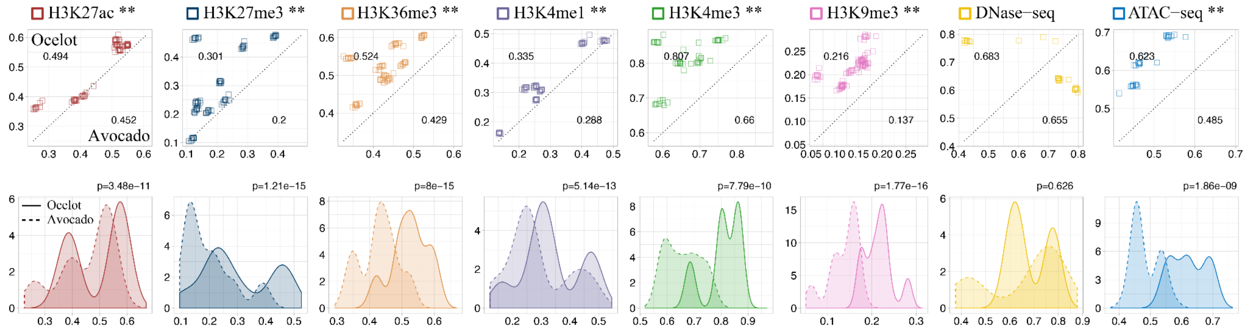
Asymmetric Predictive Relationships Across Histone Modifications

Hongyang Li[1,*], Yuanfang Guan[1,*]

1. Department of Computational Medicine and Bioinformatics, University of Michigan, 100 Washtenaw Avenue, Ann Arbor, MI 48109, USA

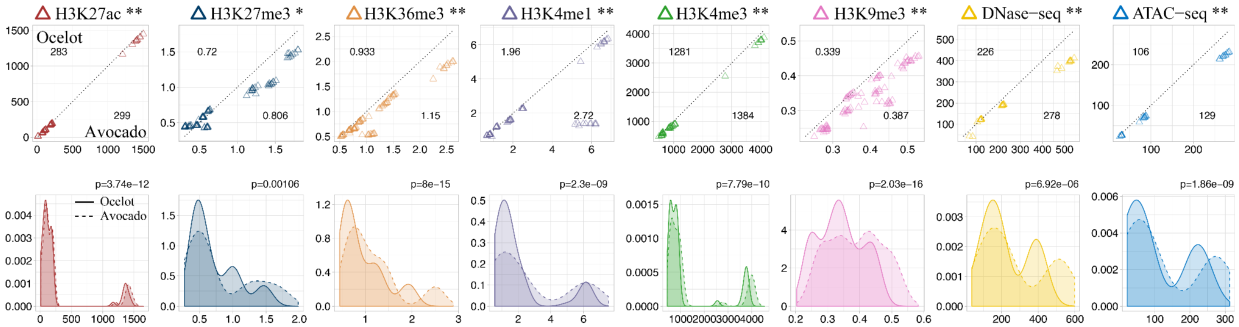
* Corresponding authors: hyangl@umich.edu or gyuanfan@umich.edu

Supplementary Figures



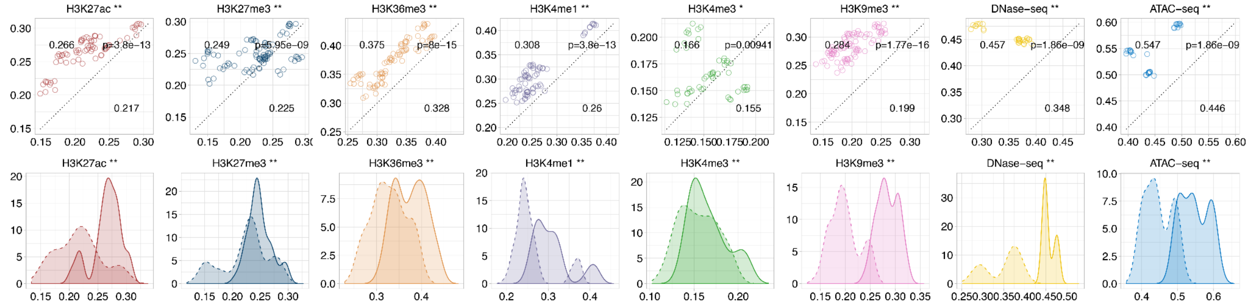
Supplementary Figure 1: Benchmarking against Avocado using the global Pearson's correlation at the mark level.

In the first row, we benchmarked our method (y-axis) against Avocado (x-axis) across the entire human genome on 51 held-out testing epigenomes covering six histone modifications, DNase-seq and ATAC-seq. Each symbol represents a testing sample. If a symbol is above the diagonal dashed line, it means our method achieves a higher predictive performance than Avocado. The paired Wilcoxon signed-rank test was used to statistically compare two methods and the p-values are shown in the figures. Significantly different results are labeled by asterisks (* p-value < 0.01 and ** p-value < 0.001). In the second row, the overall distribution of this evaluation metric is shown as density plots, where the solid lines represent our method and the dashed lines represent Avocado.



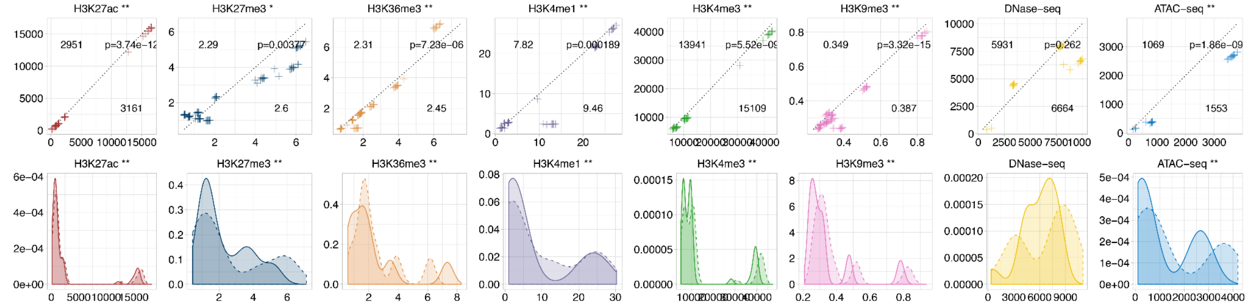
Supplementary Figure 2: Benchmarking against Avocado using the global mean squared error at the mark level.

In the first row, we benchmarked our method (y-axis) against Avocado (x-axis) across the entire human genome on 51 held-out testing epigenomes covering six histone modifications, DNase-seq and ATAC-seq. Each symbol represents a testing sample. If a symbol is above the diagonal dashed line, it means our method achieves a higher predictive performance than Avocado. The paired Wilcoxon signed-rank test was used to statistically compare two methods and the p-values are shown in the figures. Significantly different results are labeled by asterisks (* p-value < 0.01 and ** p-value < 0.001). In the second row, the overall distribution of this evaluation metric is shown as density plots, where the solid lines represent our method and the dashed lines represent Avocado.



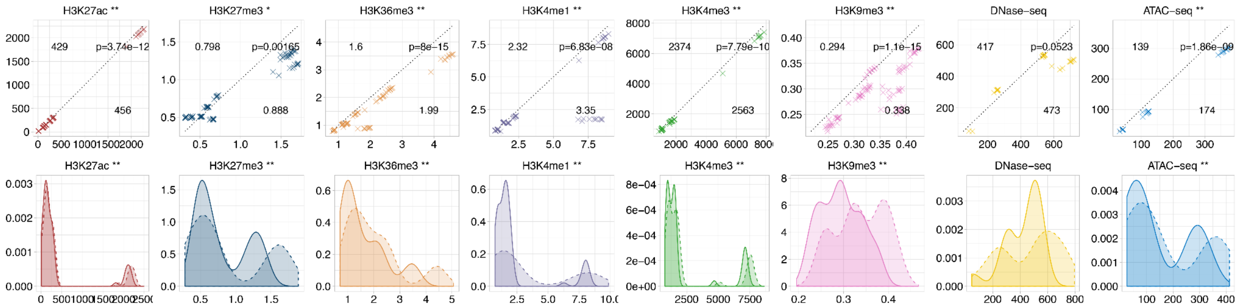
Supplementary Figure 3: Benchmarking against Avocado using the global Spearman's correlation at the mark level.

In the first row, we benchmarked our method (y-axis) against Avocado (x-axis) across the entire human genome on 51 held-out testing epigenomes covering six histone modifications, DNase-seq and ATAC-seq. Each symbol represents a testing sample. If a symbol is above the diagonal dashed line, it means our method achieves a higher predictive performance than Avocado. The paired Wilcoxon signed-rank test was used to statistically compare two methods and the p-values are shown in the figures. Significantly different results are labeled by asterisks (* p-value < 0.01 and ** p-value < 0.001). In the second row, the overall distribution of this evaluation metric is shown as density plots, where the solid lines represent our method and the dashed lines represent Avocado.



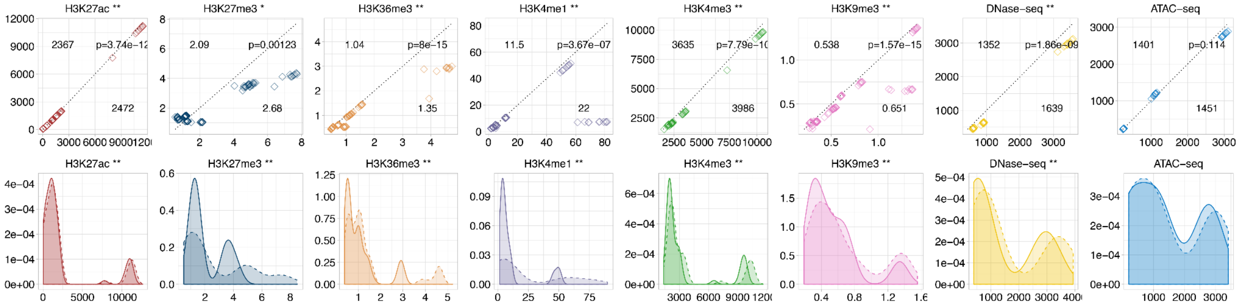
Supplementary Figure 4: Benchmarking against Avocado using the local MSEs across promoter regions at the mark level.

In the first row, we benchmarked our method (y-axis) against Avocado (x-axis) across the entire human genome on 51 held-out testing epigenomes covering six histone modifications, DNase-seq and ATAC-seq. Each symbol represents a testing sample. If a symbol is above the diagonal dashed line, it means our method achieves a higher predictive performance than Avocado. The paired Wilcoxon signed-rank test was used to statistically compare two methods and the p-values are shown in the figures. Significantly different results are labeled by asterisks (* p-value < 0.01 and ** p-value < 0.001). In the second row, the overall distribution of this evaluation metric is shown as density plots, where the solid lines represent our method and the dashed lines represent Avocado.



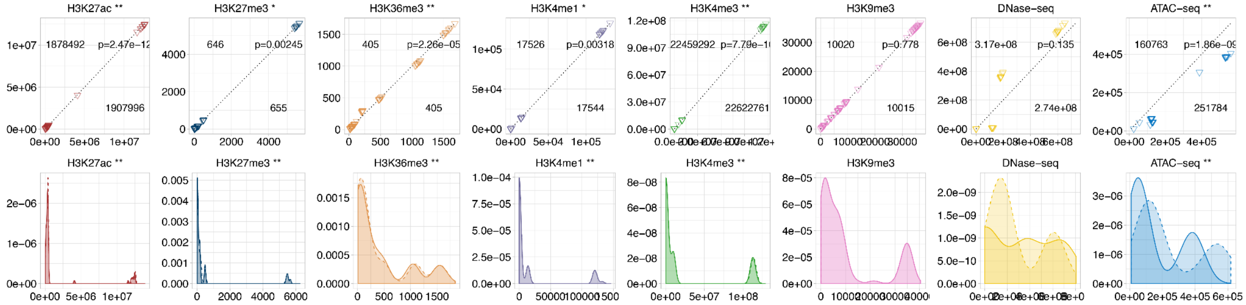
Supplementary Figure 5: Benchmarking against Avocado using the local MSEs across gene regions at the mark level.

In the first row, we benchmarked our method (y-axis) against Avocado (x-axis) across the entire human genome on 51 held-out testing epigenomes covering six histone modifications, DNase-seq and ATAC-seq. Each symbol represents a testing sample. If a symbol is above the diagonal dashed line, it means our method achieves a higher predictive performance than Avocado. The paired Wilcoxon signed-rank test was used to statistically compare two methods and the p-values are shown in the figures. Significantly different results are labeled by asterisks (* p-value < 0.01 and ** p-value < 0.001). In the second row, the overall distribution of this evaluation metric is shown as density plots, where the solid lines represent our method and the dashed lines represent Avocado.



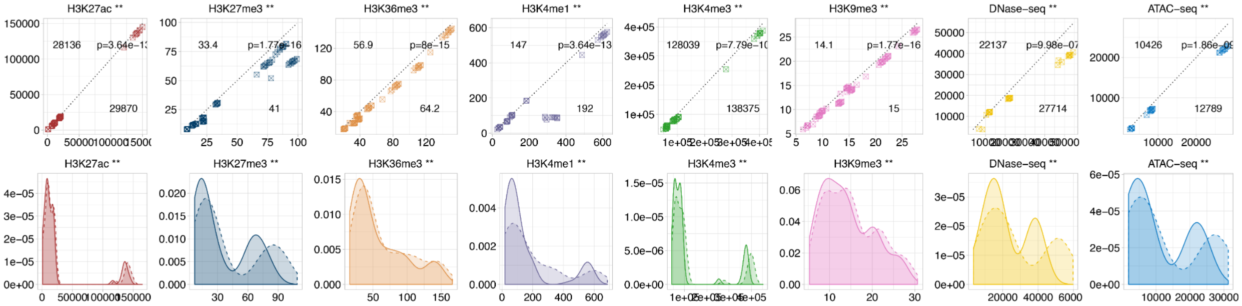
Supplementary Figure 6: Benchmarking against Avocado using the local MSEs across enhancer regions at the mark level.

In the first row, we benchmarked our method (y-axis) against Avocado (x-axis) across the entire human genome on 51 held-out testing epigenomes covering six histone modifications, DNase-seq and ATAC-seq. Each symbol represents a testing sample. If a symbol is above the diagonal dashed line, it means our method achieves a higher predictive performance than Avocado. The paired Wilcoxon signed-rank test was used to statistically compare two methods and the p-values are shown in the figures. Significantly different results are labeled by asterisks (* p-value < 0.01 and ** p-value < 0.001). In the second row, the overall distribution of this evaluation metric is shown as density plots, where the solid lines represent our method and the dashed lines represent Avocado.



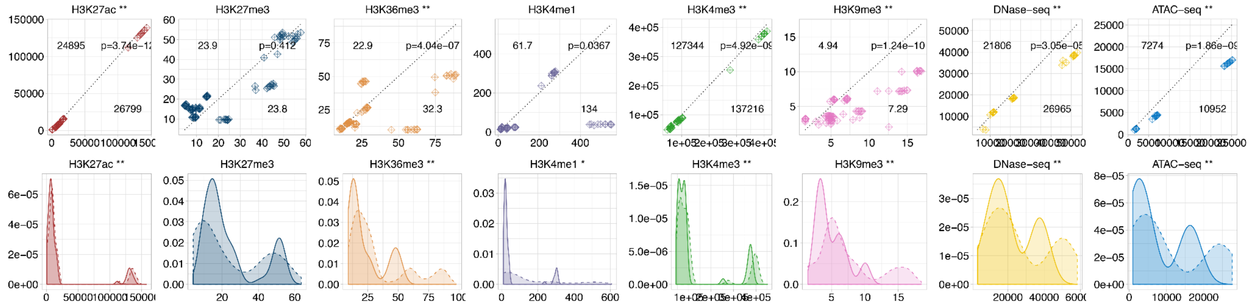
Supplementary Figure 7: Benchmarking against Avocado using the global MSE weighted by the cross-cell-type variance at the mark level.

In the first row, we benchmarked our method (y-axis) against Avocado (x-axis) across the entire human genome on 51 held-out testing epigenomes covering six histone modifications, DNase-seq and ATAC-seq. Each symbol represents a testing sample. If a symbol is above the diagonal dashed line, it means our method achieves a higher predictive performance than Avocado. The paired Wilcoxon signed-rank test was used to statistically compare two methods and the p-values are shown in the figures. Significantly different results are labeled by asterisks (* p-value < 0.01 and ** p-value < 0.001). In the second row, the overall distribution of this evaluation metric is shown as density plots, where the solid lines represent our method and the dashed lines represent Avocado.



