Supplementary Materials for

Deep learning identifies and quantifies the recombination hotspot determinants

Yu Li[1,2,3,4,#,*], Siyuan Chen[2,3,#], Trisevgeni Rapakoulia[5], Hiroyuki Kuwahara[2,3], Kevin Y. Yip[1], Xin Gao[2,3*]

[1]Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong SAR, China.

[2]Computer Science Program, Computer, Electrical and Mathematical Sciences and Engineering (CEMSE) Division, King Abdullah University of Science and Technology (KAUST), Thuwal, 23955-6900, Kingdom of Saudi Arabia

[3]KAUST Computational Bioscience Research Center (CBRC), King Abdullah University of Science and Technology

[4]The CUHK Shenzhen Research Institute, Hi-Tech Park, Nanshan, Shenzhen 518057, China.

[5]Max Planck Institute for Molecular Genetics, Ihnestrasse 63-73, 14195 Berlin, Germany.

[#]These authors contributed equally to this work.

*Correspondence: liyu@cse.cuhk.edu.hk (Y.L.) and xin.gao@kaust.edu.sa (X.G.).

Section 1. Statistics of datasets

a. The Mice Dataset construction

The Mice dataset(Lange, et al., 2016) is constructed according to the SPO11-oligo maps from C57BL/6J. Soluble protein is subjected to two successive rounds of affinity purification with a monoclonal anti-mSPO11 antibody and protein A-agarose beads. Sampled hotspots are identified with 0.000377 RPM/bp, which is 50 times of the average reads per million (RPM) within the mappable GRCm38/mm10 genome. The selected hotspots are further cropped into 1000bp for deep learning purposes, resulting in 9,620 hotspot sequences. Similarly, the coldspot region is equally selected with $1 * 10^{-7}$ times of the average reads per million (RPM) within the mappable GRCm38/mm10 genome region and cropped with the same length

b. The Yeast Dataset construction

In order to verify our method's generalization ability on species lacking the PRDM9 gene, we further construct Yeast(Mancera, et al., 2008) (Saccharomyces cerevisiae) hotspots from nearly 52,000 markers in all the four viable spores derived from 51 meiosis of an S288c/YJM789 hybrid strain. Pairs of genotype changes isolated from all other changes are called NCOs if they appear on the same spore, or COs if they appear on two spores. In total, 468 meiotic hotspots (averaging 842bp in length) that contain 92 COs and 74 NCO are cropped from the S288C Yeast reference genome. As S288C Yeast reference genome is much shorter than that of humans and mice, we define the corresponding recombination coldspots as the gap sequences between two recombination hotspots with at least 1000bp away from hotspots. Statistical comparison between the hotspots data across different species can be found in **Supplementary Table S2.**

c. The 1000 Genome Dataset construction

We utilize the population-wise recombination maps generated from the 1000 Genomes Dataset(Genomes Project, et al., 2015) to conduct the direct meiotic recombination hotspot prediction as well as downstream analysis. Fine-scale maps having an average resolution of 711 bp based on 26 diverse human populations are further merged into five super-populations: African (AFR), admixed American (AMR), East Asian (EAS), European (EUR), and South Asian (SAS). The merged five super-populations share a high Spearman correlation ranging from $\rho = 0.986$ to $\rho = 0.998$ within each category. Therefore, we further map the recombination hotspot regions within each super-population category to the GRCh38 reference genome, and generate corresponding hotspots for each population (AFR 50,049 hotspots avg 27.46cM/ Mb; AMR 18,160 hotspots avg 27.32cM/Mb; EAS 27,030 hotspots avg 38.44cM/ Mb; EUR 31,283 hotspots avg 35.05cM/ Mb; SAS 32,593 hotspots avg 33.92cM/ Mb). Similar to the Icelandic(Halldorsson, et al., 2019) dataset, we select the coldspot sequences from the lowest recombination rate regions of the recombination map (AFR 50,049 coldspots avg 0.0378cM/Mb; AMR 18,152 coldspots avg 0.039cM/Mb; EAS 27,020 coldspots avg 0.094cM/Mb; EUR 31,273 coldspots avg 0.038cM/Mb; SAS 32,583 coldspots avg 0.033cM/Mb). Statistical comparison between the generated hotspots and coldspots data across different populations can be found in **Supplementary Table S3**.

d. Sex-specific Feature

When targeting sex-specific recombination prediction, we use the ChIP-seq features as extra information for the proposed RHSNet. Following the previous research(Halldorsson, et al., 2019), we use histone modifications from ovary for the maternal map and testis for the

paternal map. On the Icelandic(Halldorsson, et al., 2019) dataset, we define the hotspot ChIP-seq feature as the closest narrow peak next to the hotspot sequence found in six different kinds of histone modifications(Dekker, et al., 2017) (H3K4me1(Yamada, et al., 2013), H3K4me3, H3K27ac(Chen, et al., 2020), H3K9me3, H3K36me3, H3K27me3). Similarly, we define the coldspot feature as the closest narrow peak next to the coldspot sequence. Specifically, the feature vector is set to zero when the actual distance exceeds 10kbp.

	5-fold Cross-validation		**Imbalanced Testing**	
	Hotspots	Coldspots	Hotspots	Coldspots
Icelandic 2019(Halldorsson, et al., 2019)	20,000 51.07cM/Mb	20,000 1.78e10-10 cM/Mb	5,467 43.18cM/Mb	95,733 0.07 cM/Mb
HapMap II 2008(Frazer, et al., 2007)	17,552 10.5cM/Mb	17,547 0.5cM/Mb	——	——
Sperm 2020(Bell, et al., 2020)	5,000 19.96cM/Mb	5,000 1e10-20 cM/Mb	——	——

