Supplemental information

Predicting A/B compartments from histone modifications using deep learning

Suchen Zheng, Nitya Thakkar, Hannah L. Harris, Susanna Liu, Megan Zhang, Mark Gerstein, Erez Lieberman Aiden, M. Jordan Rowley, William Stafford Noble, Gamze Gürsoy, and Ritambhara Singh
Predicting A/B compartments from histone modifications using deep learning

Supplementry Information

A Tables

| Cell   | Hidden layer size | Number of layers | Output layer size |
|--------|-------------------|------------------|-------------------|
| GM12878| 64                | 2                | 64                |
| K562   | 32                | 4                | 32                |
| IMR90  | 32                | 4                | 32                |
| HMEC   | 128               | 1                | 128               |
| NHEK   | 64                | 1                | 64                |
| HUVEC  | 64                | 4                | 64                |

Table S1. Hyper-parameters of CoRNN. For all cells, we trained the model using a batch size of 64, a learning rate of 0.001, and 20 epochs. Related to fig. 2.

| Cell   | A count | B count |
|--------|---------|---------|
| HUVEC  | 13265   | 13714   |
| HMEC   | 13140   | 13779   |
| IMR90  | 12691   | 14261   |
| GM12878| 12754   | 14210   |
| K562   | 13099   | 13851   |
| NHEK   | 14233   | 12707   |
| Total  | 79182   | 82522   |

Table S2. Data summary of the six selected cells. Related to fig. 2.

|               | GM12878 | K562 | IMR90 | HMEC | NHEK | HUVEC |
|---------------|---------|------|-------|------|------|-------|
| Mean Baseline | 0.905   | 0.893| 0.872 | 0.938| 0.85 | 0.916 |
| CoRNN (our model) | 0.953 | 0.938| 0.901 | 0.957| 0.857| 0.916 |
| GRU           | 0.911   | 0.939| 0.781 | 0.904| 0.745| 0.74  |
| Random Forest | 0.79    | 0.785| 0.826 | 0.723| 0.699| 0.533 |
| Random Forest (add mean) | 0.88  | 0.852| 0.837 | 0.872| 0.788| 0.81  |
| Logistic Regression | 0.533 | 0.806| 0.815 | 0.738| 0.508| 0.853 |
| Logistic Regression (add mean) | 0.837 | 0.835| 0.83  | 0.882| 0.788| 0.814 |

Table S3. CoRNN and baseline model performance. Related to fig. 3.
B Figures

Fig. S1. We compare the correlations of first-, second-, and third-order eigenvectors obtained from Hi-C correlation matrices for all 6 cell lines. Based on these correlation scores, we see that the first-order eigenvector corresponds to the chromosomal compartments in all cell lines. Related to fig. 1.

Fig. S2. We compare the correlations of first-order eigenvectors obtained from Hi-C correlation matrices for all 6 cell lines across 100kbp and 50kbp resolution. The chromosomal compartments (first-order eigenvectors) show a weaker correlation with histone modification signal at higher resolution. Related to fig. 1.
Predicting A/B compartments from histone modifications using deep learning

Fig. S3. Cross-validation scheme for training CoRNN with IMR90 as the test cell line. Only the creation of validation folds 0 and 1 are shown as examples. A similar process was used for creating both test and validation folds with other cell lines. Related to fig. 2.

Fig. S4. Testing results of CoRNN and Convolutional neural network model. Our model outperforms the convolutional neural network, thus justifying the choice of GRU as our neural network architecture. Related to fig. 3.
Zheng and Thakkar, et al.

Fig. S5. Random forest and linear regression models with concatenated 600 histone modification signal values as input. These models yielded a worse performance than the mean baseline. Therefore, we used these models with mean values of histone modification signals as inputs for baseline comparison. Related to fig. 3.

Fig. S6. Correlation of compartment values across all six cell lines. Compartment values of HMEC, NHEK, and HUVEC have high correlations compared to the GM12878, K562, and IMR90 cell lines. This observation suggests that these cell lines are easier to predict using the mean compartment value baseline. Related to fig. 2.
Predicting A/B compartments from histone modifications using deep learning

Fig. S7. Testing results of CoRNN and baselines using AUPR score. Our model outperforms the mean baseline for five out of six cell lines. Related to fig. 3.

Fig. S8. Picking strong compartments. We select strong compartments as those with absolute values > (mean – std.deviation) for all the compartment values. Related to fig. 5.
Fig. S9. Testing performance of CoRNN and mean baseline on independent human muscle and colon tissue samples measured in AUPRC. CoRNN predicts A/B compartments for both tissues with higher scores compared to the mean baseline. Related to fig. 6.