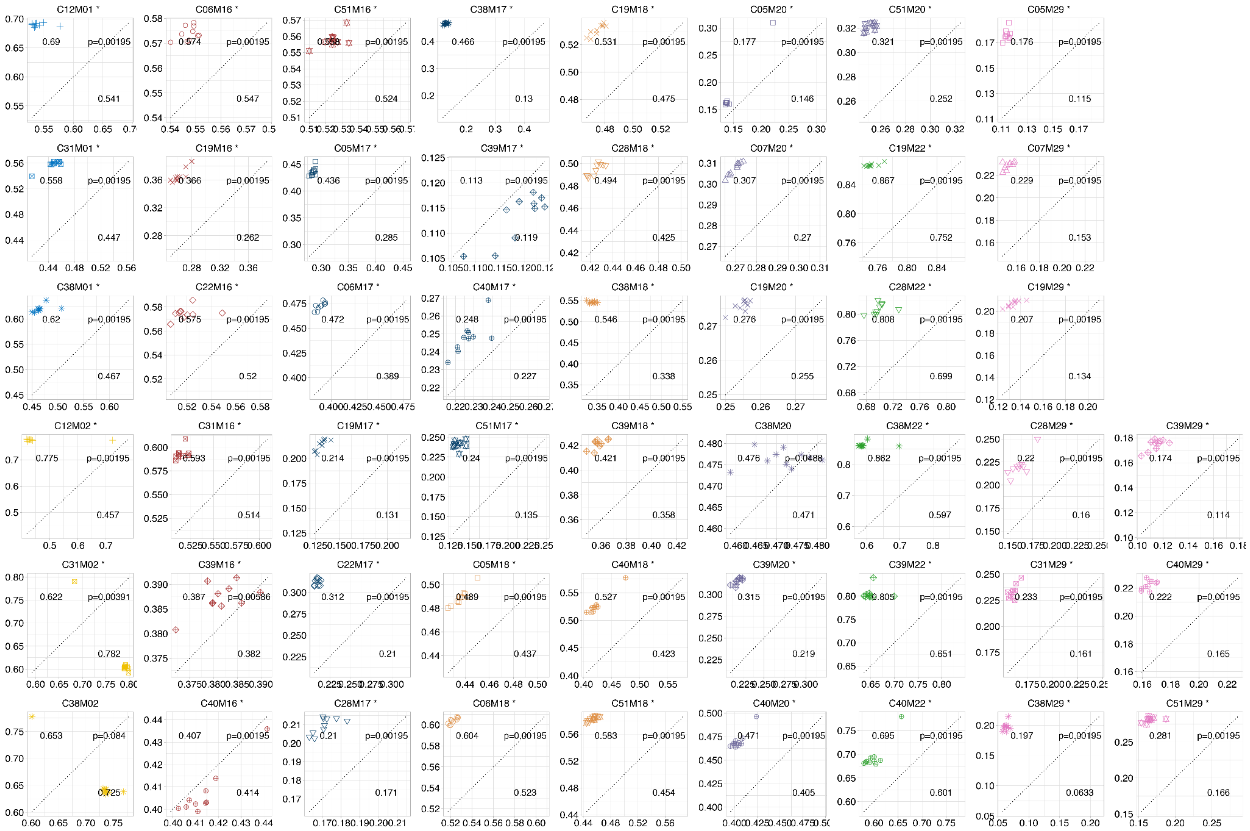
Supplementary Figure 8: Benchmarking against Avocado using the local MSE across genomic regions with top 1% observed values.

In the first row, we benchmarked our method (y-axis) against Avocado (x-axis) across the entire human genome on 51 held-out testing epigenomes covering six histone modifications, DNase-seq and ATAC-seq. Each symbol represents a testing sample. If a symbol is above the diagonal dashed line, it means our method achieves a higher predictive performance than Avocado. The paired Wilcoxon signed-rank test was used to statistically compare two methods and the p-values are shown in the figures. Significantly different results are labeled by asterisks (* p-value < 0.01 and ** p-value < 0.001). In the second row, the overall distribution of this evaluation metric is shown as density plots, where the solid lines represent our method and the dashed lines represent Avocado.



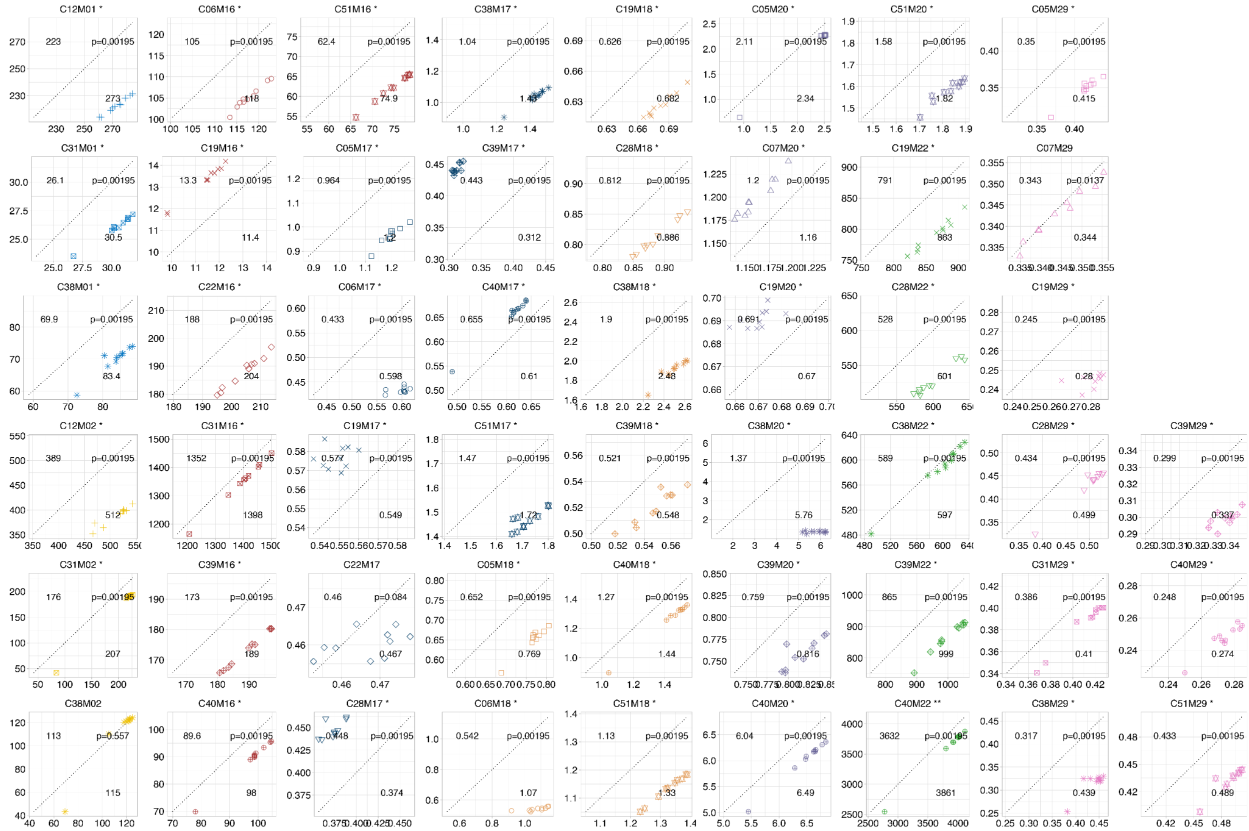
Supplementary Figure 9: Benchmarking against Avocado using the local MSE across genomic regions with top 1% predicted values at the mark level.

In the first row, we benchmarked our method (y-axis) against Avocado (x-axis) across the entire human genome on 51 held-out testing epigenomes covering six histone modifications, DNase-seq and ATAC-seq. Each symbol represents a testing sample. If a symbol is above the diagonal dashed line, it means our method achieves a higher predictive performance than Avocado. The paired Wilcoxon signed-rank test was used to statistically compare two methods and the p-values are shown in the figures. Significantly different results are labeled by asterisks (* p-value < 0.01 and ** p-value < 0.001). In the second row, the overall distribution of this evaluation metric is shown as density plots, where the solid lines represent our method and the dashed lines represent Avocado.



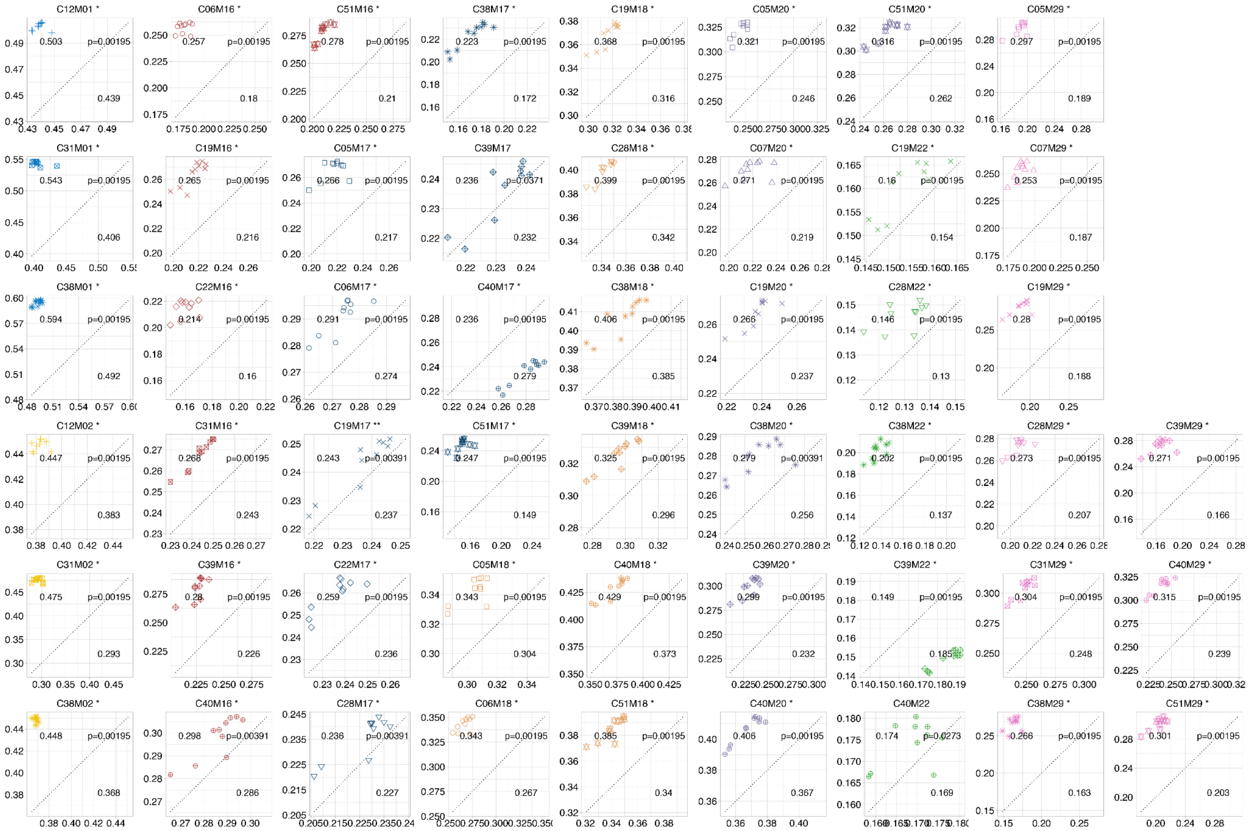
Supplementary Figure 10: Benchmarking against Avocado using the global Pearson's correlation for each testing mark-cell type pair.

We benchmarked our method (y-axis) against Avocado (x-axis) across the entire human genome on 51 held-out testing epigenomes covering six histone modifications, DNase-seq and ATAC-seq. Each figure contains 10 bootstrap sampling results from the human genome. The color represents the mark type and the symbol shape represents the cell type. If a symbol is above the diagonal dashed line, it means our method achieves a higher predictive performance than Avocado. The paired Wilcoxon signed-rank test was used to statistically compare two methods and the p-values are shown in the figures. Significantly different results are labeled by asterisks (* p-value < 0.01 and ** p-value < 0.001).



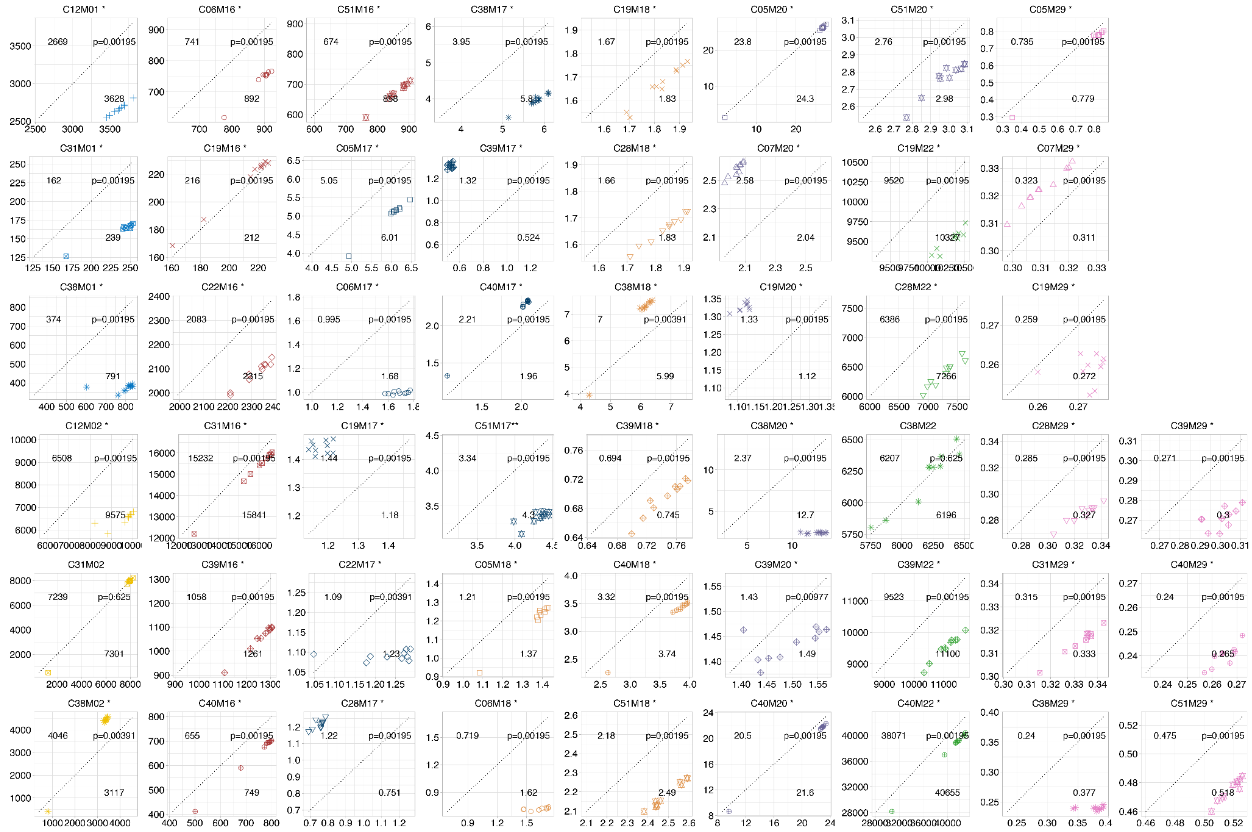
Supplementary Figure 11: Benchmarking against Avocado using the global MSE for each testing mark-cell type pair.

We benchmarked our method (y-axis) against Avocado (x-axis) across the entire human genome on 51 held-out testing epigenomes covering six histone modifications, DNase-seq and ATAC-seq. Each figure contains 10 bootstrap sampling results from the human genome. The color represents the mark type and the symbol shape represents the cell type. If a symbol is above the diagonal dashed line, it means our method achieves a higher predictive performance than Avocado. The paired Wilcoxon signed-rank test was used to statistically compare two methods and the p-values are shown in the figures. Significantly different results are labeled by asterisks (* p-value < 0.01 and ** p-value < 0.001).



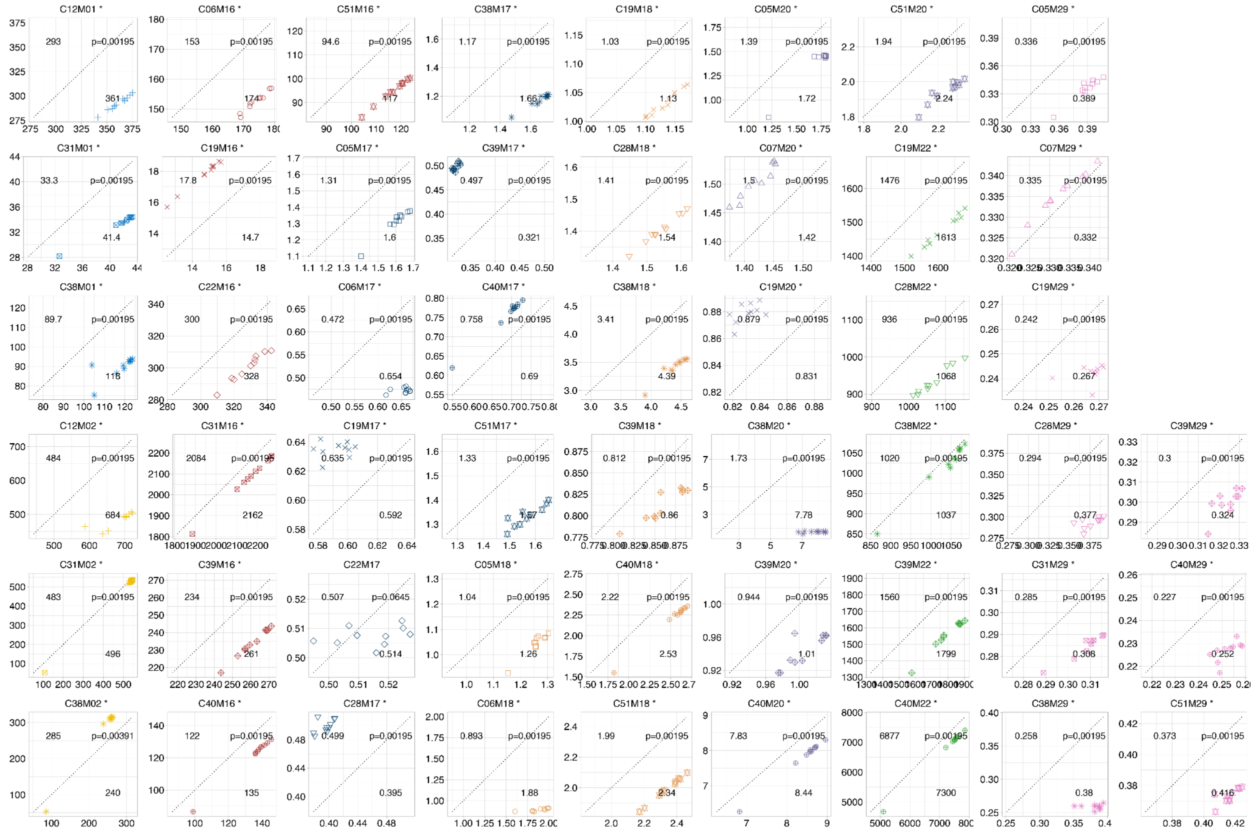
Supplementary Figure 12: Benchmarking against Avocado using the global Spearman's correlation for each testing mark-cell type pair.