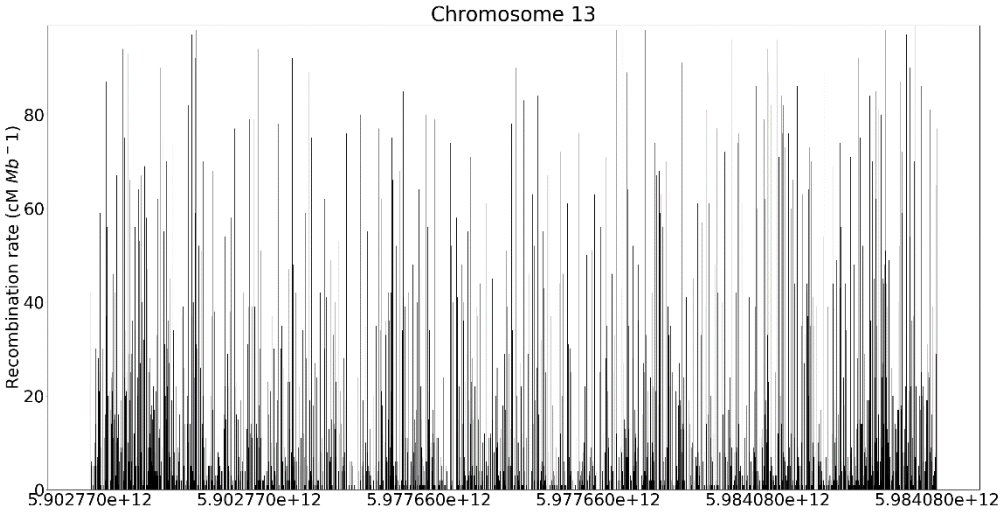
Supplementary Table S1. Statistical comparison of our hotspots/coldspots construction across different studies on the human genome, with an imbalanced testing dataset on Icelandic(Halldorsson, et al., 2019) 2019 dataset. The average recombination rate of each study is attached under each row.

	Hotspots	**Coldspots**
Icelandic Paternal ♂	15,000 44.13cM/Mb	15,000 1.2e10-14cM/Mb
Icelandic Maternal ♀	20,000 48.28cM/Mb	20,000 1.8e10-11cM/Mb
Mouse(Lange, et al., 2016)	9,620	9,620
Yeast	468	468

Supplementary Table S2. Statistical comparison of dataset construction across different sexes and different species.

Population Code	Population	Hotspots	Coldspots
AFR	African	50,049 27.45cM/Mb	50,035 0.0378cM/Mb
AMR	Admixed American	18,160 27.32cM/Mb	18,152 0.0390cM/Mb
EAS	East Asian	27,030 38.44cM/Mb	27.020 0.094cM/Mb
EUR	European	31,283 35.50cM/Mb	81,273 0.0380cM/Mb
SAS	South Asian	32,593 33.91cM/Mb	32,583 0.0383cM/Mb

Supplementary Table S3. Statistical comparison of the 1000 Genome(Genomes Project, et al., 2015) dataset construction across five populations with corresponding recombination rate over each population.

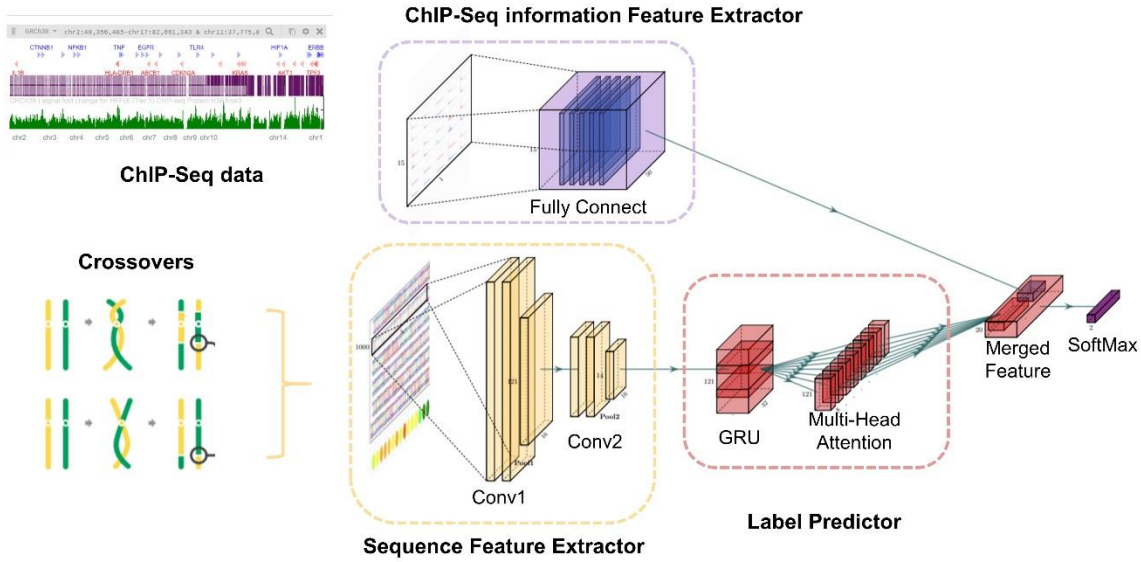


Supplementary Figure S1. The recombination rate distribution over chromosome 13 from the Icelandic dataset. The resolution is set as 100kbp.

Section 2. Supplementary methods

a. The deep learning network architecture

The deep learning network architecture consists of two independent feature extractors. The sequence feature extractor first encodes DNA sequences into one-hot matrices. To represent each nucleotide, we define the encoding as a vector of size 4, A as (1 0 0 0), T as (0 0 0 1), G as (0 0 1 0), and C as (0 1 0 0). Next, two convolution layers are introduced as a feature extractor to encode the one-hot matrix into a relatively shorter feature vector. Utilizing sequential neural networks known as Gated Recurrent Units (GRU) and multi-head Attention mechanism, we take advantage of the combination of sequential model and attention model to capture the deep contextual information of the input sequence.
The ChIP-seq feature extractor takes the sequence's nearest corresponding ChIP-seq(Roadmap Epigenomics, et al., 2015) information (including the peak value, score, and signal value from 6 different histone modifications: H3K4me1, H3K4me3, H3K27ac, H3K9me3, H3K36me3, and H3K27me3) as input. The ChIP-seq information is encoded as high-dimensional feature vectors and fed into the network. We apply fully connected layers with dropout at this part of the network to prevent over-fitting during training. Finally, high-dimensional features containing both sequences and their surrounding histone H3 protein information are passed to the final SoftMax layer, producing the final prediction.



