Specifically, we calculate the SICP values for each bin by dividing the number of contacts within the small domain by the total number of contacts in the bin. This results in a vector that represents the SICP feature set for the chromatin bin. On the chromosome contact matrix C_{n*n}, the contact probability of a small intra-domain is calculated similarly to the bin contact probability as follows:

$$SICP_i = \frac{\sum_{i=k-1}^{k+1} \sum_{j=i}^{k+1} C_{ij}}{Total\left(C_{n*n}\right)} \quad (4)$$

where C_{k*k} represents the currently calculated chromatin domain center and the value range of k is [1, n]. It is important to note that to calculate the SICP features, it is necessary to fill matrix C with 0 elements, transforming it into matrix B. This enables the formation of a complete small domain, even when k is equal to 1 and n. Similar to BCP, we integrate multiple sets of small intra-domain contact probabilities for entire chromatin as the SICP feature set of a mouse cell.

CNN model based on multi-feature fusion

In this section, we present a deep learning-based CNN model that leverages multi-feature fusion to accurately predict cell cycle phases. This model incorporates convolution and feature fusion modules into its architecture. The convolution module generates identical network models for the three feature sets CDD, BCP, and SICP. This module consists of a CNN layer, batch norm layer, max pooling layer, and dropout layer which are stacked twice to generate more complicated features. The CNN layer uses a one-dimensional convolution kernel with a kernel size of 7 and a channel size of 32 to collect features from various input feature sets. The batch norm layer prevents gradient explosion and disappearance, while the max pooling layer reduces feature dimension, preserves key features, scales back model calculations, avoids overfitting, and enhances generalizability. We use ReLU as the activation function to connect the batch norm layer and max pooling layer[54], adding nonlinear components to enhance the model's expression capability. The dropout layer effectively prevents model overfitting by discarding some neurons during forward propagation with a predetermined probability. Finally, a flattening operation is applied to the data produced from the second dropout layer to combine the data from all channels into a vector.

In the feature fusion module, the three vectors corresponding to the three feature sets generated by the convolution module are combined into a single vector. The scores from various categories are then mapped using a linear layer, followed by the "log_softmax" function[55–57]. The ultimate classification outcome corresponds to the index with the highest score. Using the "log_softmax" function avoids the value overflow problem and facilitates the calculation of the loss function. To address the issue of imbalanced samples and enhance the overall performance of the model, we adopt the focal loss function[58,59] as the loss function of the model. The focal loss is calculated as in Eq. 5.

$$FL(p_t) = -\alpha_t (1 - p_t)^{\gamma} log(p_t) \quad (5)$$

where the value range of γ is [0,5], p_t means the probability that the model predicts the current sample as phase t, and $-log(p_t)$ is utilized to calculate the cross entropy loss. If the p_t corresponding to the current sample phase is smaller, it means that the prediction result of the model is more inaccurate, then the coefficient $(1 - p_t)^{\gamma}$ of the difficult sample will increase, and the difficult sample will lose. α_t represents the weight coefficient corresponding to phase t, and the value is the number of cells contained in phase t. This model is trained by minimizing the focal loss.

Additionally, to prevent overfitting of the model to the training set during model building, we employ an early stopping[60–64] mechanism and a 5-fold cross-validation approach. The model training is terminated if the loss of the model on the validation set does not reduce for 10 consecutive epochs (Fig. 1 and Supplementary Fig. 3).

Statistics and reproducibility

All statistical analyses were conducted using Python version 3.9.15. scHiCyclePred was tested and successfully executed on the independent server, and the same results were produced as with the original experiments.

Reporting summary

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

Data availability

Nagano_dataset is downloaded from https://github.com/tanaylab/schic2?tab=readme-ov-file. Liu_dataset is downloaded from NCBI GEO (accession number GSE223917). Numerical source data for graphs presented in the main figures can be found in Supplementary Data.

Code availability

The source code for scHiCyclePred is freely available on GitHub (https://github.com/HaoWuLab-Bioinformatics/scHiCyclePred) and its Zenodo (https://doi.org/10.5281/zenodo.12721771)[65].