We benchmarked our method (y-axis) against Avocado (x-axis) across the entire human genome on 51 held-out testing epigenomes covering six histone modifications, DNase-seq and ATAC-seq. Each figure contains 10 bootstrap sampling results from the human genome. The color represents the mark type and the symbol shape represents the cell type. If a symbol is above the diagonal dashed line, it means our method achieves a higher predictive performance than Avocado. The paired Wilcoxon signed-rank test was used to statistically compare two methods and the p-values are shown in the figures. Significantly different results are labeled by asterisks (* p-value < 0.01 and ** p-value < 0.001).



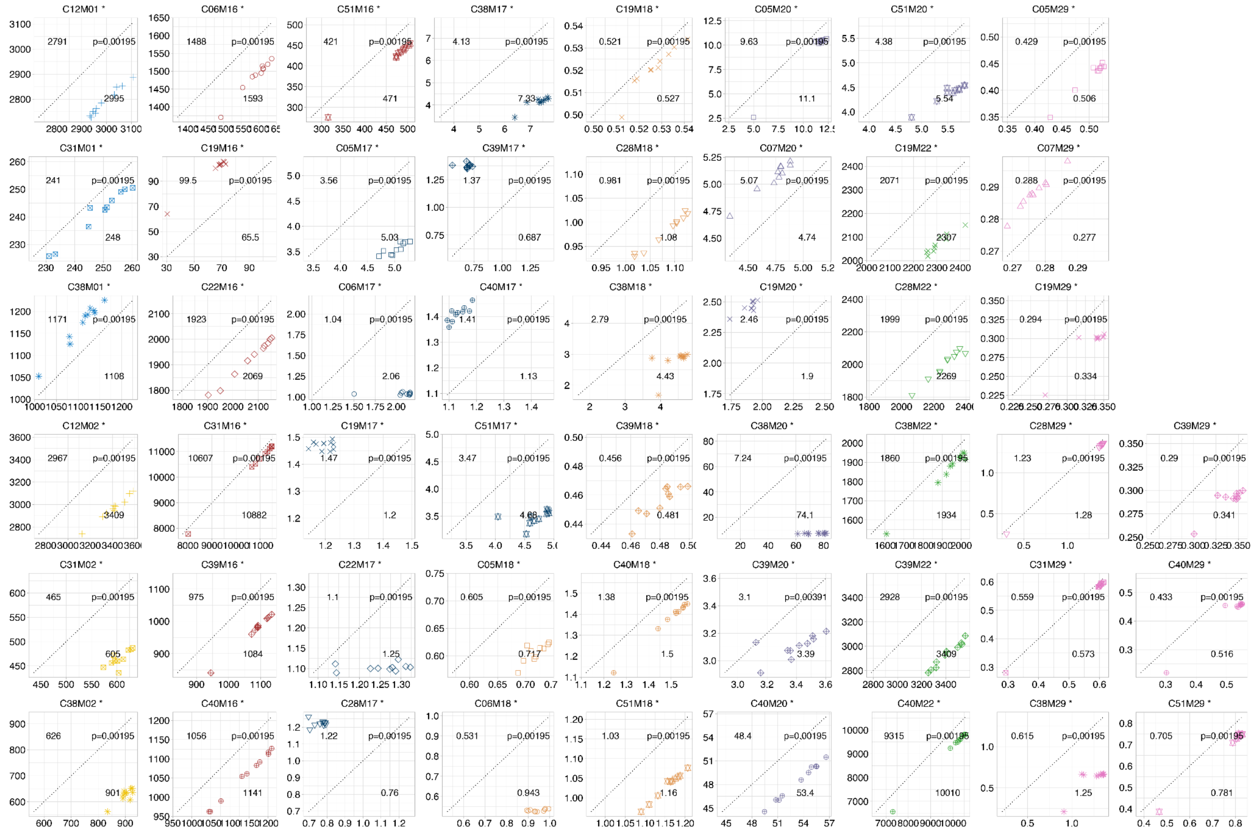
Supplementary Figure 13: Benchmarking against Avocado using the local MSEs across promoter regions for each testing mark-cell type pair.

We benchmarked our method (y-axis) against Avocado (x-axis) across the entire human genome on 51 held-out testing epigenomes covering six histone modifications, DNase-seq and ATAC-seq. Each figure contains 10 bootstrap sampling results from the human genome. The color represents the mark type and the symbol shape represents the cell type. If a symbol is above the diagonal dashed line, it means our method achieves a higher predictive performance than Avocado. The paired Wilcoxon signed-rank test was used to statistically compare two methods and the p-values are shown in the figures. Significantly different results are labeled by asterisks (* p-value < 0.01 and ** p-value < 0.001).



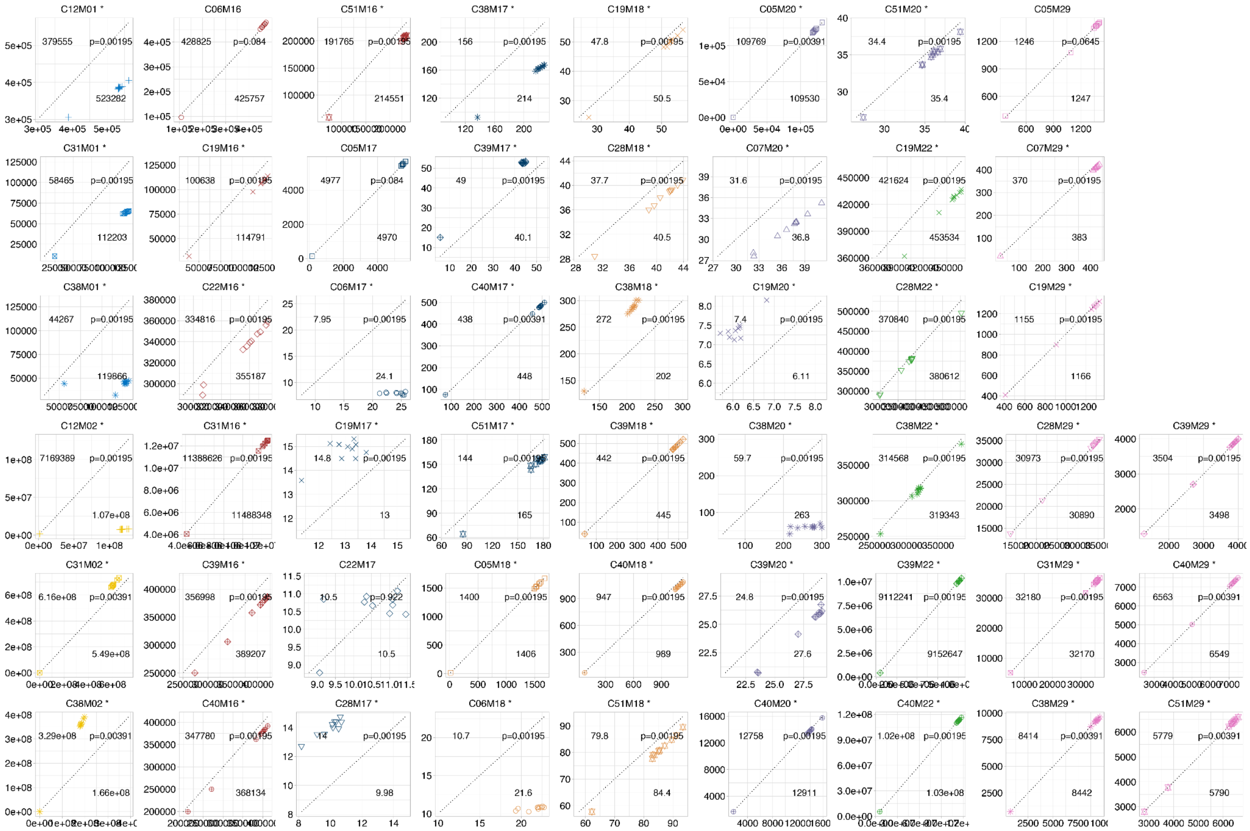
Supplementary Figure 14: Benchmarking against Avocado using the local MSEs across gene regions for each testing mark-cell type pair.

We benchmarked our method (y-axis) against Avocado (x-axis) across the entire human genome on 51 held-out testing epigenomes covering six histone modifications, DNase-seq and ATAC-seq. Each figure contains 10 bootstrap sampling results from the human genome. The color represents the mark type and the symbol shape represents the cell type. If a symbol is above the diagonal dashed line, it means our method achieves a higher predictive performance than Avocado. The paired Wilcoxon signed-rank test was used to statistically compare two methods and the p-values are shown in the figures. Significantly different results are labeled by asterisks (* p-value < 0.01 and ** p-value < 0.001).



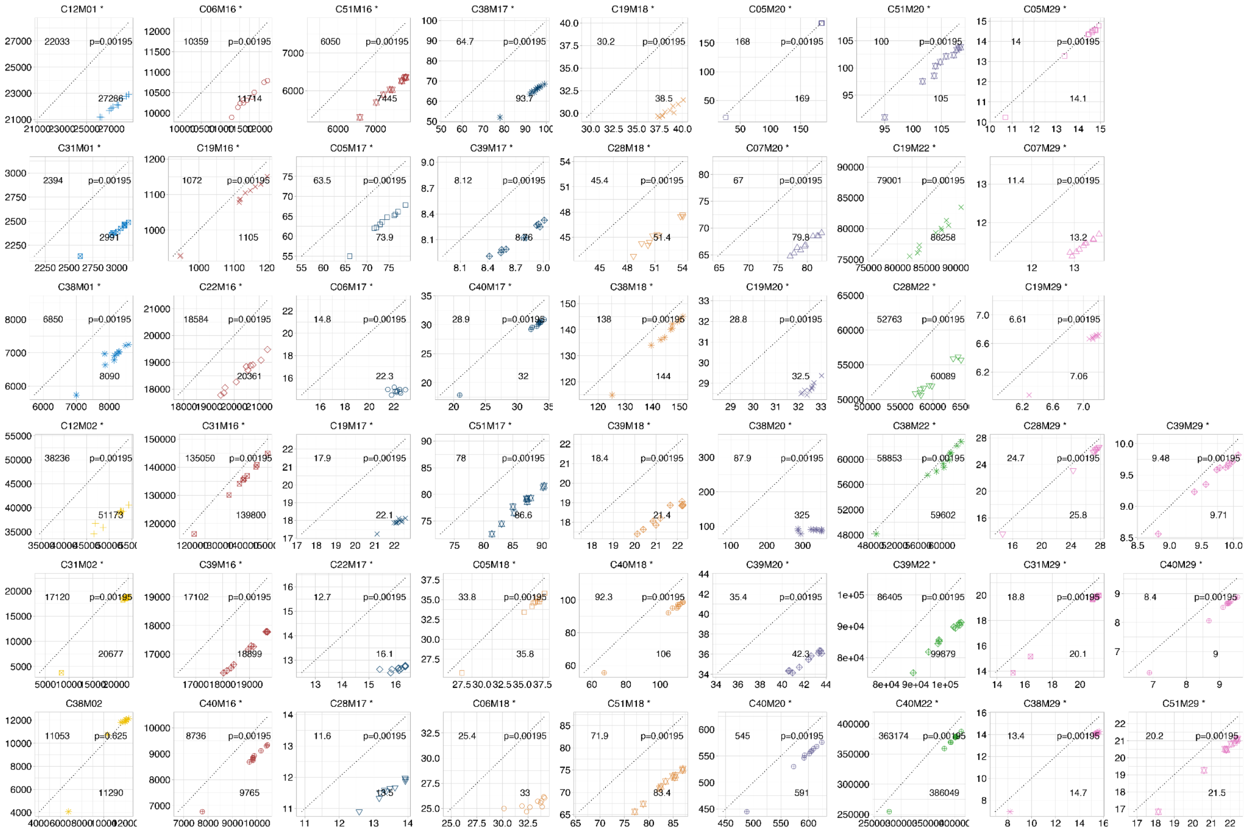
Supplementary Figure 15: Benchmarking against Avocado using the local MSEs across enhancer regions for each testing mark-cell type pair.

We benchmarked our method (y-axis) against Avocado (x-axis) across the entire human genome on 51 held-out testing epigenomes covering six histone modifications, DNase-seq and ATAC-seq. Each figure contains 10 bootstrap sampling results from the human genome. The color represents the mark type and the symbol shape represents the cell type. If a symbol is above the diagonal dashed line, it means our method achieves a higher predictive performance than Avocado. The paired Wilcoxon signed-rank test was used to statistically compare two methods and the p-values are shown in the figures. Significantly different results are labeled by asterisks (* p-value < 0.01 and ** p-value < 0.001).



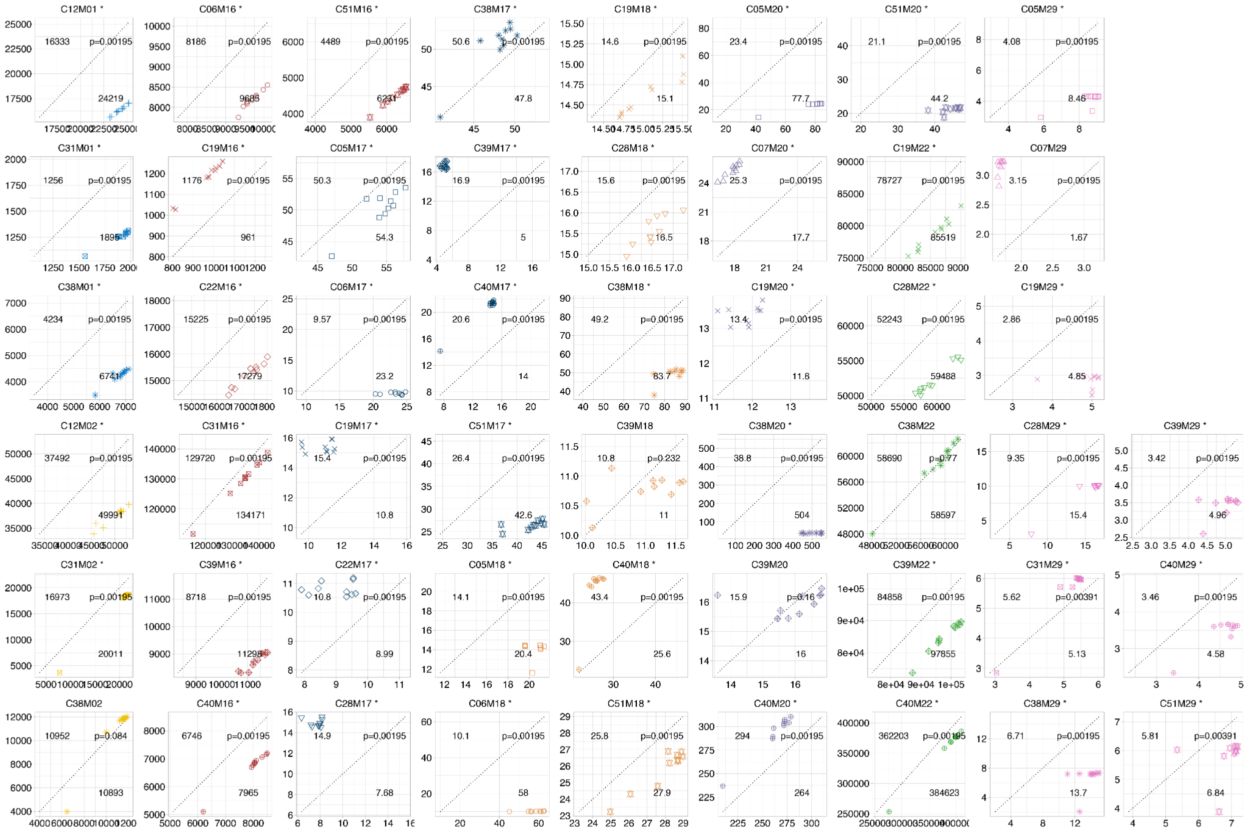
Supplementary Figure 16: Benchmarking against Avocado using the global MSE weighted by the cross-cell-type variance for each testing mark-cell type pair.