Supplementary Figure S2. The detailed identification model of our proposed RHSNet-chip framework. The input sequence would first go through two 1-D convolutional layers as the sequence feature extractor. Then it will go through a Gated Recurrent Unit (GRU) for capturing long-range information, and a multi-head attention layer for detecting interactions within the sequence. In parallel, the ChIP-seq information would go through a fully connected network. Finally, the sequence feature and the ChIP-seq feature would be merged and give out the final prediction via the SoftMax activation.

b. Parameter settings and implementation details

The baseline CNN has two plain convolution layers connected with the final SoftMax layer. The first 1D convolution layer is designed with a filter of length16 and a kernel size of 30 with ReLU activation. The proposed RHSNet connects a Gated Recurrent Units (GRU) and a Multi-head attention layer after the 2-layer feature extractor. The number of neurons assigned in GRU is designed as 16. We use four multi-heads with the head size as 4 in the multi-head Attention mechanism.

Stochastic gradient descent with the momentum parameter as 0.9 and dynamically updated weight decay is used to optimize the learning process. The batch size is selected to be 64. The initial learning rate is set to be $1 * e^{-3}$. The dropout ratio is set to be 0.1 after the first and the second convolution layer to prevent over-fitting.

c. ChIP-seq feature extraction and importance score board

We quantify the distribution of three Chromatin Immunoprecipitation Sequencing (ChIP-seq) features (Score, Signal Value, Peak Value) extracted from the peaks of signal enrichment based on six different kinds of histone modifications(Dekker, et al., 2017) (H3K4me1(Yamada, et al., 2013), H3K4me3, H3K27ac(Chen, et al., 2020), H3K9me3, H3K36me3, H3K27me3), on different cell lines. Specifically, for paternal recombination map, we select the histone modifications from homo sapiens testis tissue (male adults, see Supplementary Fig. S8A). For maternal recombination map, we select the histone modifications from homo sapiens ovary tissue (female adults, see Supplementary Fig. S8B). Score is an integer value ranging from 0 to 1000, representing the significant score of each peak. Signal Value measures the average enrichment for the related peak region. Peak Value is the point-source called for this peak. It is the 0-based offset from the chromosome starting point and is set to -1 if no point-source is called. We take the log mean feature values of each histone modification and choose the nearest conservative peaks of each hotspot/coldspot clip.

Within each type of histone modification, p-values are obtained using the two-tailed Student's t-test. The calculated t-statistics of H3K4m3-signalValue ($8.03 * 10^{-7}$), H3K4m3-peakValue ($1.16 * 10^{-9}$), and H3K36me3-peakValue ($3.91 * 10^{-2}$) show the significant statistical difference between hotspots and coldspots within 18 features.

To further quantitively perceive the difference between the above 18 features and provide a more intuitive impression of their significance, we define the importance scoreboard not only as an index for measuring the usefulness of each feature, but also as an important index for measuring the quality and statistical significance of the feature's contribution to the improvement of the final prediction. The score is calculated by backpropagating the activation through the entire network to the ChIP-seq feature extraction branch of the RHSNet deep learning model. The greater the contribution score, the higher likelihood that this feature, along with its histone modification, plays a critical role in the prediction process.

d. Motif embedding and outlier detection

To intuitively visualize the discovered motif, we utilize DNABERT(Ji, et al., 2021), which extracts the short- to long-term patterns of each enriched motif into a fixed-size embedding vector. The embedded vector is further fitted into t-distributed Stochastic Neighbor Embedding(van der Maaten and Hinton, 2008) (t-SNE) to visualize the 2-dimensional embedding vector and investigate their divergence across different sexes, populations, and species. We use heatmaps of 2-mers to illustrate the physical meaning of each motif embedding cluster. Practically, we calculate the frequency of the 16 possible 2-mers within the [A, T, C, G] alphabets that appear in each population, sex and species, and visualize them with the saliency heatmap.

The outliers within each cluster are defined by the Local Outlier Factor (LOF)(Breunig, et al., 2000). This algorithm is an unsupervised anomaly detection method that computes the local density deviation of a given data point with respect to its neighbors. It considers as outliers the samples that have a substantially lower density than their neighbors. During the outlier

detection, we compute the locality based on the given k-nearest neighbors, whose distance is used to estimate the local density. By comparing the local density of a sample to the local densities of its 50 neighbors, we can identify samples that have a substantially lower density than their neighbors. These motifs are considered outlier motifs.
Regarding calculating the divergence of the embedded vectors distributing within 2-D space, we calculate the average distance of each embedded vector of selected species/sex/population to the central point of each embedding space. For example, the vectors of maternal crossover sequences have an average distance of 0.0447 ($\pm 1.9 * 10^{-2}$), which is much larger and has grater dispersion than that of paternal crossovers: $0.0355 \pm 1.6 * 10^{-2}$.

e. Evaluation Criteria

We use multiple evaluation methods to quantify the prediction results generated by different prediction methods according to the number of TP (True Positives), FP (False Positives), FN (False Negatives), and TN (True Negatives) samples.

Accuracy (ACC) can judge the performance of our model, but there is a serious flaw: in the case of imbalanced positive and negative samples, the category with a large proportion will often become the most important factor affecting accuracy. Therefore, sometimes, it might not reflect the overall prediction performance of the model. Accuracy is defined as follows:

$$Acc = \frac{TP + TN}{TP + FP + FN + TN} \tag{1}$$

F1 score (F1-Score) is a weighted average of recall and precision, and is defined as follows:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

$$F1\ score = \frac{2*Precision*Recall}{Precision+Recall} \tag{4}$$

Matthews correlation coefficient (MCC) is an index used in machine learning to measure the binary classification performance of the predictor. It is generally considered to be a relatively balanced evaluation metric, and it can be applied even when the number of positive and negative classes is extremely imbalanced. MCC is essentially a coefficient describing the correlation between the actual classification and the predicted classification. Its value range is [-1,1]. A value of 1 indicates a perfect prediction of the subject, and a value of 0 indicates that the predicted result is not as good as the random predicted result. -1 means that the predicted classification is exactly the opposite of the actual classification. MCC is defined as:

$$MCC = \frac{TP * TN - TP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}} \tag{5}$$

ROC curve and the corresponding AUROC score are other evaluation indexes. The larger the area under the curve (AUC), or the curve closer to the upper left corner (true positive rate=1, false-positive rate=0), the better the model's prediction in the task.

f. Imbalanced Testing

The recombination hotspots prediction is a problem that the number of positive samples (hotspots) is much less than that of negative samples (coldspots), which makes it difficult for the predictor to achieve high sensitivity. In average maps of the Icelandic2019(Halldorsson, et al., 2019) dataset, we found 130,172 hot spots and 755,041 cold spots with extreme diverse recombination rate from1.7e10−17cm/Mb to 56,242.73cm/Mb. Those sequences with lengths longer than 1kb were discarded during training, and the remaining hotspot samples are 17 times less than the cold spots samples (see **Supplementary Table S1**).