Received: 30 October 2023; Accepted: 24 July 2024;
Published online: 31 July 2024

References

1. Israels, E. D. & Israels, L. G. The cell cycle [J]. *Oncologist* **5**, 510–513 (2000).
2. Barron, M. & Li, J. Identifying and removing the cell-cycle effect from single-cell RNA-Sequencing data [J]. *Sci. Rep.* **6**, 33892 (2016).
3. Caldon, C. E., Sutherland, R. L. & Musgrove, E. A. Cell cycle proteins in epithelial cell differentiation: implications for breast cancer [J]. *Cell Cycle* **9**, 1918–1928 (2010).
4. Raj, A., Tyagi, S., van den Bogaard, P., Rifkin, S. A. & van Oudenaarden, A. Imaging individual mRNA molecules using multiple singly labeled probes. *Nat. Methods* **5**, 877–879 (2008).
5. Setty, M. et al. Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat. Biotechnol.* **34**, 637–645 (2016).
6. Ye, Y., Gao, L. & Zhang, S. Circular trajectory reconstruction uncovers cell-cycle progression and regulatory dynamics from single-cell Hi-C maps [J]. *Adv. Sci.* **6**, 1900986 (2019).
7. Buettner, F. et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells [J]. *Nat. Biotechnol.* **33**, 155–160 (2015).
8. Wang, Y. et al. A multi-view latent variable model reveals cellular heterogeneity in complex tissues for paired multimodal single-cell data. *Bioinformatics.* **39**, btad005 (2023).
9. Tuch, B. B. et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).