We benchmarked our method (y-axis) against Avocado (x-axis) across the entire human genome on 51 held-out testing epigenomes covering six histone modifications, DNase-seq and ATAC-seq. Each figure contains 10 bootstrap sampling results from the human genome. The color represents the mark type and the symbol shape represents the cell type. If a symbol is above the diagonal dashed line, it means our method achieves a higher predictive performance than Avocado. The paired Wilcoxon signed-rank test was used to statistically compare two methods and the p-values are shown in the figures. Significantly different results are labeled by asterisks (* p-value < 0.01 and ** p-value < 0.001).



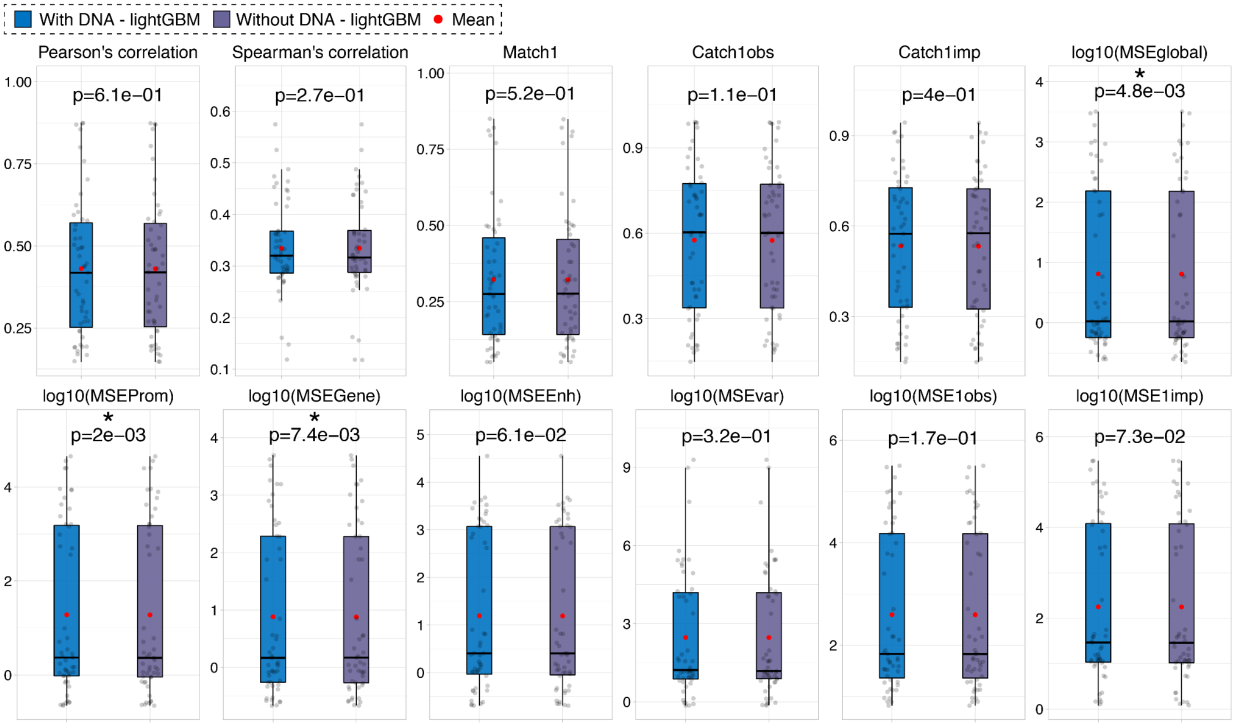
Supplementary Figure 17: Benchmarking against Avocado using the local MSE across genomic regions with top 1% observed values for each testing mark-cell type pair.

We benchmarked our method (y-axis) against Avocado (x-axis) across the entire human genome on 51 held-out testing epigenomes covering six histone modifications, DNase-seq and ATAC-seq. Each figure contains 10 bootstrap sampling results from the human genome. The color represents the mark type and the symbol shape represents the cell type. If a symbol is above the diagonal dashed line, it means our method achieves a higher predictive performance than Avocado. The paired Wilcoxon signed-rank test was used to statistically compare two methods and the p-values are shown in the figures. Significantly different results are labeled by asterisks (* p-value < 0.01 and ** p-value < 0.001).



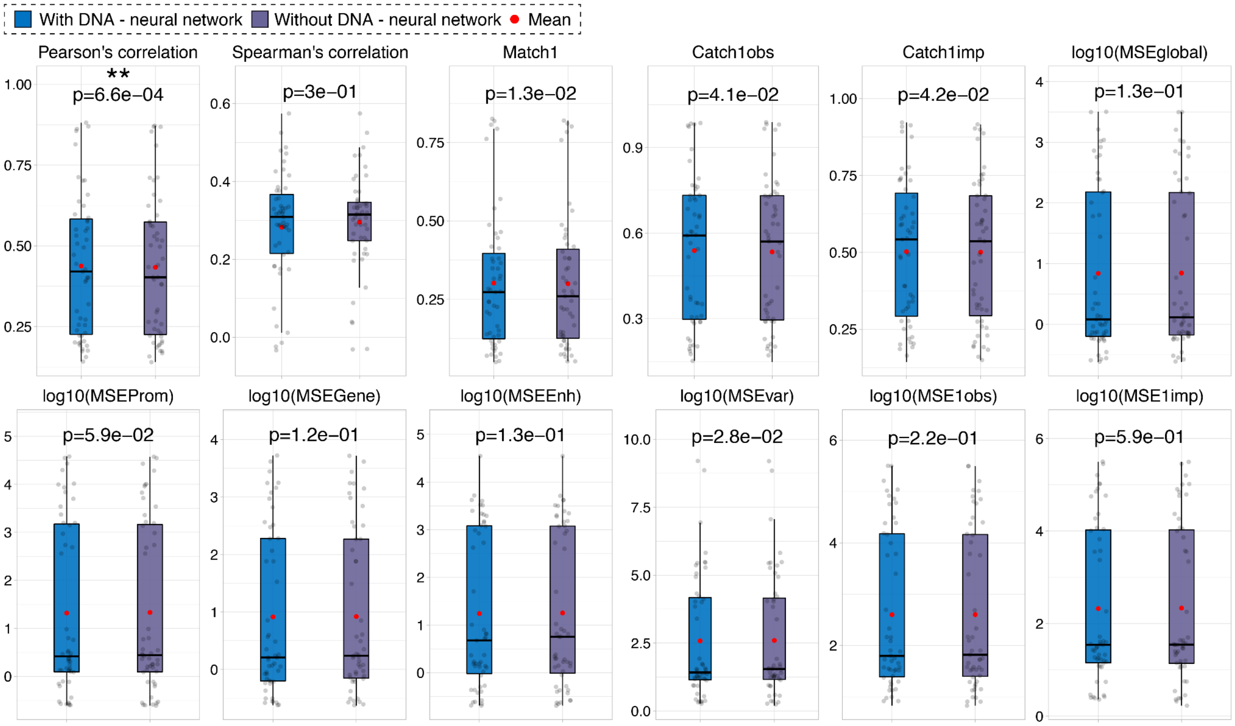
Supplementary Figure 18: Benchmarking against Avocado using the local MSE across genomic regions with top 1% predicted values for each testing mark-cell type pair.

We benchmarked our method (y-axis) against Avocado (x-axis) across the entire human genome on 51 held-out testing epigenomes covering six histone modifications, DNase-seq and ATAC-seq. Each figure contains 10 bootstrap sampling results from the human genome. The color represents the mark type and the symbol shape represents the cell type. If a symbol is above the diagonal dashed line, it means our method achieves a higher predictive performance than Avocado. The paired Wilcoxon signed-rank test was used to statistically compare two methods and the p-values are shown in the figures. Significantly different results are labeled by asterisks (* p-value < 0.01 and ** p-value < 0.001).



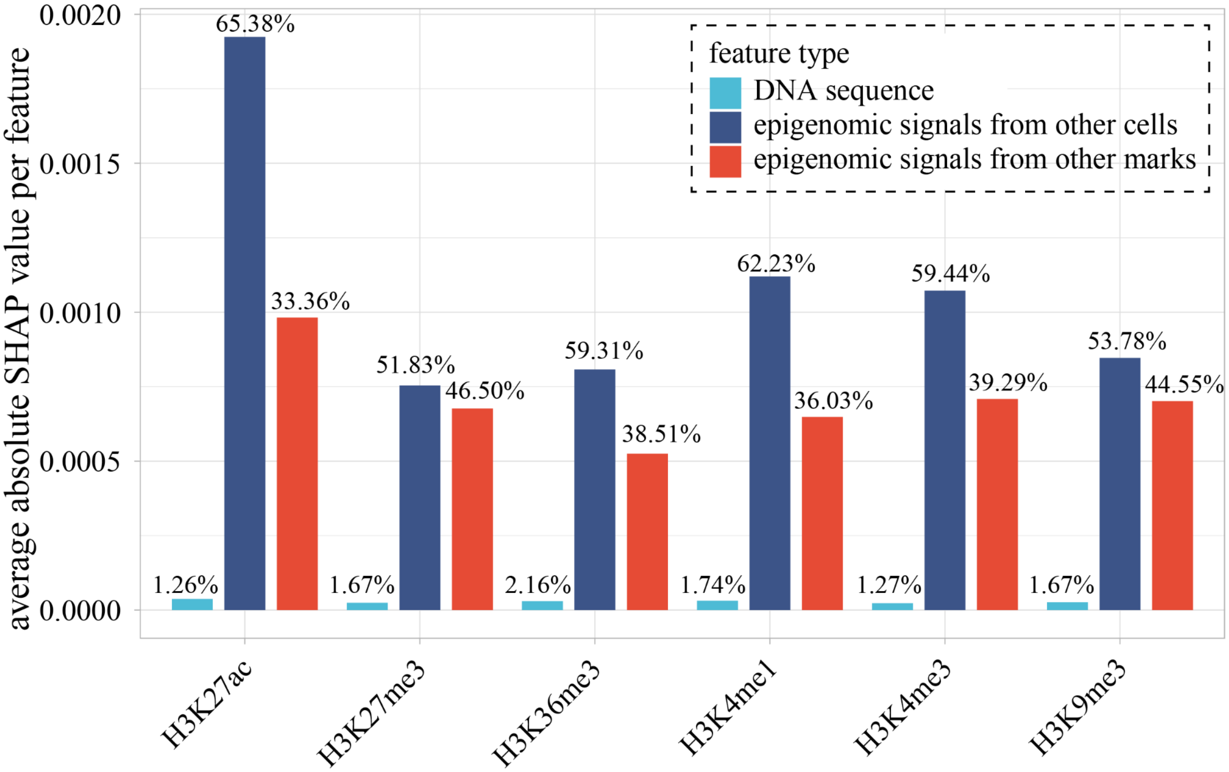
Supplementary Figure 19: Predictive performance comparison between lightGBM models with and without DNA sequence features

Predictive performance was evaluated on the largest Chromosome 1 using 12 scoring metrics on the ENCODE Imputation Challenge testing set of 51 mark-cell type pairs. For each metric, the paired one-sided Wilcoxon signed-rank test was performed between two methods across 51 testing mark-cell type pairs. The significantly different ones are labeled by asterisks (* p-value < 0.01 and ** p-value < 0.001).



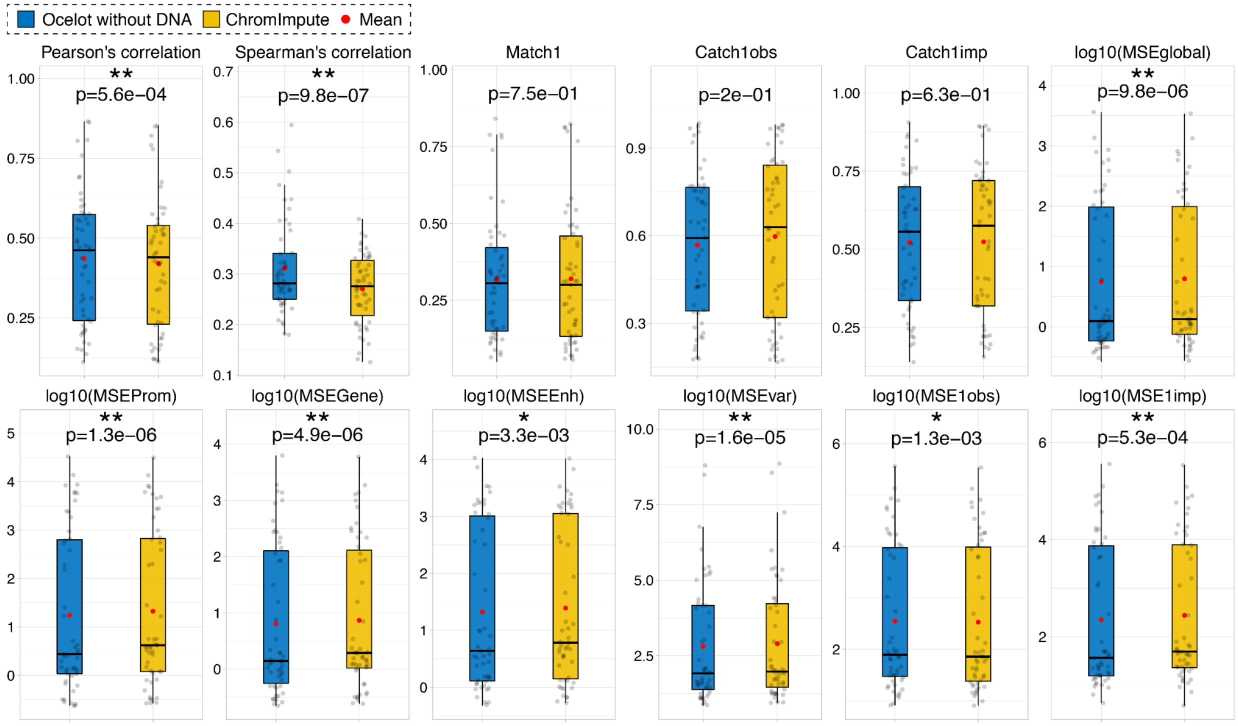
Supplementary Figure 20: Predictive performance comparison between neural network models with and without DNA sequence features

Predictive performance was evaluated on the largest Chromosome 1 using 12 scoring metrics on the ENCODE Imputation Challenge testing set of 51 mark-cell type pairs. For each metric, the paired one-sided Wilcoxon signed-rank test was performed between two methods across 51 testing mark-cell type pairs. The significantly different ones are labeled by asterisks (* p-value < 0.01 and ** p-value < 0.001).



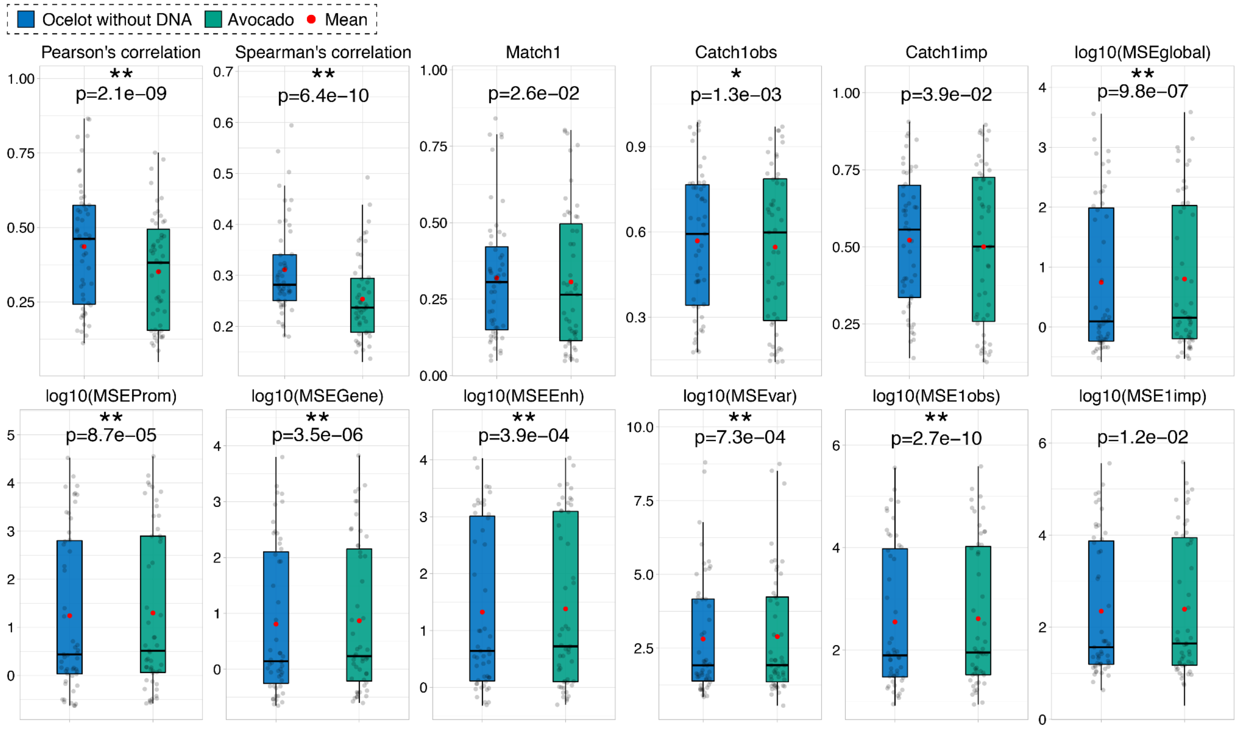
Supplementary Figure 21: The contributions to predictions of different types of features.