As shown in **Supplementary Figure S4,** after the balanced training, we test our model throughout all the sequence samples across the entire genome from the paternal and maternal maps at the Icelandic 2019(Halldorsson, et al., 2019) dataset. The area under the receiver operating characteristic (AUROC) was calculated to evaluate the algorithm's performance under an imbalanced prediction task. The RHSNet approach achieves the best AUCROC score of 0.689.

g. Recombination Rate Comparison

The illustration of the recombination rate distribution over the detected paternal and maternal motifs (**Supplementary Figure S14**) shows that the maternal recombination rates are relatively higher than that in paternal crossovers within each rate interval.

h. Hit@20,50,100 Evaluation

Similar to the Recommendation System (RS) ranking evaluation index, the motif detection method proposed in RHSNet that recommends prediction motifs could be evaluated similarly. First, we calculate the enrichment factor for each detected motif through the contribution score of that slot over the entire input sequence. In this way, each detected motif will get an enrichment factor score. Furthermore, we sort these scores in descending order so that the motifs with the highest enrichment factor would be ranked in the front.

According to the above ranking function, we can count whether the PRDM9-A/C allele exists for each detected motif is in the top 20 of the sequence, and if so, we add one count to Hit@20. In the end, the top 20 number/total is Hit@20. Similarly, Hit@50 and Hit@100 are the top 50/100 detected PRDM9-A/C alleles over the total number of motifs. Furthermore, the value of Hit@20 may exceed 20 because the detected motif is usually 21bp long, and it might contain more than one 12-bp motifs in one sequence.

Intuitively speaking, the key factor of predicting an input sequence as the recombination hotspot will give much credit to the PRDM9-A/C allele. As showed in **Supplementary Table S6,** the PRDM9-A allele is ranked pretty high in Hit@10/20/100 evaluation, demonstrating that RHSNet could precisely identify the key factor of the recombination hotspot determinant. Also, the number of the RHSNet-detected PRDM9-A allele is approximately nine times larger than that of the PRDM9-C allele. Such a result is consistent with previous studies that the PRDM9-A motif plays a role in approximately 40% of hotspots and is proposed to be involved in initiation specification or other aspects of recombination activity.

i. Maximum Mean Discrepancy (MMD) calculation

As a kernel-based distance calculation metric, Maximum Mean Discrepancy (MMD) can accurately quantify the similarity and the distance between two vector distributions. When measuring the difference between two embedding distributions, because we need to measure

the non-parametric distribution distance between the source and target data, namely, P and Q, we use MMD. The calculation is done by the following equation:

$$MMD(P,Q) = \left\| \frac{1}{n}\sum_{i=1}^{n} \emptyset(x_i) - \frac{1}{m}\sum_{j=1}^{m} \emptyset(y_j) \right\|_H^2 \quad (6)$$

For example, when comparing the embedding distance between paternal motif vectors (P) and maternal motif vectors (Q), we map P and Q (Original $\boldsymbol{X}$ space) from the original embedding space to another space $\mathbf{H}$ ($\mathbf{Hilbert\ space}$) through the function $\emptyset: \boldsymbol{X} \rightarrow \boldsymbol{H}$. Then, we can calculate the mean difference between P and Q in the H space feature dimension. When utilizing MMD as the evaluation metric for calculating the distance between two distributions within the embedding space, we can determine whether the two distributions are similar. Quantitatively, the $\boldsymbol{MMD}(\boldsymbol{Human}, \boldsymbol{Mouse})$ is 0.0221 ($p = 0.9774$), which is much smaller than $\boldsymbol{MMD}(\boldsymbol{Human}, \boldsymbol{Yeast})$=0.3729 ($p = 8.5 * 10^{-3}$).

j. PRDM9-A/C allele identification

All the PRDM9-A/C alleles are identified through rigorous multiple sequence alignment from both ground-truth hotspots and detected binding motifs. The major PRDM9-A allele: CCNCCNTNNCCNC and its reverse: GGNGGNANNGGNG, as well as PRDM9-C allele: CCGCNGTNNNCGT and its reverse: GGCGNCANNNGCA, are selected as the reference sequences. Each discovered motif would be conducted a pairwise sequence alignment with each allele using a dynamic programming algorithm. The selected motif would only be considered containing the major allele A when having the minimum alignment score of 8, which is chosen as the certainty of the 8 certain bases: CC-CC-T--CC-C and GG-GG-A--GG-G. The identification rule is also applied for the identification of the rare allele C.

k. Explanation for back-propagation based reference sequence selection

Base on the Homo sapiens (human) genome assembly GRCh38 (hg38) from Genome Reference Consortium . We have calculated the frequencies of A, C, G, T, N nucleotides(where N represents any nucleotide) for 22 autosomes as:
'A': 812507870, 'C': 563229147,'G': 565528321, 'T': 815065703, 'N': 118670481,'all': 2875001522.
Therefore, we have calculated the corresponding frequency:
A:0.2826, C:0.1959,G:0.1967, T:0.2835,N:0.0412.
Since we don't indicate N nucleotides in our filter-based back propagation, we approximate the frequencies for A,C,G and T nucleotides of 0.3, 0.2, 0.2, and 0.3.

l. Explanation for choosing Wn from [0.1,0.2,0.4]

In this manuscript, we have all the data represented as 1000bp sequence. Therefore, the sample signal can be considered a 1 second signal with 1000hz sampling frequency ($f_s = 1000$).
For example, when the users are looking for 10bp-length motif, we want to filter out the insignificant peaks with length less than 10bp. Therefore, we can select the cut-off frequency: fc= $\frac{1}{\frac{10}{1000}} = 100$Hz. So, the user can select $W_n = 2 * f_c \,/\, f_s$= 2*(100)/1000 = 0.2 to capture interesting motifs with approximately 10bp length. Similarly, the user can set W_n=0.1 to capture motifs close to 50bp length, and W_n=0.4 for motifs close to 5bp length.