10. Bendall, S. C. et al. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science* **332**, 687–696 (2011).
11. Liu, Z. et al. Reconstructing cell cycle pseudo-time-series via single-cell transcriptome data[J]. *Nat. Commun.* **8**, 22 (2017).
12. Ji, Z. & Ji, H. TSCAN: pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Res.* **44**, e117 (2016).
13. Trapnell, C. et al. Pseudo-temporal ordering of individual cells reveals dynamics and regulators of cell fate decisions. *Nat. Biotechnol.* **32**, 381–386 (2014).
14. Bendall, S. et al. Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* **157**, 714–725 (2014).
15. Scialdone, A. et al. Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods* **85**, 54–61 (2015).
16. Ersoy, I., Bunyak, F., Chagin, V., Cardoso, M. C. & Palaniappan, K. Segmentation and classification of cell cycle phases in fluorescence imaging. in *International Conference on Medical Image Computing and Computer-Assisted Intervention(MICCAI).* 617–624 (Springer, Berlin, Heidelberg, 2009).
17. Du, T. H., Puah, W. C. & Wasser, M. Cell cycle phase classification in 3D in vivo microscopy of Drosophila embryogenesis [J]. *BMC Bioinform.* **12**, 1–9 (2011).
18. Schönenberger, F., Deutzmann, A., Ferrando-May, E. & Merhof, D. Discrimination of cell cycle phases in PCNA-immunolabeled cells [J]. *BMC Bioinform.* **16**, 1–10 (2015).
19. Hsiao, C. J. et al. Characterizing and inferring quantitative cell cycle phase in single-cell RNA-seq data analysis [J]. *Genome Res.* **30**, 611–621 (2020).
20. Kowalczyk, M. S. et al. Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells [J]. *Genome Res.* **25**, 1860–1872 (2015).
21. Wu, H. et al. scHiCStackL: a stacking ensemble learning-based method for single-cell Hi-C classification using cell embedding [J]. *Brief. Bioinform.* **23**, bbab396 (2022).
22. Nagano, T. et al. Cell-cycle dynamics of chromosomal organization at single-cell resolution [J]. *Nature* **547**, 61–67 (2017).
23. Liu, Z. et al. Linking genome structures to functions by simultaneous single-cell Hi-C and RNA-seq. *Science* **380**, 1070–1076 (2023).
24. Lundberg, S. M. & Lee, S. I. A unified approach to interpreting model predictions [J]. *Adv. Neural Inf. Process. Syst*. **30**, 4765–4774 (2017).
25. Hashemzadeh, H., Shojaeilangari, S. & Allahverdi, A. A combined microfluidic deep learning approach for lung cancer cell high throughput screening toward automatic cancer screening applications. *Sci. Rep.* **11**, 9804 (2021).
26. Liang, X. et al. Large-scale comparative review and assessment of computational methods for anti-cancer peptide identification. *Brief Bioinform*. https://doi.org/10.1093/bib/bbaa312 (2020).
27. Noble, W. S. What is a support vector machine? [J]. *Nat. Biotechnol.* **24**, 1565–1567 (2006).
28. LaValley, M. P. Logistic regression [J]. *Circulation* **117**, 2395–2399 (2008).
29. Rigatti, S. J. Random forest [J]. *J. Insurance Med.* **47**, 31–39 (2017).
30. Zhang, P. & Wu, H. IChrom-deep: an attention-based deep learning model for identifying chromatin interactions [J]. *IEEE J. Biomed. Health Inform.* **27**, 1–12 (2023).
31. Hashimoto, H. et al. Time-lapse imaging of cell cycle dynamics during development in living cardiomyocyte[J]. *J. Mol. Cell. Cardiol.* **72**, 241–249 (2014).
32. Zhang, J. M., Wei, Q., Zhao, X. & Paterson, B. M. Coupling of the cell cycle and myogenesis through the cyclin D1-dependent interaction of MyoD with cdk4[J]. *EMBO J.* **18**, 926–933 (1999).
33. Jia, W. et al. Non-canonical roles of PFKFB3 in regulation of cell cycle through binding to CDK4[J]. *Oncogene* **37**, 1685–1698 (2018).
34. Yasuhara, N., Takeda, E., Inoue, H., Kotera, I. & Yoneda, Y. Importin α/β-mediated nuclear protein import is regulated in a cell cycle-dependent manner[J]. *Exp. cell Res.* **297**, 285–293 (2004).
35. Zou, Y. et al. Characterization of nuclear localization signal in the N terminus of CUL4B and its essential role in cyclin E degradation and cell cycle progression[J]. *J. Biol. Chem.* **284**, 33320–33332 (2009).
36. Flores-Delgado, G., Liu, C. W. Y., Sposto, R. & Berndt, N. A limited screen for protein interactions reveals new roles for protein phosphatase 1 in cell cycle control and apoptosis[J]. *J. Proteome Res.* **6**, 1165–1175 (2007).
37. Yu, C. et al. BTG4 is a meiotic cell cycle–coupled maternal-zygotic-transition licensing factor in oocytes[J]. *Nat. Struct. Mol. Biol.* **23**, 387–394 (2016).
38. Hirai, M. et al. Adaptor proteins NUMB and NUMBL promote cell cycle withdrawal by targeting ERBB2 for degradation[J]. *J. Clin. Investig.* **127**, 569–582 (2017).
39. Li, M. et al. Somatostatin receptor-1 induces cell cycle arrest and inhibits tumor growth in pancreatic cancer[J]. *Cancer Sci.* **99**, 2218–2223 (2008).
40. Lu, H. et al. Cell cycle-dependent phosphorylation regulates RECQL4 pathway choice and ubiquitination in DNA double-strand break repair[J]. *Nat. Commun.* **8**, 2039 (2017).
41. Illenberger, S. et al. The endogenous and cell cycle-dependent phosphorylation of tau protein in living cells: implications for Alzheimer's disease[J]. *Mol. Biol. Cell* **9**, 1495–1512 (1998).
42. Dekker, J., Marti-Renom, M. & Mirny, L. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data [J]. *Nat. Rev. Genet.* **14**, 390–403 (2013).
43. Naumova, N. et al. Organization of the mitotic chromosome [J]. *Science* **342**, 948–953 (2013).
44. Ay, F. & Noble, W. S. Analysis methods for studying the 3D architecture of the genome [J]. *Genome Biol.* **16**, 183 (2015).
45. Yang, T. et al. HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient [J]. *Genome Res.* **27**, 1939–1949 (2017).
46. Zhou, X., Shi, Z., Wu, Y., Zhao, J. & Wu, H. scHiCSC: a novel single-cell Hi-C clustering framework by contact-weight-based smoothing and feature fusion, 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). *IEEE,* 44–50 (2022).
47. Gawad, C., Koh, W. & Quake, S. R. Single-cell genome sequencing: current state of the science[J]. *Nat. Rev. Genet.* **17**, 175–188 (2016).
48. Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C. & Teichmann, S. A. The technology and biology of single-cell RNA sequencing[J]. *Mol. Cell* **58**, 610–620 (2015).
49. Schwartzman, O. & Tanay, A. Single-cell epigenomics: techniques and emerging applications[J]. *Nat. Rev. Genet.* **16**, 716–726 (2015).
50. Kelsey, G., Stegle, O. & Reik, W. Single-cell epigenomics: recording the past and predicting the future[J]. *Science* **358**, 69–75 (2017).
51. Liu, Y., Zhou, Y., Wen, S. & Tang, C. A strategy on selecting performance metrics for classifier evaluation [J]. *Int. J. Mob. Comput. Multimed. Commun.* **6**, 20–35 (2014).
52. Zhang, P., Zhang, H. & Wu, H. iPro-WAEL: a comprehensive and robust framework for identifying promoters in multiple species [J]. *Nucleic Acids Res.* **50**, 10278–10289 (2022).
53. Liu, J., Lin, D., Yardımcı, G. G. & Noble, W. S. Unsupervised embedding of single-cell Hi-C data [J]. *Bioinformatics* **34**, i96–i104 (2018).
54. Zhang, P. et al. CLNN-loop: a deep learning model to predict CTCF-mediated chromatin loops in the different cell lines and CTCF-binding sites (CBS) pair types [J]. *Bioinformatics* **38**, 4497–4504 (2022).
55. Zhang, H. et al. Hyperspectral classification based on lightweight 3-D-CNN with transfer learning. *IEEE Trans. Geosci. Remote Sens.* **57**, 5813–5828 (2019).
56. Tripathy, A., Yelick, K. & Buluc, A. Reducing communication in graph neural network training. *SC20: International Conference for High*