We used three types of input features in Ocelot: (1) DNA sequence, (2) epigenomic signals from other cell types, and (3) epigenomic signals from other marks. For each histone mark, the average absolute SHAP value per feature was calculated. The percentages are shown on top of each bar.



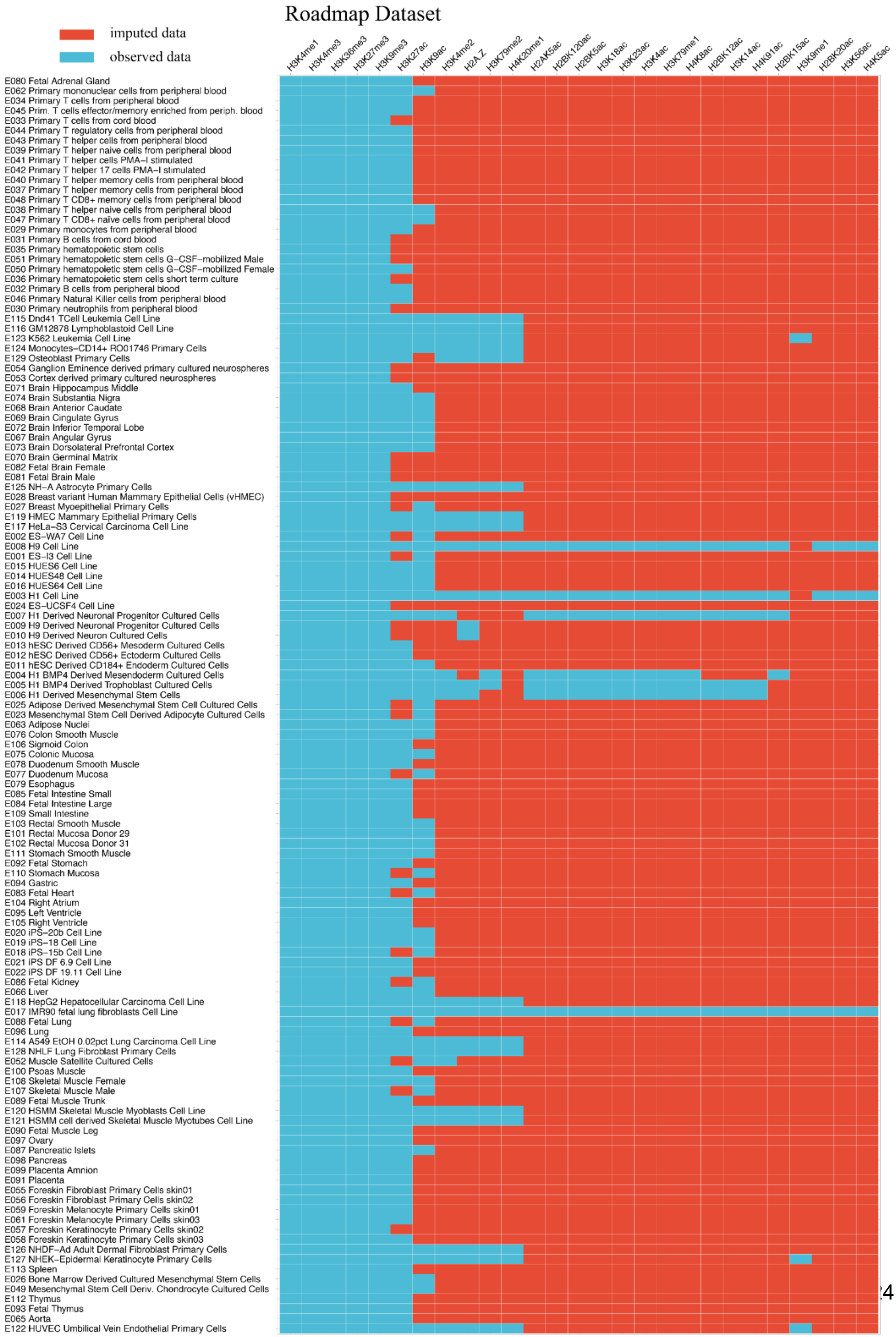
Supplementary Figure 22: Predictive performance comparison between Ocelot without DNA sequence features and ChromImpute

Predictive performance was evaluated on the whole human genome using 12 scoring metrics on the ENCODE Imputation Challenge testing set of 51 mark-cell type pairs. For each testing pair, we calculated genome-wide evaluation scores through concatenating signals of 23 chromosomes. For each metric, the paired one-sided Wilcoxon signed-rank test was performed between two methods across 51 testing mark-cell type pairs. The significantly different ones are labeled by asterisks (* p-value < 0.01 and ** p-value < 0.001).



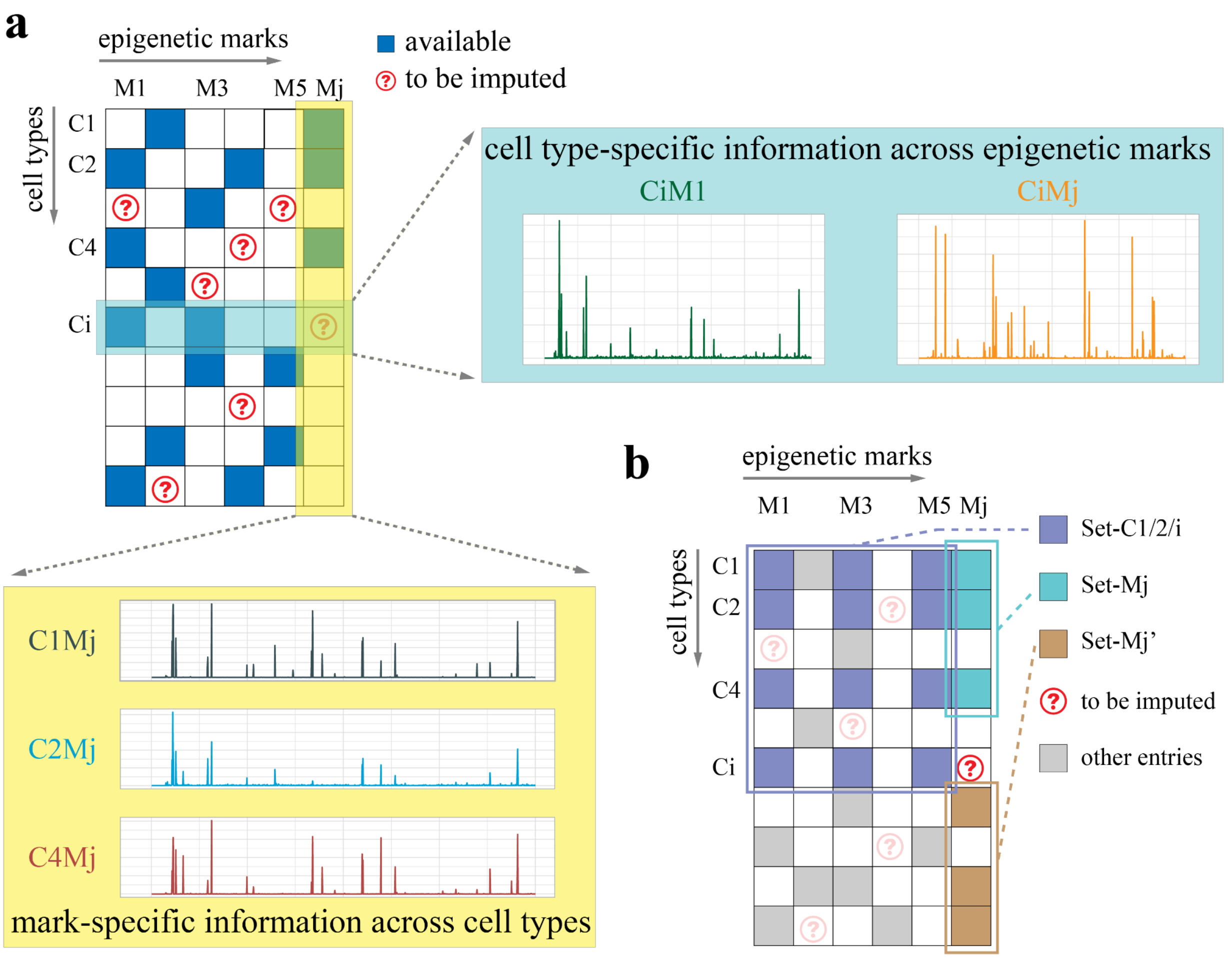
Supplementary Figure 23: Predictive performance comparison between Ocelot without DNA sequence features and Avocado

Predictive performance was evaluated on the whole human genome using 12 scoring metrics on the ENCODE Imputation Challenge testing set of 51 mark-cell type pairs. For each testing pair, we calculated genome-wide evaluation scores through concatenating signals of 23 chromosomes. For each metric, the paired one-sided Wilcoxon signed-rank test was performed between two methods across 51 testing mark-cell type pairs. The significantly different ones are labeled by asterisks (* p-value < 0.01 and ** p-value < 0.001).


Roadmap Dataset
imputed data
observed data
H3K4me1
H3K4me3
H3K36me3
H3K27me3
H3K9me3
H3K27ac
H3K9ac
H3K4me2
H2A.Z
H3K79me2
H4K20me1
H2AK5ac
H2BK120ac
H2BK5ac
H3K18ac
H3K23ac
H3K4ac
H3K79me1
H4K8ac
H2BK12ac
H3K14ac
H4K91ac
H2BK15ac
H3K9me1
H2BK20ac
H3K56ac
H4K5ac
E080 Fetal Adrenal Gland
E062 Primary mononuclear cells from peripheral blood
E034 Primary T cells from peripheral blood
E045 Prim. T cells effector/memory enriched from periph. blood
E033 Primary T cells from cord blood
E044 Primary T regulatory cells from peripheral blood
E043 Primary T helper cells from peripheral blood
E039 Primary T helper naive cells from peripheral blood
E041 Primary T helper cells PMA–I stimulated
E042 Primary T helper 17 cells PMA–I stimulated
E040 Primary T helper memory cells from peripheral blood
E037 Primary T helper memory cells from peripheral blood
E048 Primary T CD8+ memory cells from peripheral blood
E038 Primary T helper naive cells from peripheral blood
E047 Primary T CD8+ naïve cells from peripheral blood
E029 Primary monocytes from peripheral blood
E031 Primary B cells from cord blood
E035 Primary hematopoietic stem cells
E051 Primary hematopoietic stem cells G–CSF–mobilized Male
E050 Primary hematopoietic stem cells G–CSF–mobilized Female
E036 Primary hematopoietic stem cells short term culture
E032 Primary B cells from peripheral blood
E046 Primary Natural Killer cells from peripheral blood
E030 Primary neutrophils from peripheral blood
E115 Dnd41 TCell Leukemia Cell Line
E116 GM12878 Lymphoblastoid Cell Line
E123 K562 Leukemia Cell Line
E124 Monocytes–CD14+ RO01746 Primary Cells
E129 Osteoblast Primary Cells
E054 Ganglion Eminence derived primary cultured neurospheres
E053 Cortex derived primary cultured neurospheres
E071 Brain Hippocampus Middle
E074 Brain Substantia Nigra
E068 Brain Anterior Caudate
E069 Brain Cingulate Gyrus
E072 Brain Inferior Temporal Lobe
E067 Brain Angular Gyrus
E073 Brain Dorsolateral Prefrontal Cortex
E070 Brain Germinal Matrix
E082 Fetal Brain Female
E081 Fetal Brain Male
E125 NH–A Astrocyte Primary Cells
E028 Breast variant Human Mammary Epithelial Cells (vHMEC)
E027 Breast Myoepithelial Primary Cells
E119 HMEC Mammary Epithelial Primary Cells
E117 HeLa–S3 Cervical Carcinoma Cell Line
E002 ES–WA7 Cell Line
E008 H9 Cell Line
E001 ES–I3 Cell Line
E015 HUES6 Cell Line
E014 HUES48 Cell Line
E016 HUES64 Cell Line
E003 H1 Cell Line
E024 ES–UCSF4 Cell Line
E007 H1 Derived Neuronal Progenitor Cultured Cells
E009 H9 Derived Neuronal Progenitor Cultured Cells
E010 H9 Derived Neuron Cultured Cells
E013 hESC Derived CD56+ Mesoderm Cultured Cells
E012 hESC Derived CD56+ Ectoderm Cultured Cells
E011 hESC Derived CD184+ Endoderm Cultured Cells
E004 H1 BMP4 Derived Mesendoderm Cultured Cells
E005 H1 BMP4 Derived Trophoblast Cultured Cells
E006 H1 Derived Mesenchymal Stem Cells
E025 Adipose Derived Mesenchymal Stem Cell Cultured Cells
E023 Mesenchymal Stem Cell Derived Adipocyte Cultured Cells
E063 Adipose Nuclei
E076 Colon Smooth Muscle
E106 Sigmoid Colon
E075 Colonic Mucosa
E078 Duodenum Smooth Muscle
E077 Duodenum Mucosa
E079 Esophagus
E085 Fetal Intestine Small
E084 Fetal Intestine Large
E109 Small Intestine
E103 Rectal Smooth Muscle
E101 Rectal Mucosa Donor 29
E102 Rectal Mucosa Donor 31
E111 Stomach Smooth Muscle
E092 Fetal Stomach
E110 Stomach Mucosa
E094 Gastric
E083 Fetal Heart
E104 Right Atrium
E095 Left Ventricle
E105 Right Ventricle
E020 iPS–20b Cell Line
E019 iPS–18 Cell Line
E018 iPS–15b Cell Line
E021 iPS DF 6.9 Cell Line
E022 iPS DF 19.11 Cell Line
E086 Fetal Kidney
E066 Liver
E118 HepG2 Hepatocellular Carcinoma Cell Line
E017 IMR90 fetal lung fibroblasts Cell Line
E088 Fetal Lung
E096 Lung
E114 A549 EtOH 0.02pct Lung Carcinoma Cell Line
E128 NHLF Lung Fibroblast Primary Cells
E052 Muscle Satellite Cultured Cells
E100 Psoas Muscle
E108 Skeletal Muscle Female
E107 Skeletal Muscle Male
E089 Fetal Muscle Trunk
E120 HSMM Skeletal Muscle Myoblasts Cell Line
E121 HSMM cell derived Skeletal Muscle Myotubes Cell Line
E090 Fetal Muscle Leg
E097 Ovary
E087 Pancreatic Islets
E098 Pancreas
E099 Placenta Amnion
E091 Placenta
E055 Foreskin Fibroblast Primary Cells skin01
E056 Foreskin Fibroblast Primary Cells skin02
E059 Foreskin Melanocyte Primary Cells skin01
E061 Foreskin Melanocyte Primary Cells skin03
E057 Foreskin Keratinocyte Primary Cells skin02
E058 Foreskin Keratinocyte Primary Cells skin03
E126 NHDF–Ad Adult Dermal Fibroblast Primary Cells
E127 NHEK–Epidermal Keratinocyte Primary Cells
E113 Spleen
E026 Bone Marrow Derived Cultured Mesenchymal Stem Cells
E049 Mesenchymal Stem Cell Deriv. Chondrocyte Cultured Cells
E112 Thymus
E093 Fetal Thymus
E065 Aorta
E122 HUVEC Umbilical Vein Endothelial Primary Cells

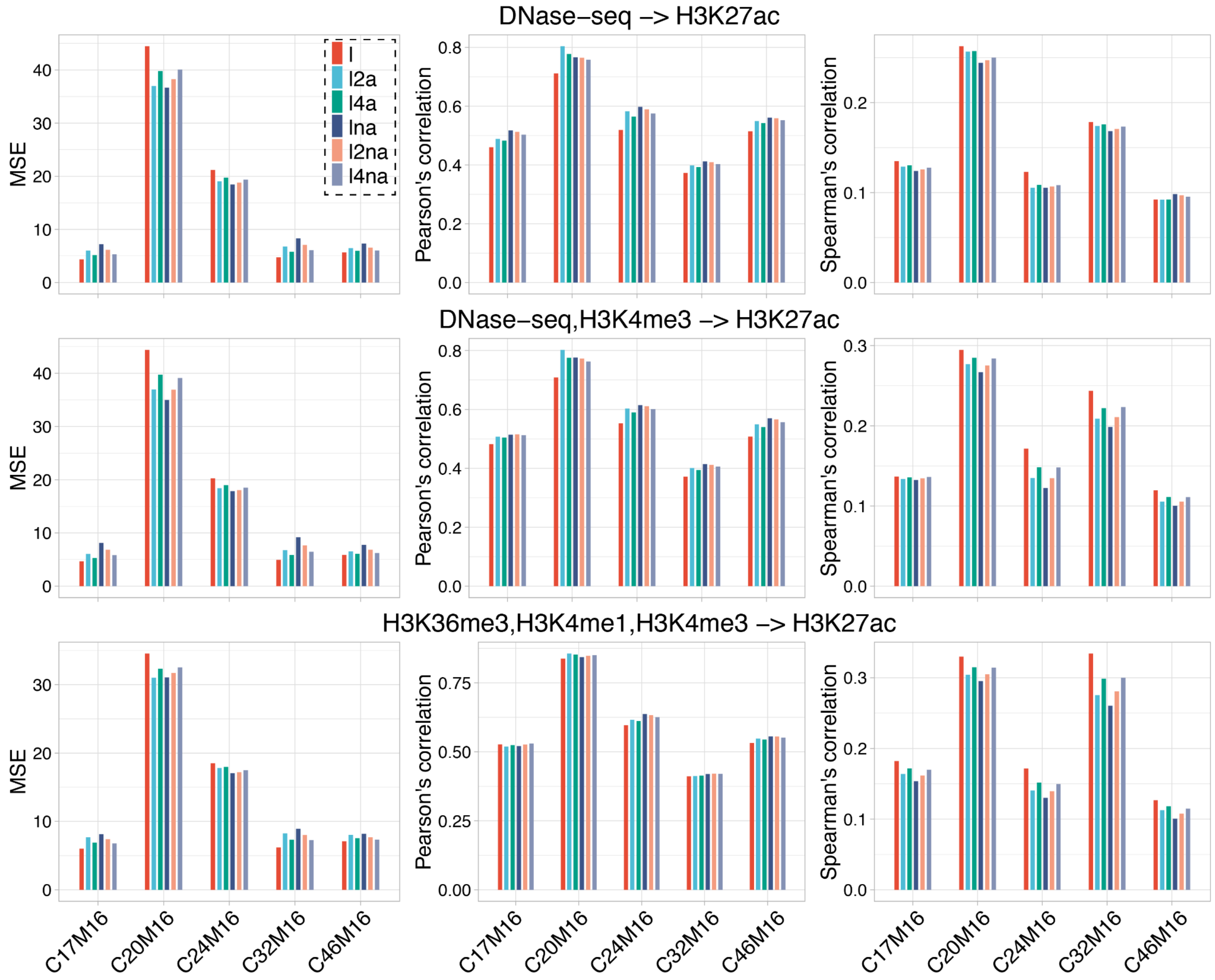
Supplementary Figure 24: Application of Ocelot to impute missing entries and complete the Roadmap Epigenomics dataset.