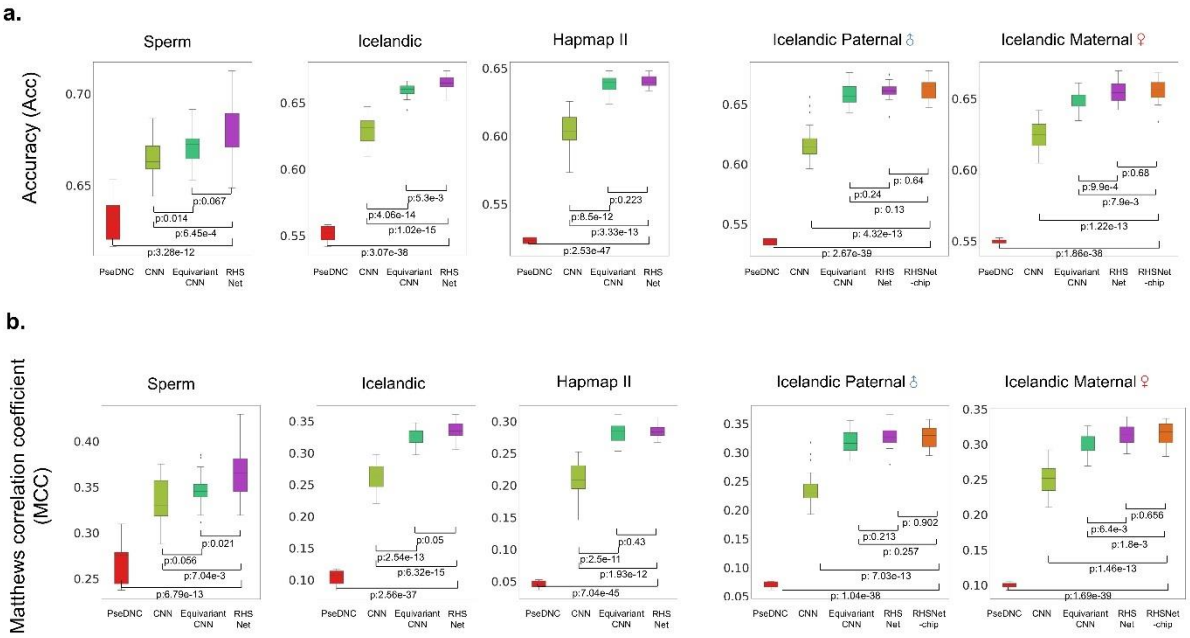
Section 3. Supplementary identification results

a. Detailed overall classification performance statistics on different datasets

Data	Method	F1 Score %	Accuracy %	MCC %
Icelandic(Halldorsson, et al., 2019)	RHSNet	**63.61 (±1.94)**	**66.47 (±0.62)**	**33.45 (±1.41)**
	Equivariant CNN(Brown and Lunter, 2019)	62.66 (±2.09)	65.93 (±0.51)	32.54 (±1.35)
	CNN	60.30 (±2.51)	62.96 (±0.99)	26.29 (±2.09)
	PseDNC(Chen, et al., 2013)	53.79(±0.58)	55.13(±0.62)	10.3(±1.26)
Paternal Map ♂ Icelandic(Halldorsson, et al., 2019)	RHSNet-chip	**64.66(±1.37)**	**66.20 (±0.58)**	**32.31 (±1.06)**
	RHSNet	63.17 (±1.58)	66.11 (±0.72)	32.70 (±1.84)
	Equivariant CNN(Brown and Lunter, 2019)	63.08 (±1.58)	65.79 (±0.92)	31.99 (±2.01)
	CNN	60.31 (±2.62)	61.80 (±1.57)	23.80 (±3.18)
	PseDNC(Chen, et al., 2013)	52.48(±0.40)	53.53(±0.30)	7.08(±0.58)
Maternal Map ♀ Icelandic(Halldorsson, et al., 2019)	RHSNet-chip	**64.21 (±1.94)**	**66.09 (±0.85)**	**31.89 (±1.77)**
	RHSNet	63.17 (±2.11)	65.47 (±0.75)	31.34 (±1.48)
	Equivariant CNN(Brown and Lunter, 2019)	61.99 (±2.46)	64.74 (±0.66)	29.98 (±1.42)
	CNN	60.55 (±1.80)	62.35 (±0.98)	24.95 (±2.11)
	PseDNC(Chen, et al., 2013)	53.88(±0.46)	54.96 (±0.22)	9.95(±0.42)
Hapmap II (Frazer, et al., 2007)	RHSNet	**61.49 (±1.72)**	**64.02 (±0.44)**	**29.39 (±1.08)**
	Equivariant CNN(Brown and Lunter, 2019)	60.07 (±2.09)	63.81 (±0.64)	28.25 (±1.54)