Performance Computing, Networking, Storage and Analysis. 1–14 (IEEE Press, Atlanta, Georgia, 2020).
57. Wu, Z., Qu, X., Huang, J. & Wu, X. In-air handwritten chinese text recognition with attention convolutional recurrent network. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, (Cham: Springer Nature Switzerland; 2023).
58. Lin, T., Goyal, P., Girshick, R., He, K. & Dollar, P. Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **42**, 318–327 (2020).
59. Li, X. et al. Generalized focal loss: towards efficient representation learning for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 1–14 (2023).
60. Yao, Y., Rosasco, L. & Caponnetto, A. On early stopping in gradient descent learning. *Construct. Approx.* **26**, 289–315 (2007).
61. Prechelt, L. Early stopping: But when? (Berlin: Springer; 1998).
62. Ji, F., Zhang, X. & Zhao, J. α-EGAN: α-Energy distance GAN with an early stopping rule. *Comput. Vis. Image Underst.* **234**, 103748 (2023).
63. Walter, S. D. et al. Randomized trials with provision for early stopping for benefit (or harm): The impact on the estimated treatment effect. *Stat. Med.* **38**, 2524–2543 (2019).
64. Cataltepe, Z., Abu-Mostafa, Y. S. & Magdon-Ismail, M. No free lunch for early stopping. *Neural Comput.* **11**, 995–1009 (1999).
65. Wu, Y. et al. scHiCyclePred: a deep learning framework for predicting cell cycle phases from single-cell Hi-C data using multi-scale interaction information [code] Zenodo https://doi.org/10.5281/zenodo.12721771.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grant No. 62272278), the National Key Research and Development Program (Grant No. 2021YFF0704103), and Guangdong Basic and Applied Basic Research Foundation (Grant No. 2024A1515012775), held by H.W. This work is partially supported by grants from the Canadian Institutes of Health Research (CIHR) (PJT-180505), the Fonds de recherche du Québec - Santé (FRQS) (295298, 295299), the Natural Sciences and Engineering Research Council of Canada (NSERC) (RGPIN2022-04399), and the Meakins-Christie Chair in Respiratory Research, held by J.D. The funders did not play any role in the design of the study, the collection, analysis, and interpretation of data, or the writing of the manuscript.

Author contributions

H.W. and J.D. contributed to designing the study. Y.W., Z.S., and X.Z. constructed the models and performed the related experiments. Y.W., Z.S., X.Z., and P.Z. analyzed the scHi-C data. X.Y. and P.Z. performed a genetic analysis. Y.W., Z.S., and X.Z. wrote the manuscript. Y.W., Z.S., and X.Y. revised the manuscript based on comments from all reviewers. All authors read and approved the final manuscript.

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/s42003-024-06626-3.

Correspondence and requests for materials should be addressed to Jun Ding or Hao Wu.

Peer review information *Communications Biology* thanks Lindsay Lee, Ali Al-Fatlawi and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editors: Adib Keikhosravi and David Favero. A peer review file is available.