The Roadmap Epigenomics dataset covers 27 histone marks in 127 cell and tissue conditions after excluding four histone marks (H3K23me2, H2AK9ac, H3T11ph, H4K12ac) that only have one or two observed signal tracks. This is because with too few observed whole-genome tracks, we could not train a solid lightGBM or neural network model. A total of 974 (28.40%) whole-genome profiles were observed (blue blocks) and used to build machine learning models to predict the remaining 2,455 (71.60%) missing profiles (red blocks).



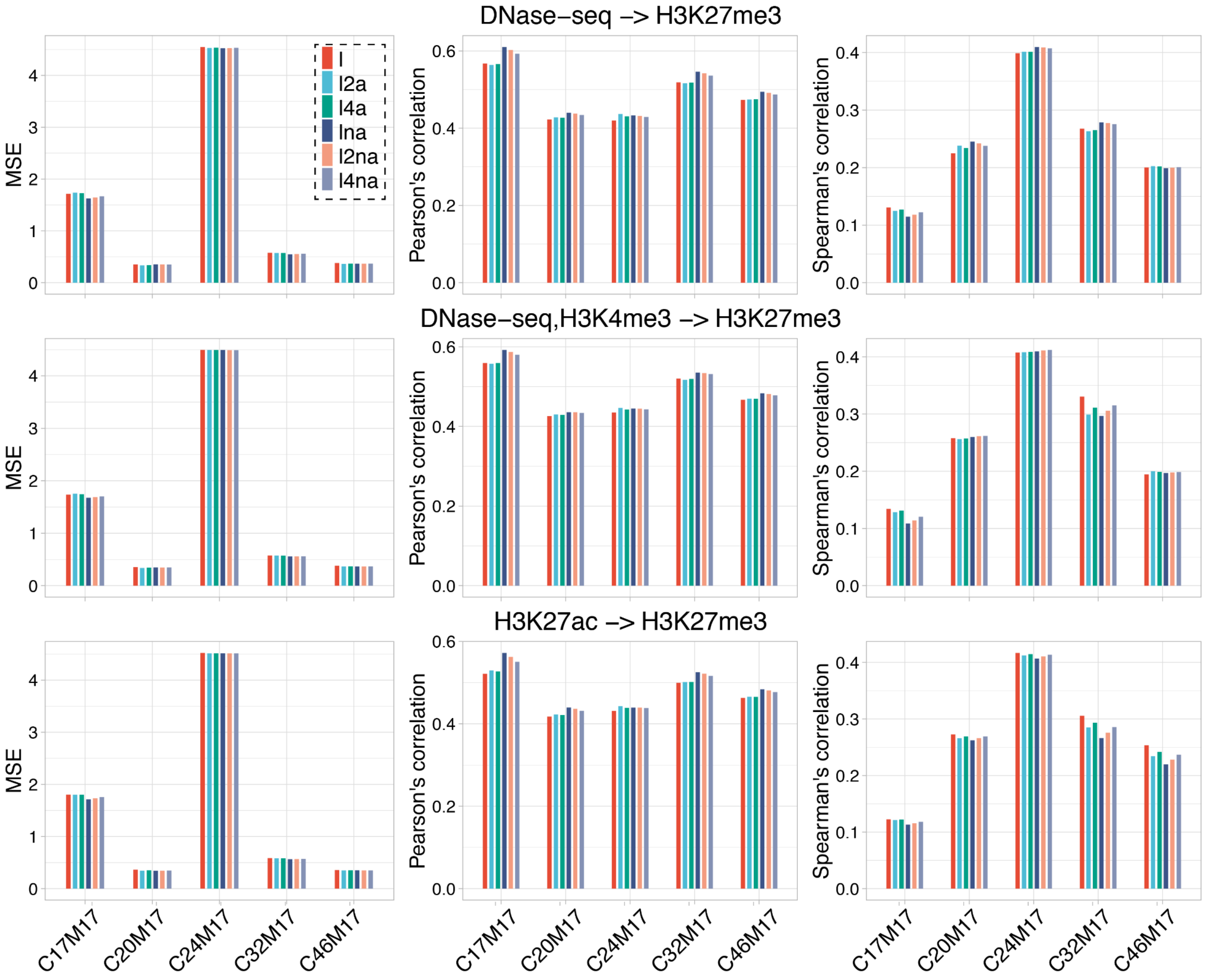
Supplementary Figure 25: Data partition for machine learning models.

a, The mark-specific and cell type-specific information is integrated into machine learning models to improve predictive performance. **b,** Based on the target entry to be imputed and available entries, the cell type - mark matrix can be partitioned into subset, serving as training targets, mark-specific features or cell type-specific features.



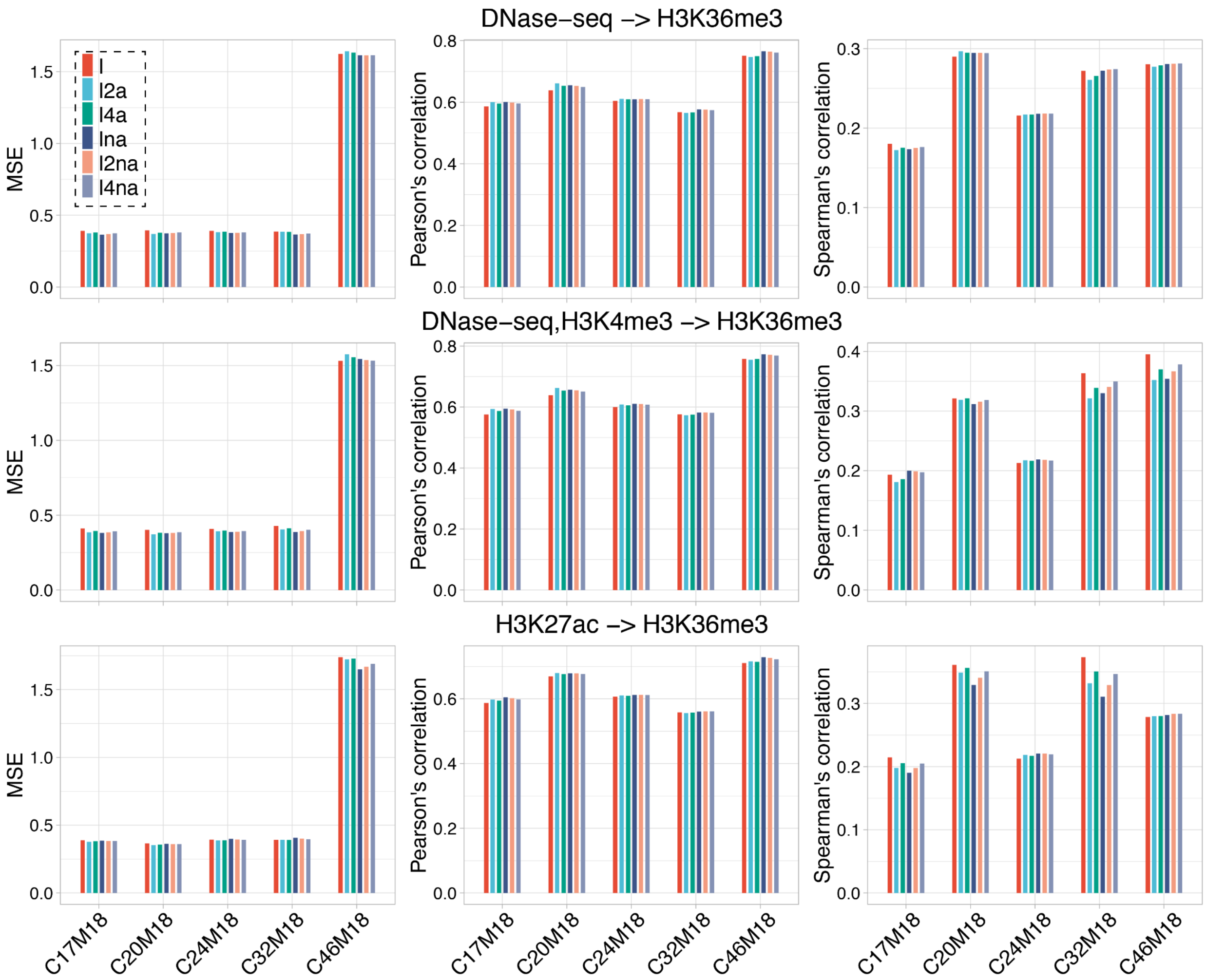
Supplementary Figure 26: Cross-validation results of predicting H3K27ac based on the challenge training data

Within the challenge training data, five cell types (C17, C20, C24, C32, C46) were held out as the testing set to evaluate predictive performance of different model design and ensemble weights. Three columns represent three evaluation metrics: (1) MSEglobal, (2) Pearson’s correlation, and (3) Spearman’s correlation. Multiple rows represent different model designs, where one or multiple marks are used as feature marks in machine learning models. Different colors represent different ensemble weights between lightGBM (“l”), neural network (“n”) and average signal track (“a”). For example, “l4na” represents the ensemble weights of l:n:a = 4:1:1.



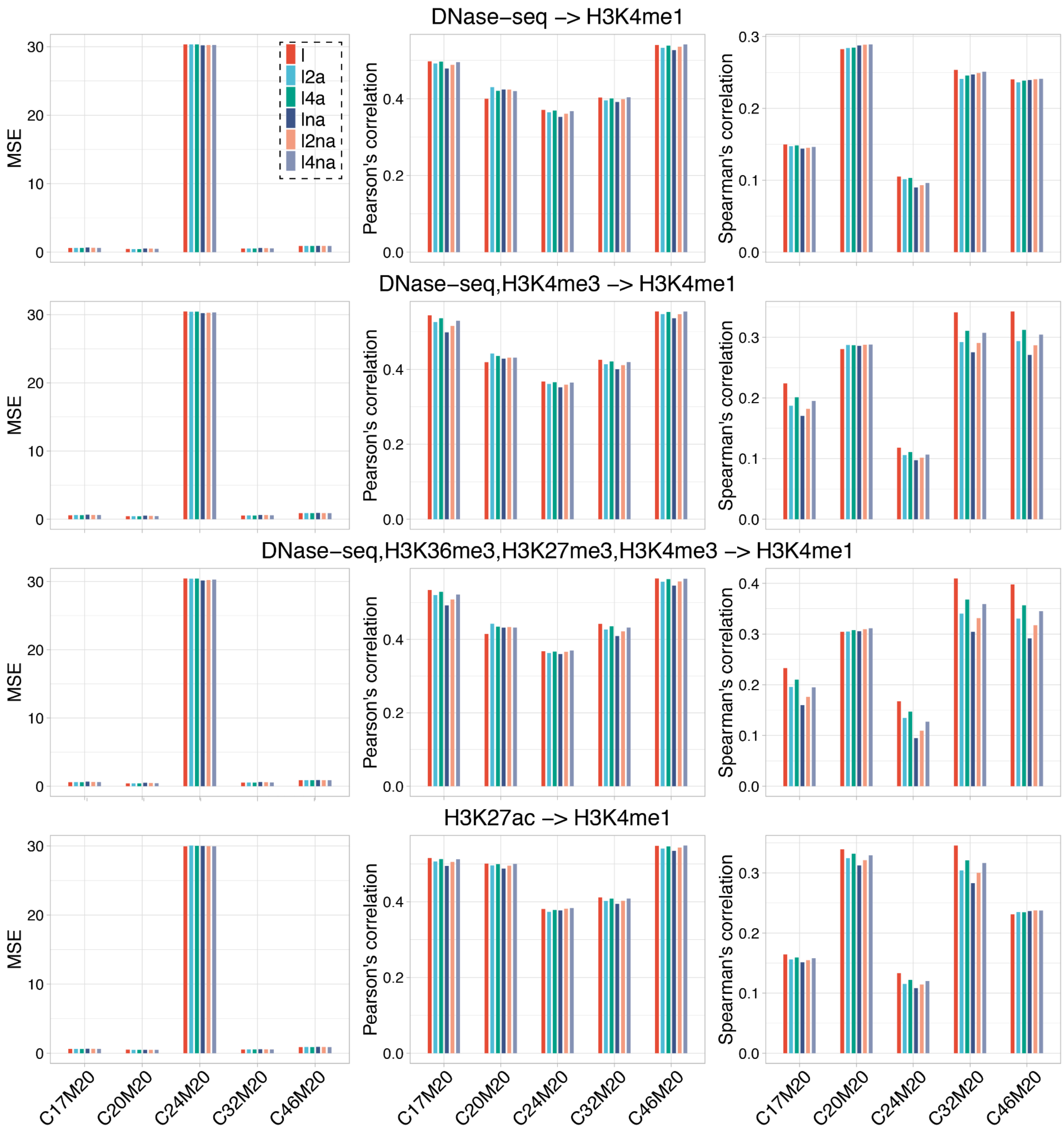
Supplementary Figure 27: Cross-validation results of predicting H3K27me3 based on the challenge training data

Within the challenge training data, five cell types (C17, C20, C24, C32, C46) were held out as the testing set to evaluate predictive performance of different model design and ensemble weights. Three columns represent three evaluation metrics: (1) MSEglobal, (2) Pearson's correlation, and (3) Spearman's correlation. Multiple rows represent different model designs, where one or multiple marks are used as feature marks in machine learning models. Different colors represent different ensemble weights between lightGBM ("l"), neural network ("n") and average signal track ("a"). For example, "l4na" represents the ensemble weights of l:n:a = 4:1:1.



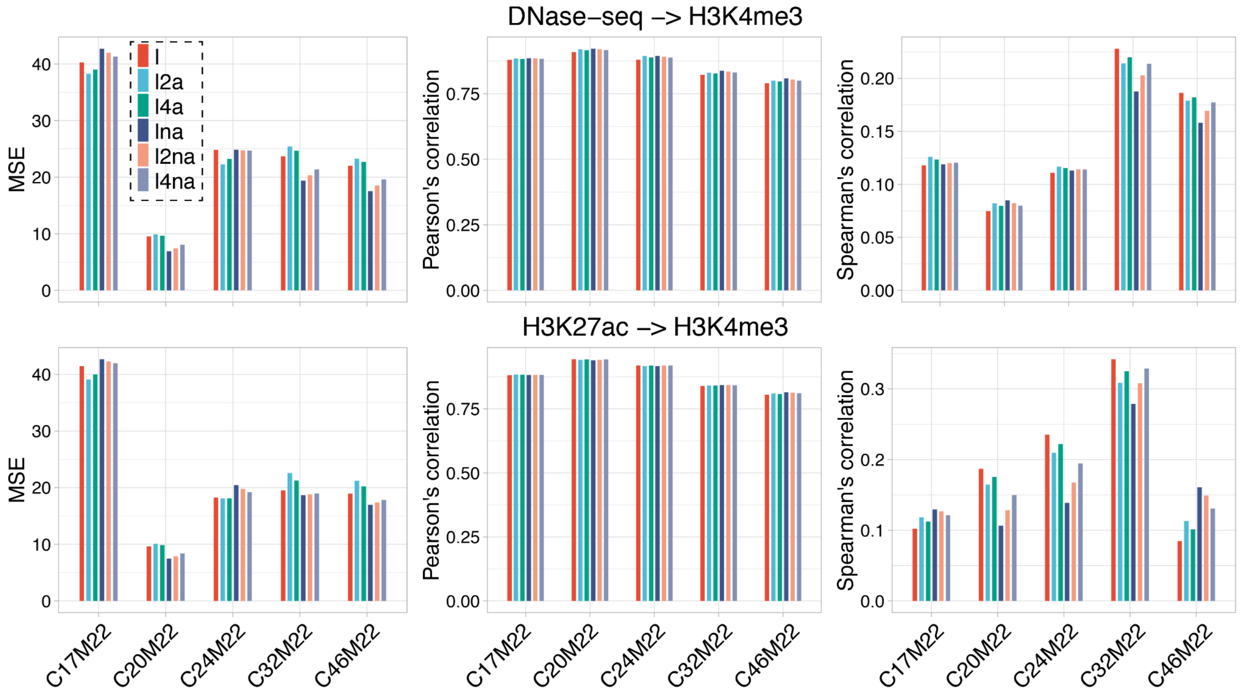
Supplementary Figure 28: Cross-validation results of predicting H3K36me3 based on the challenge training data

Within the challenge training data, five cell types (C17, C20, C24, C32, C46) were held out as the testing set to evaluate predictive performance of different model design and ensemble weights. Three columns represent three evaluation metrics: (1) MSEglobal, (2) Pearson’s correlation, and (3) Spearman’s correlation. Multiple rows represent different model designs, where one or multiple marks are used as feature marks in machine learning models. Different colors represent different ensemble weights between lightGBM (“l”), neural network (“n”) and average signal track (“a”). For example, “l4na” represents the ensemble weights of l:n:a = 4:1:1.