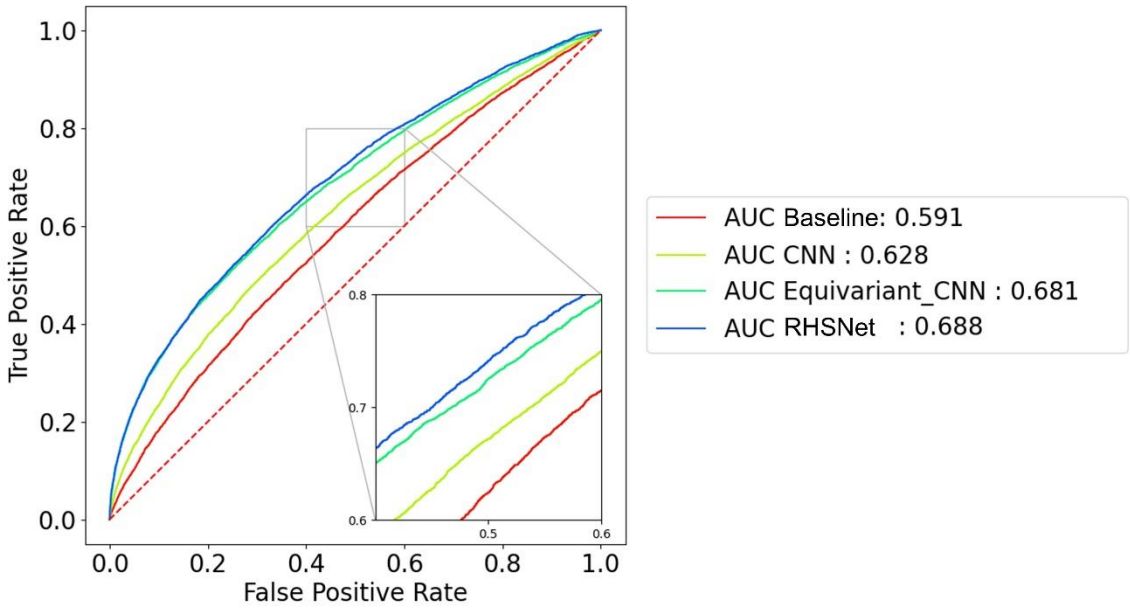
	CNN	57.89 (±2.68)	60.28 (±1.37)	20.84 (±2.88)
	PseDNC(Chen, et al., 2013)	52.72(±0.48)	52.27 (±0.28)	4.56 (±0.58)
Sperm(Frazer, et al., 2007)	RHSNet	**68.92 (±3.51)**	**67.95 (±1.55)**	**36.55 (±3.0)**
	Equivariant CNN(Brown and Lunter, 2019)	67.87 (±2.41)	67.15 (±0.93)	34.62 (±1.71)
	CNN	68.33 (±2.43)	66.32 (±1.12)	33.21 (±2.43)
	PseDNC(Chen, et al., 2013)	66.08(±1.03)	63.25 (±1.32)	26.91 (±2.60)
Mouse(Lange, et al., 2016)	RHSNet	**87.31 (±0.78)**	**86.32 (±0.72)**	**73.59 (±1.54)**
	Equivariant CNN(Brown and Lunter, 2019)	73.93 (±4.48)	75.41 (±3.32)	51.32 (±6.52)
	CNN	76.77 (±1.71)	76.25 (±0.94)	52.07 (±1.94)
	PseDNC(Chen, et al., 2013)	64.04 (±1.01)	63.89 (±0.82)	27.85 (±1.669)

Supplementary Table S4. Detailed classification performance of the proposed RHSNet-chip and RHSNet, compared with the baseline CNN model , PseDNC(Chen, et al., 2013) and Equivariant CNN(Brown and Lunter, 2019). RHSNet shows outstanding prediction performance on multiple benchmark datasets: Icelandic(Halldorsson, et al., 2019), HapMap II(Frazer, et al., 2007), Sperm(Frazer, et al., 2007), and Mouse(Lange, et al., 2016), which are across different studies, sexes, and species.



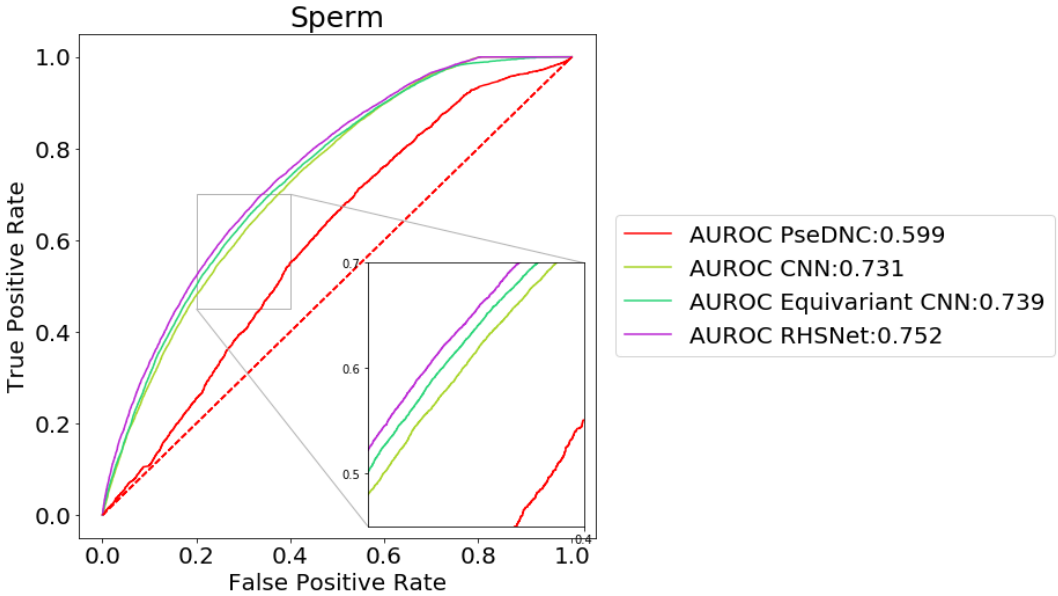
Supplementary Figure S3. Detailed data statistical performance of RHSNet across different studies and sexes. **(a)** Boxplot of accuracy (Acc) distribution that distinguishes RHSNet from Baseline CNN and Equivariant CNN(Brown and Lunter, 2019) in multiple trials of 5-fold cross-validation experiments. **(b)** Boxplot of Matthews correlation coefficient (MCC) distribution that distinguishes RHSNet from Baseline CNN and Equivariant CNN in multiple trials of 5-fold cross-validation experiments.

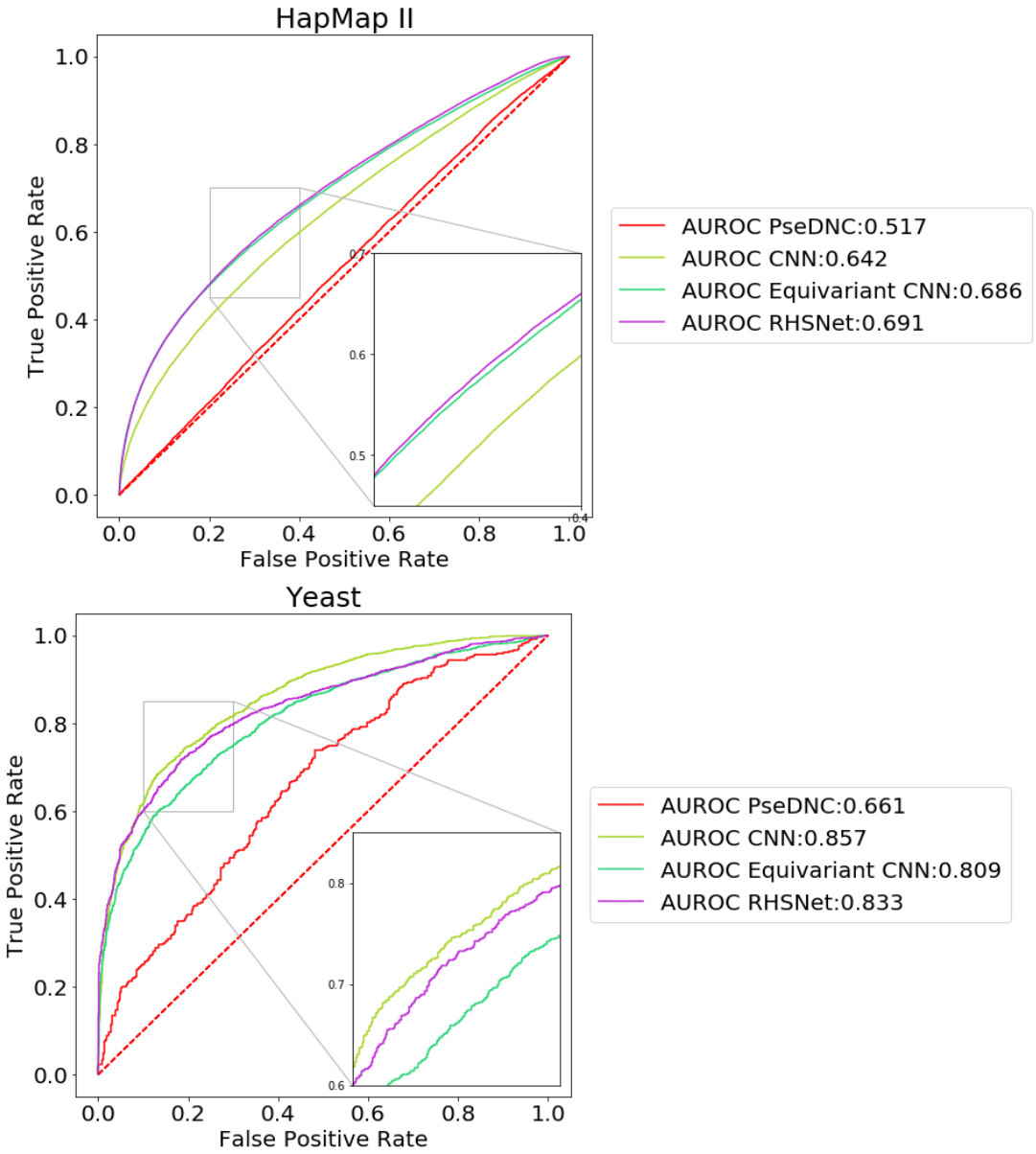
b. Imbalanced testing on Icelandic dataset



Supplementary Figure S4. In the Icelandic 2019 dataset, we show the ROC curve and the AUROC score of 4 prediction methods during imbalanced testing. RHSNet achieves the best prediction performance with an AUROC score of 0.688.

c. ROC comparison against Equivariant CNN(Brown and Lunter, 2019) PseDNC(Chen, et al., 2013) and on multiple datasets across different studies and species





Supplementary Figure S5. In all the datasets, we show the ROC curve and the AUROC score of the proposed RHSNet and the existing state-of-the-art method Equivariant CNN and SVM based method PseDNC. RHSNet has relatively higher prediction performance in each dataset.

d. Detailed comparison against PseDNC with different parameters

Data	Method	lambda	weight	Overall Accuracy%
Icelandic(Halldorsson, et al., 2019)	RHSNet	--	--	**66.47 (±0.62)**
	Equivariant CNN	--	--	65.93 (±0.51)
	CNN	--	--	60.30 (±2.51)
	PseDNC + SVM	3	0.05	55.13(±0.62)
	PseDNC + SVM	5	0.05	55.07 (±0.68)
	PseDNC + SVM	10	0.05	55.33(±0.76)
Paternal Map ♂ Icelandic(Halldorsson, et al., 2019)	RHSNet	--	--	**66.23 (±0.87)**
	Equivariant CNN	--	--	66.11 (±0.72)
	CNN	--	--	65.79 (±0.92)
	PseDNC + SVM	3	0.05	53.53(±0.30)
	PseDNC + SVM	5	0.05	53.16 (±0.38)
	PseDNC + SVM	10	0.05	54.63(±0.46)
Maternal Map ♀ Icelandic(Halldorsson, et al., 2019)	RHSNet	--	--	**65.57 (±0.78)**
	Equivariant CNN	--	--	65.47 (±0.75)
	CNN	--	--	64.74 (±0.66)
	PseDNC + SVM	3	0.05	54.96(±0.22)
	PseDNC + SVM	5	0.05	55.07 (±0.34)
	PseDNC + SVM	10	0.05	55.14(±0.36)
HapMap II(Frazer, et al., 2007)	RHSNet	--	--	**64.02 (±0.44)**
	Equivariant CNN	--	--	63.81 (±0.64)
	CNN	--	--	60.28 (±1.37)
	PseDNC + SVM	3	0.05	52.27(±0.28)
	PseDNC + SVM	5	0.05	52.53(±0.36)
	PseDNC + SVM	10	0.05	52.87(±0.42)