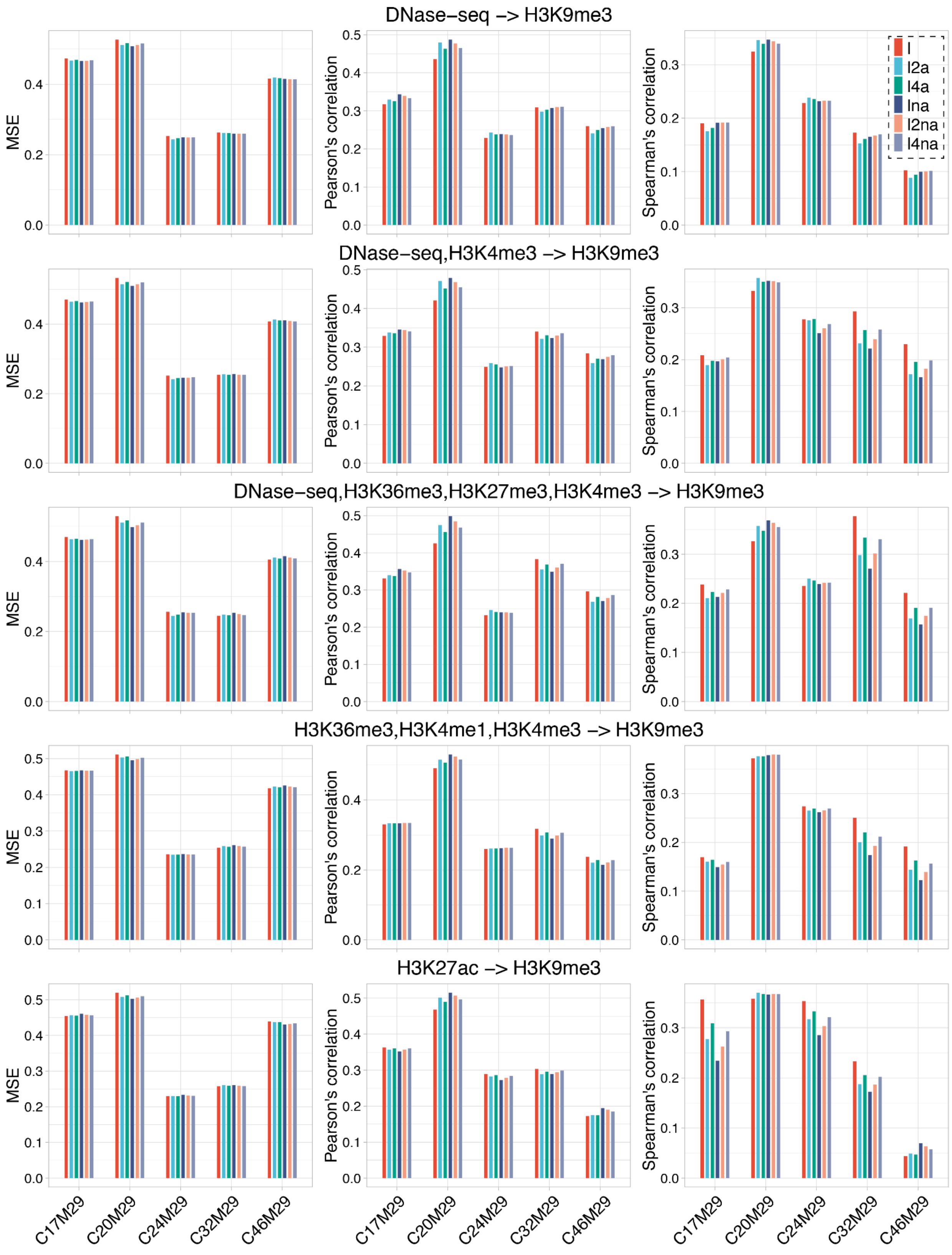
Supplementary Figure 29: Cross-validation results of predicting H3K4me1 based on the challenge training data

Within the challenge training data, five cell types (C17, C20, C24, C32, C46) were held out as the testing set to evaluate predictive performance of different model design and ensemble weights. Three columns represent three evaluation metrics: (1) MSEglobal, (2) Pearson's correlation, and (3) Spearman's correlation. Multiple rows represent different model designs, where one or multiple marks are used as feature marks in machine learning models. Different colors represent different ensemble weights between lightGBM ("l"), neural network ("n") and average signal track ("a"). For example, "l4na" represents the ensemble weights of l:n:a = 4:1:1.



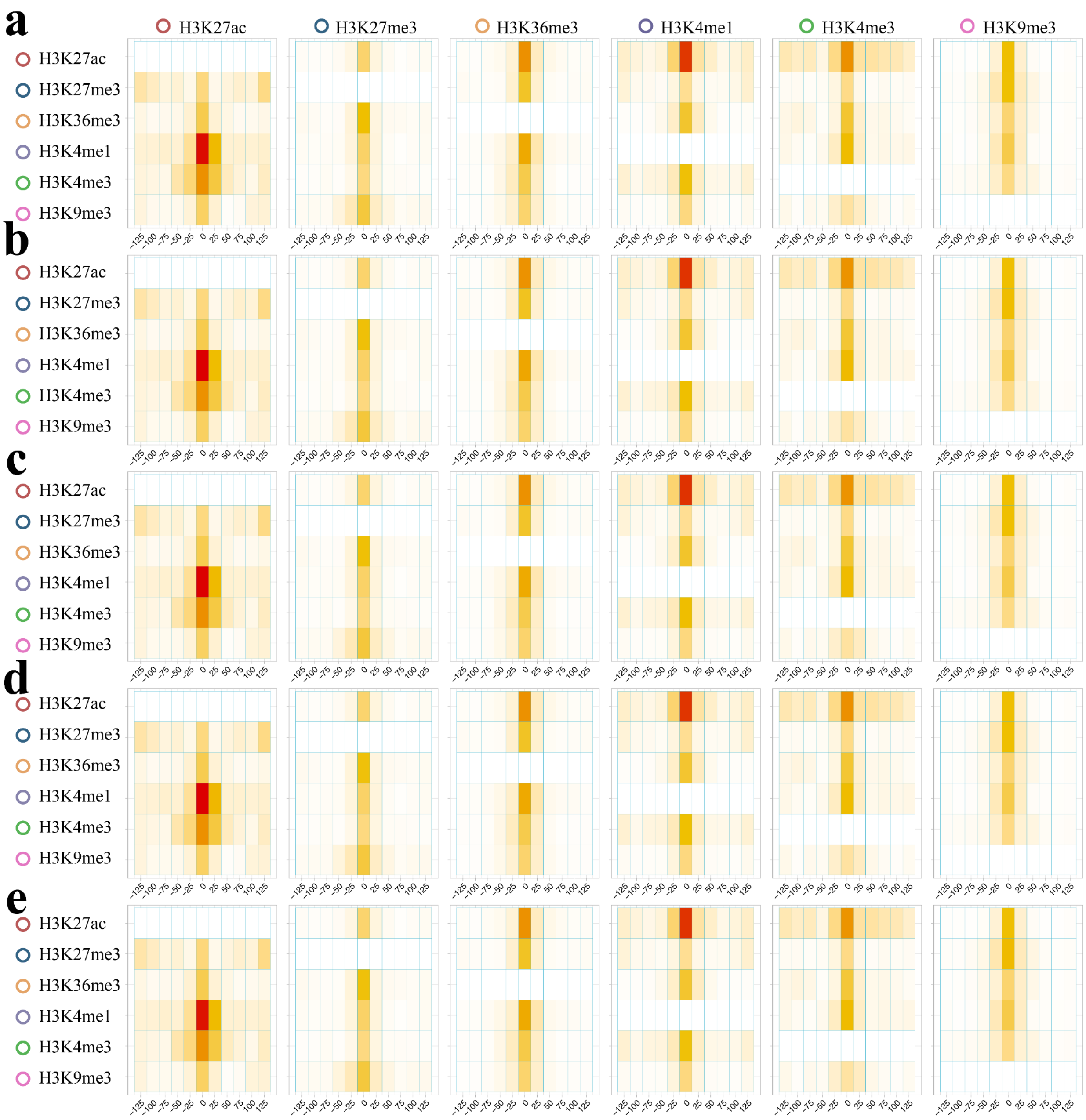
Supplementary Figure 30: Cross-validation results of predicting H3K4me3 based on the challenge training data

Within the challenge training data, five cell types (C17, C20, C24, C32, C46) were held out as the testing set to evaluate predictive performance of different model design and ensemble weights. Three columns represent three evaluation metrics: (1) MSEglobal, (2) Pearson's correlation, and (3) Spearman's correlation. Multiple rows represent different model designs, where one or multiple marks are used as feature marks in machine learning models. Different colors represent different ensemble weights between lightGBM ("l"), neural network ("n") and average signal track ("a"). For example, "l4na" represents the ensemble weights of l:n:a = 4:1:1.


DNase−seq −> H3K9me3
DNase−seq,H3K4me3 −> H3K9me3
DNase−seq,H3K36me3,H3K27me3,H3K4me3 −> H3K9me3
H3K36me3,H3K4me1,H3K4me3 −> H3K9me3
H3K27ac −> H3K9me3
MSE
Pearson's correlation
Spearman's correlation
0.0
0.1
0.2
0.3
0.4
0.5
l
l2a
l4a
lna
l2na
l4na
C17M29
C20M29
C24M29
C32M29
C46M29

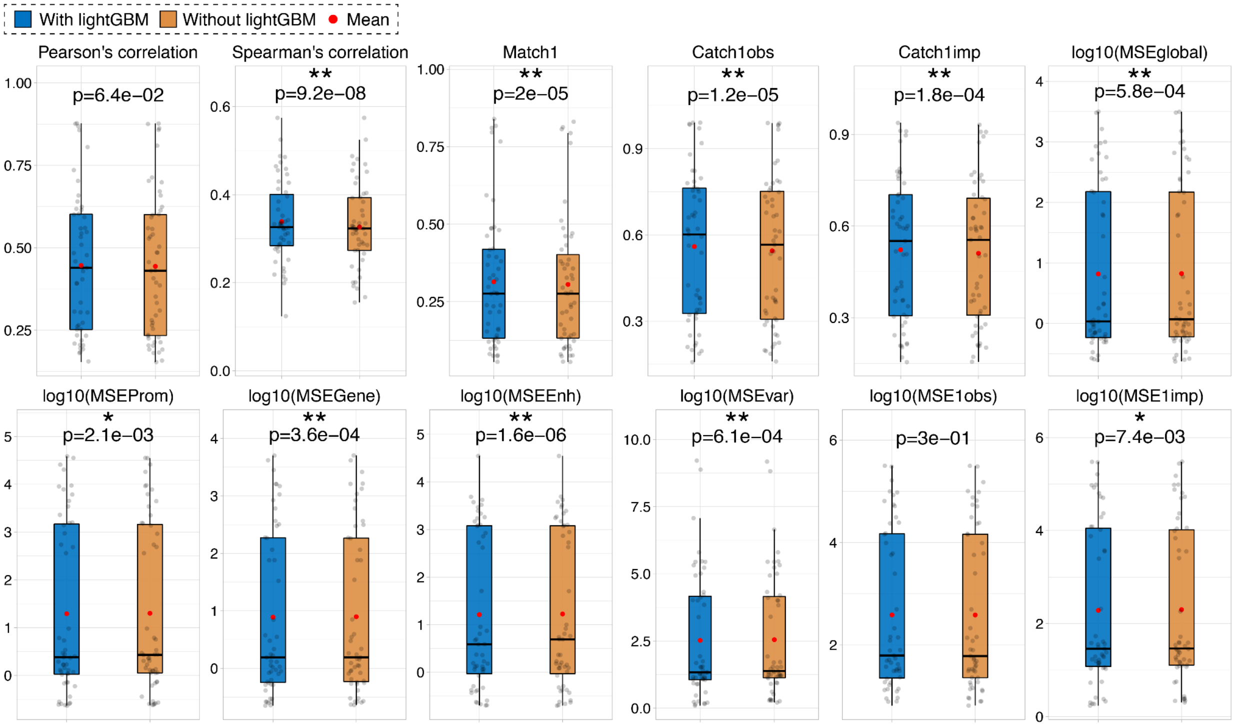
Supplementary Figure 31: Cross-validation results of predicting H3K9me3 based on the challenge training data

Within the challenge training data, five cell types (C17, C20, C24, C32, C46) were held out as the testing set to evaluate predictive performance of different model design and ensemble weights. Three columns represent three evaluation metrics: (1) MSEglobal, (2) Pearson's correlation, and (3) Spearman's correlation. Multiple rows represent different model designs, where one or multiple marks are used as feature marks in machine learning models. Different colors represent different ensemble weights between lightGBM ("l"), neural network ("n") and average signal track ("a"). For example, "l4na" represents the ensemble weights of l:n:a = 4:1:1.



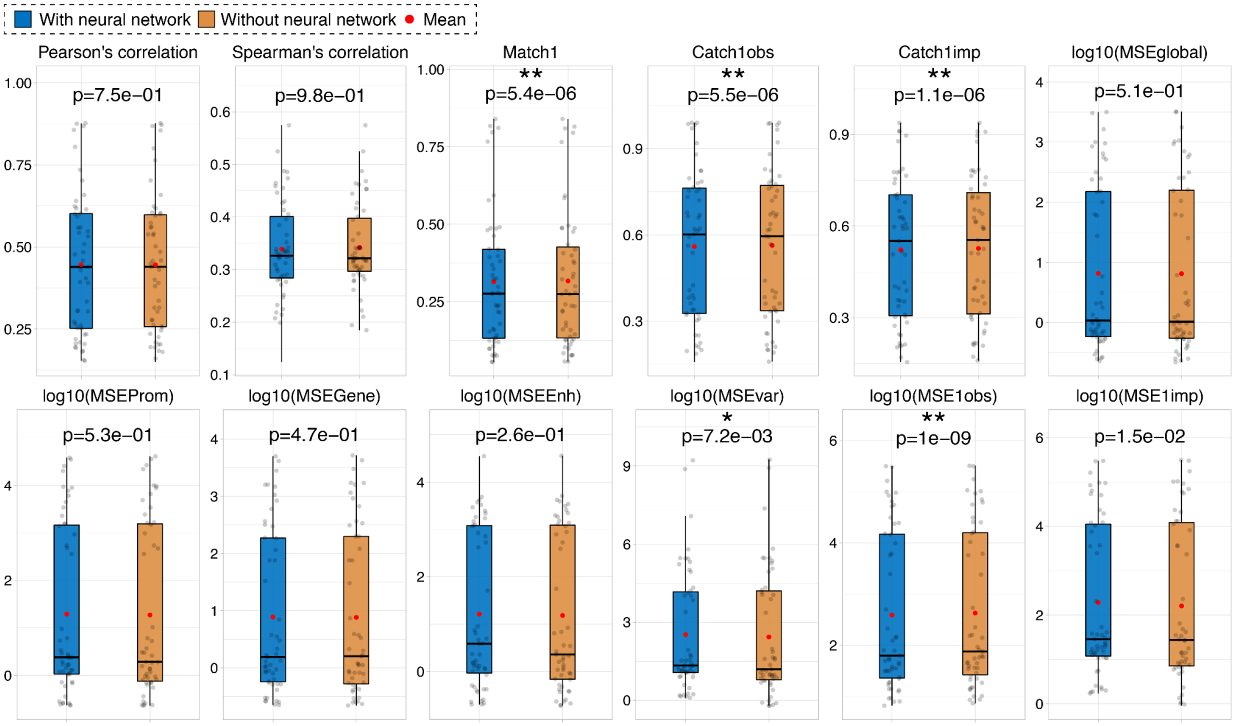
Supplementary Figure 32: The SHAP heatmaps among six histone marks based on different subsets of testing cell types.

Five cell types (C17, C20, C24, C32, C46) were held out as the testing set to perform SHAP analysis. In contrast to the SHAP heatmap in **Fig. 3a** that was derived from all five cell types, here only a subset of four cell types were used to calculate the SHAP heatmap. C17, C20, C24, C32 and C46 were excluded from the SHAP analysis in panels **a, b, c, d,** and **e,** respectively.



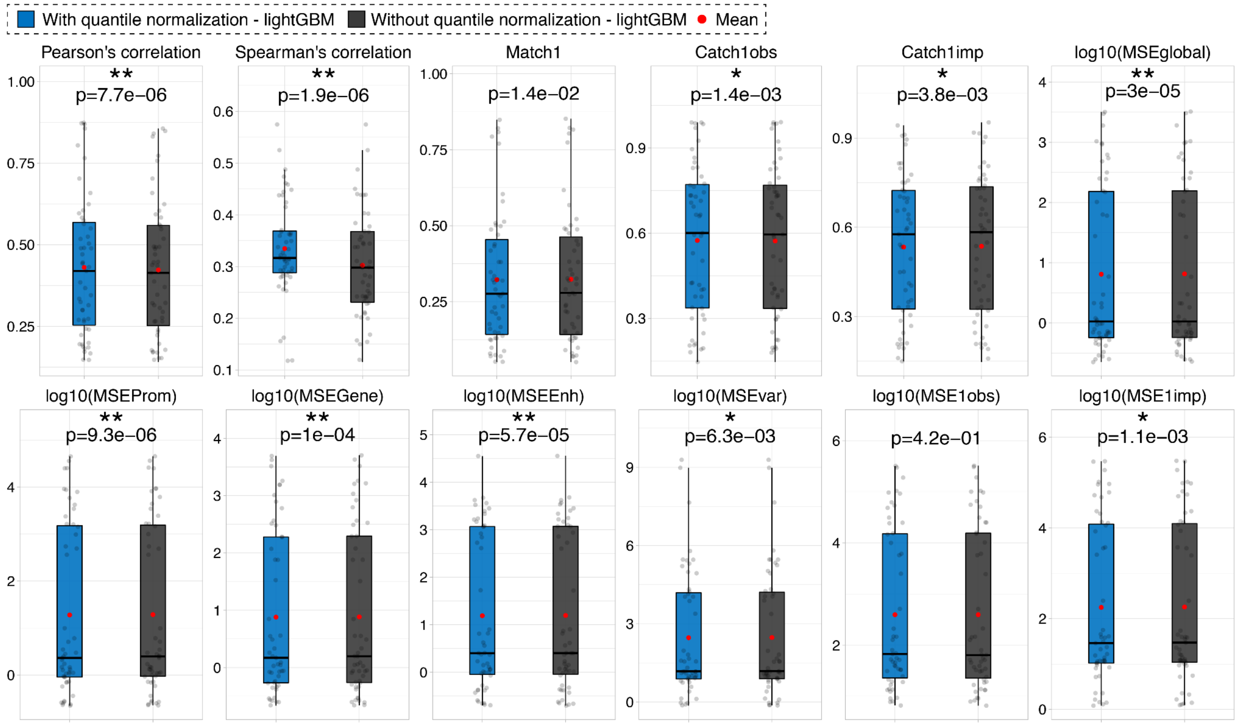
Supplementary Figure 33: Predictive performance comparison between ensemble models with and without lightGBM models

Predictive performance was evaluated on the largest Chromosome 1 using 12 scoring metrics on the ENCODE Imputation Challenge testing set of 51 mark-cell type pairs. For each metric, the paired one-sided Wilcoxon signed-rank test was performed between two methods across 51 testing mark-cell type pairs. The significantly different ones are labeled by asterisks (* p-value < 0.01 and ** p-value < 0.001).



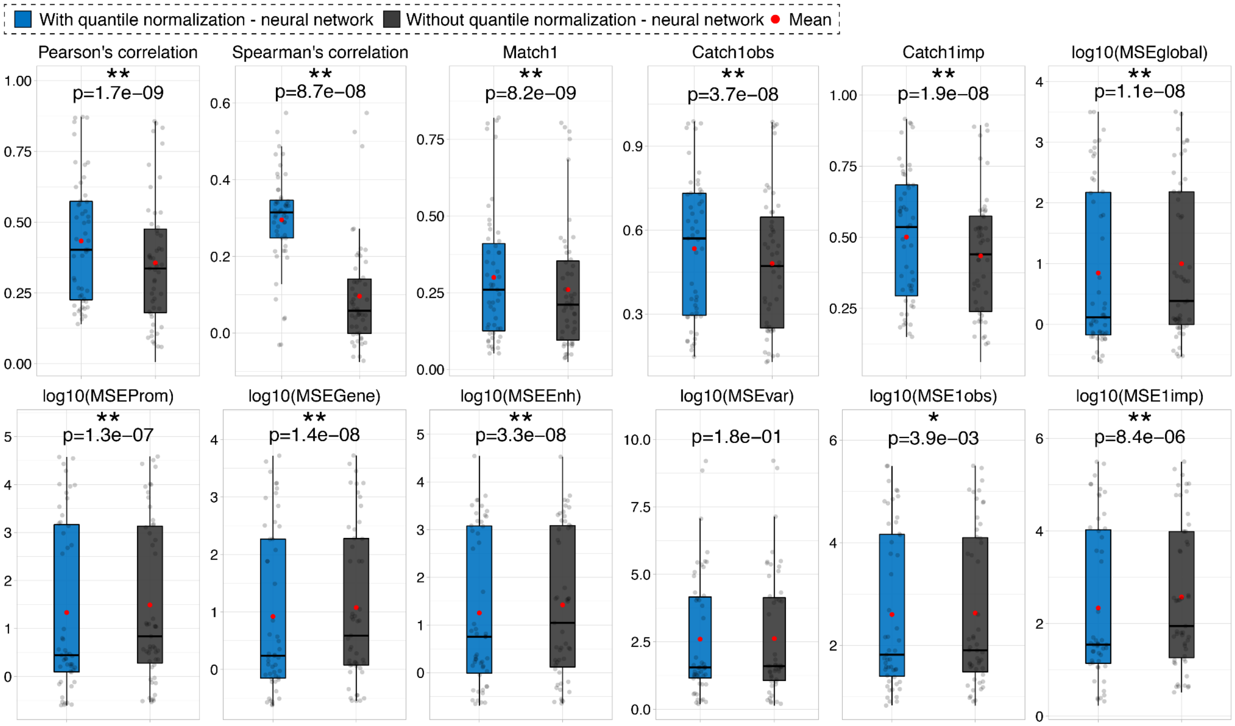
Supplementary Figure 34: Predictive performance comparison between ensemble models with and without neural network models

Predictive performance was evaluated on the largest Chromosome 1 using 12 scoring metrics on the ENCODE Imputation Challenge testing set of 51 mark-cell type pairs. For each metric, the paired one-sided Wilcoxon signed-rank test was performed between two methods across 51 testing mark-cell type pairs. The significantly different ones are labeled by asterisks (* p-value < 0.01 and ** p-value < 0.001).



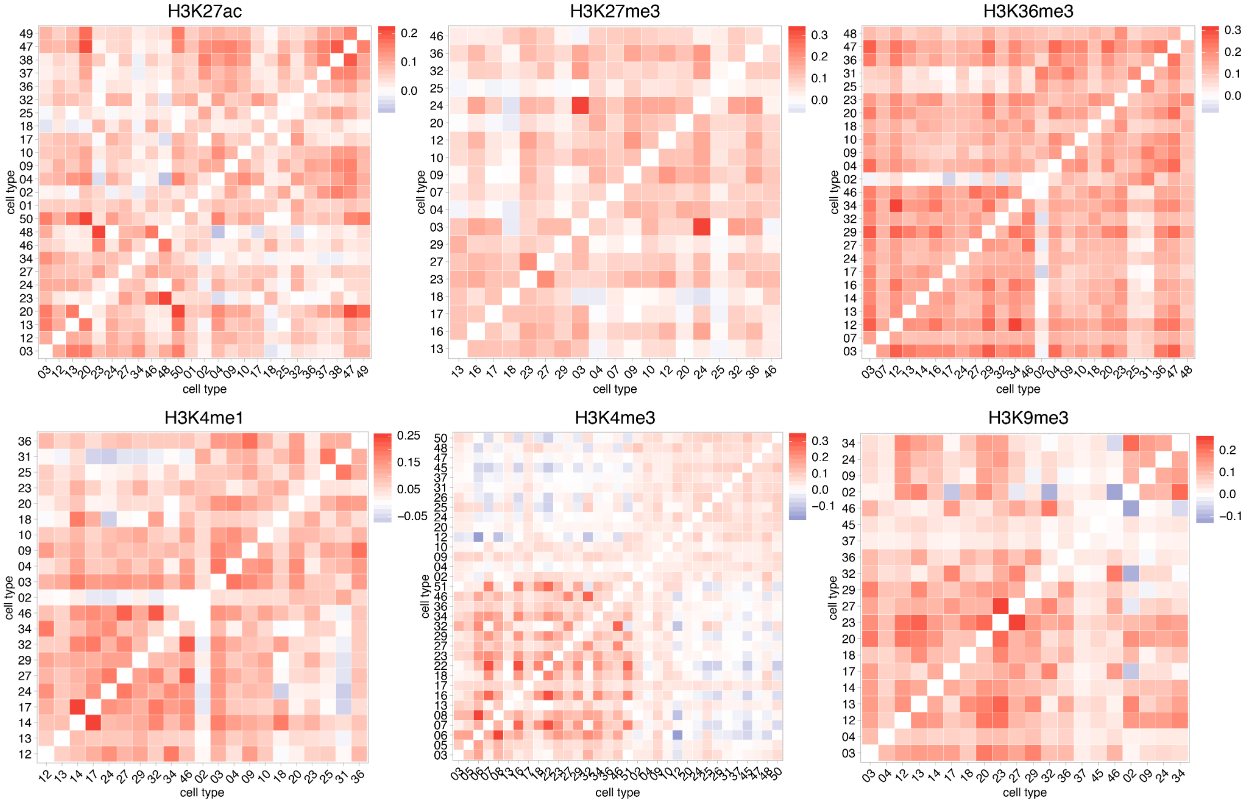
Supplementary Figure 35: Predictive performance comparison between lightGBM models with and without quantile normalization

Predictive performance was evaluated on the largest Chromosome 1 using 12 scoring metrics on the ENCODE Imputation Challenge testing set of 51 mark-cell type pairs. For each metric, the paired one-sided Wilcoxon signed-rank test was performed between two methods across 51 testing mark-cell type pairs. The significantly different ones are labeled by asterisks (* p-value < 0.01 and ** p-value < 0.001).



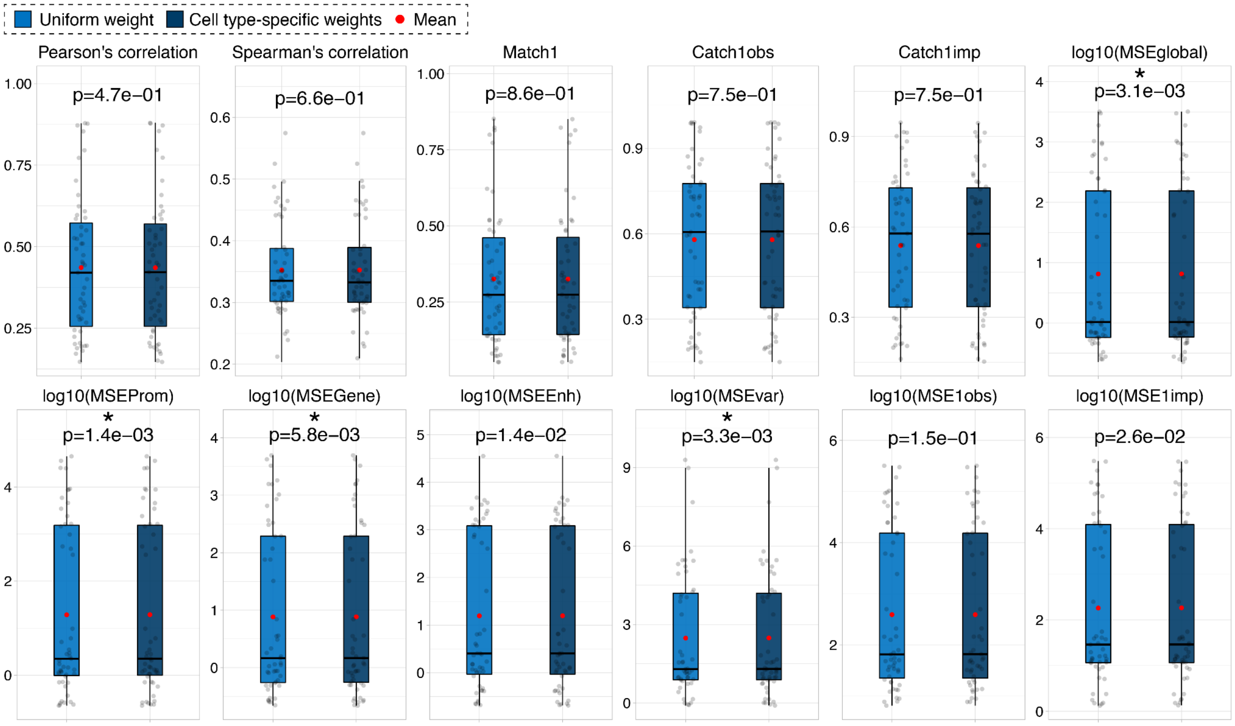
Supplementary Figure 36: Predictive performance comparison between neural network models with and without quantile normalization

Predictive performance was evaluated on the largest Chromosome 1 using 12 scoring metrics on the ENCODE Imputation Challenge testing set of 51 mark-cell type pairs. For each metric, the paired one-sided Wilcoxon signed-rank test was performed between two methods across 51 testing mark-cell type pairs. The significantly different ones are labeled by asterisks (* p-value < 0.01 and ** p-value < 0.001).



Supplementary Figure 37: The pairwise Pearson's correlations of epigenomic signals among multiple training cell types.

For each histone mark and each pair of available cell types, the pairwise Pearson correlation of epigenomics signals was calculated. Positive and negative correlations are shown in red and blue respectively.



Supplementary Figure 38: Predictive performance comparison between ensemble predictions based on a uniform weight and ensemble predictions based on cell type-specific weights.

The cell type-specific weights were defined by Pearson's correlations among cell types of marks that were used as features in machine learning models. Predictive performance was evaluated on the largest Chromosome 1 using 12 scoring metrics on the ENCODE Imputation Challenge testing set of 51 mark-cell type pairs. For each metric, the paired one-sided Wilcoxon signed-rank test was performed between two methods across 51 testing mark-cell type pairs. The significantly different ones are labeled by asterisks (* p-value < 0.01 and ** p-value < 0.001).

Supplementary Table Legends

Supplementary Table 1: The SHAP values from 10 bootstrap samplings.

The results from 10 bootstraps are shown in rows labeled by “bootstrap 0”, “bootstrap 1”, …, “bootstrap 9”. The average values are shown in the last 10 rows labeled by “average”. The histone marks in the first row represent that they are serving as the target being predicted and the histone marks along the column represent that they are serving as predictors to predict others. The distances are calculated from the target bin to be predicted.

Supplementary Table 2: Predictive performance comparison between Ocelot, ChromImpute and Avocado.

Supplementary Table 3: Predictive performance comparison between lightGBM models with and without DNA sequence features.

Supplementary Table 4: Predictive performance comparison between neural network (NN) models with and without DNA sequence features.

Supplementary Table 5: Predictive performance comparison between Ocelot without DNA sequence features, ChromImpute and Avocado.

Supplementary Table 6: The data matrix of the ENCODE Imputation Challenge.

The marks are shown in the column and the tissue and cell types are shown in the row. The green blocks with “T” are the training data and the red blocks with “B” are the blind testing data. The other empty blocks are unavailable mark-cell type combinations.

Supplementary Table 7: The model design and ensemble weights of Ocelot final submission in the ENCODE Imputation Challenge.

Based on the availability of training data in **Supplementary Table 6**, we design a specific model that includes as many feature marks as possible in predicting a target mark in a specific cell type. The ensemble weights of (1) lightGBM, (2) Neural Network (NN), and (3) the quantile normalized average signal of a mark in all training cell types were selected based on the cross-validation result on the training data during the challenge.

Supplementary Table 8: The SHAP values of 5 subsets.

Supplementary Table 9: The PPI values of 5 subsets.

Supplementary Table 10: Predictive performance comparison between ensemble models, ensemble models without lightGBM models, and ensemble models without neural network (NN) models.

Supplementary Table 11: Predictive performance comparison between lightGBM models with and without quantile normalization (QN).

Supplementary Table 12: Predictive performance comparison between neural network (NN) models with and without quantile normalization (QN).

Supplementary Table 13: Predictive performance comparison between lightGBM predictions based on a uniform weight and lightGBM predictions based on cell type-specific weights.

Supplementary Table 14: The predictive performance of the top two solutions in the ENCODE Imputation Challenge.

Nine columns represent nine evaluation metrics used in the challenge and the median scores across 51 testing mark-cell type pairs are listed. The percentage improvements (increase in correlation metrics and decrease in MSE metrics) are shown in the third row.

Supplementary Table 15: The accession numbers of the ENCODE3 data used in this study.