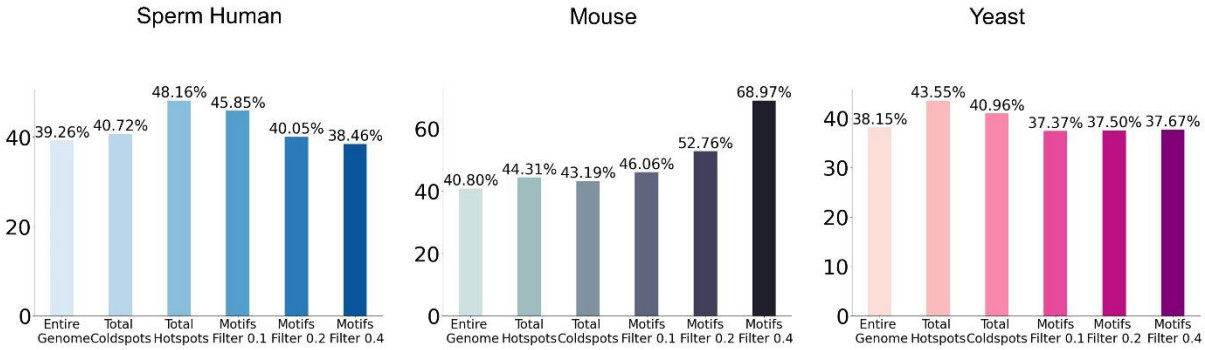
	RHSNet	--	--	**86.32 (±0.72)**
	Equivariant CNN	--	--	75.41 (±3.32)
Sperm(Frazer, et al.,	CNN	--	--	76.25 (±0.94)
2007)	PseDNC + SVM	3	0.05	63.25(±1.32)
	PseDNC + SVM	5	0.05	63.51 (±0.67)
	PseDNC + SVM	10	0.05	63.22(±0.72)
	RHSNet	--	--	**73.59 (±1.54)**
	Equivariant CNN	--	--	51.32 (±6.52)
Mouse(Lange, et al.,	CNN	--	--	52.07 (±1.94)
2016)	PseDNC + SVM	3	0.05	63.89(±0.82)
	PseDNC + SVM	5	0.05	64.29 (±0.53)
	PseDNC + SVM	10	0.05	63.82 (±0.61)

Supplementary Table S5. Statistical results on the multiple dataset, comparing the proposed RHSNet with multiple experimental settings of PseDNC(Chen, et al., 2013) + Support Vector Machine (SVM(Cherkassky, 1997)) classifier. Multiple experiments with different sets of parameter lambda for pseudo-feature extraction are conducted. The deep learning-based methods show a significant edge over the SVM-based classifier.

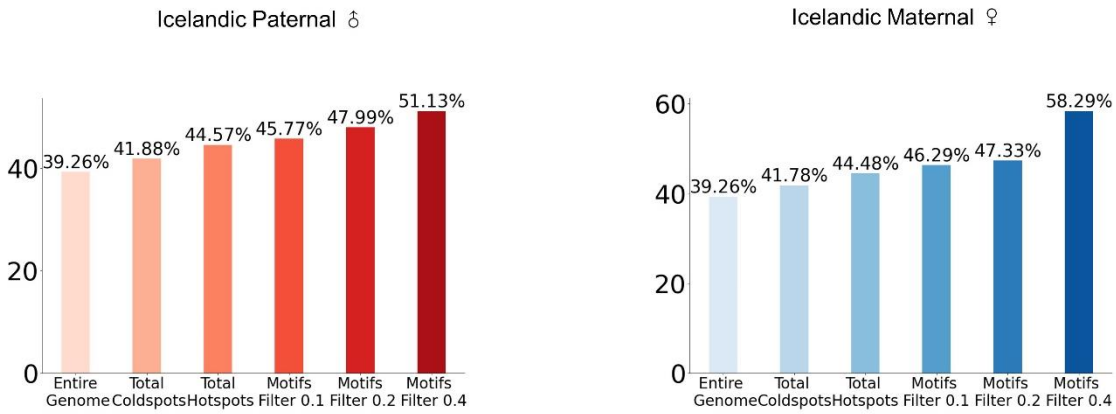
Section 4. Supplementary quantification results

a. RHSNet's sensitivity to sex differences

The recombination differs between males and females of the same species, including humans and mice(Brick, et al., 2018; Halldorsson, et al., 2016). The females show a higher overall recombination rate and more complex crossovers than the males(Bherer, et al., 2017; Halldorsson, et al., 2019; Kong, et al., 2010), despite the elusive mechanisms behind the differences(Halldorsson, et al., 2019). Although most of the recombination hotspots (>88%) are shared between males and females, the strongest hotspots tend to be sex-specific(Brick, et al., 2018). In our study, the most significant motifs detected from the Icelandic males are conserved PRMD9 binding motifs (see **Supplementary Figure S10B**), which is consistent with the finding that the PRDM9-binding sites are frequently methylated at male-biased hotspots(Brick, et al., 2018), although the motifs can be different in different species. On the other hand, the contributing determinant motifs of the Icelandic females are less conservative (enrichment factor: 8.76 ± 1.31) than the paternal ones (enrichment factor: 15.31 ± 2.01). Also, in the female, the determinant motifs are much more diverse than those in males, which may result from the distinct methylation mechanism (unlike males, DNA methylation increases in the region ± 75bp adjacent to the PRDM9-binding sites), more complex crossover, and higher evolution speed(Baudat, et al., 2013; Brick, et al., 2018; Halldorsson, et al., 2019) (see **Figure 3C, Supplementary Figure S11**, female rate: $52.48 \pm 67.29\ cM/mb$; male rate: $39.53 \pm 40.63\ cM/mb$). Although the data themselves may not be sufficient to illustrate the mechanism behind the sex biases in recombination, the identified and quantified determinants suggest that, in females, diverse factors, including PRDM9 and SPO11 (the rank 2 motifs), control the female-biased hotspots, while, in males, the hotspots tend to be PRDM9-directed.



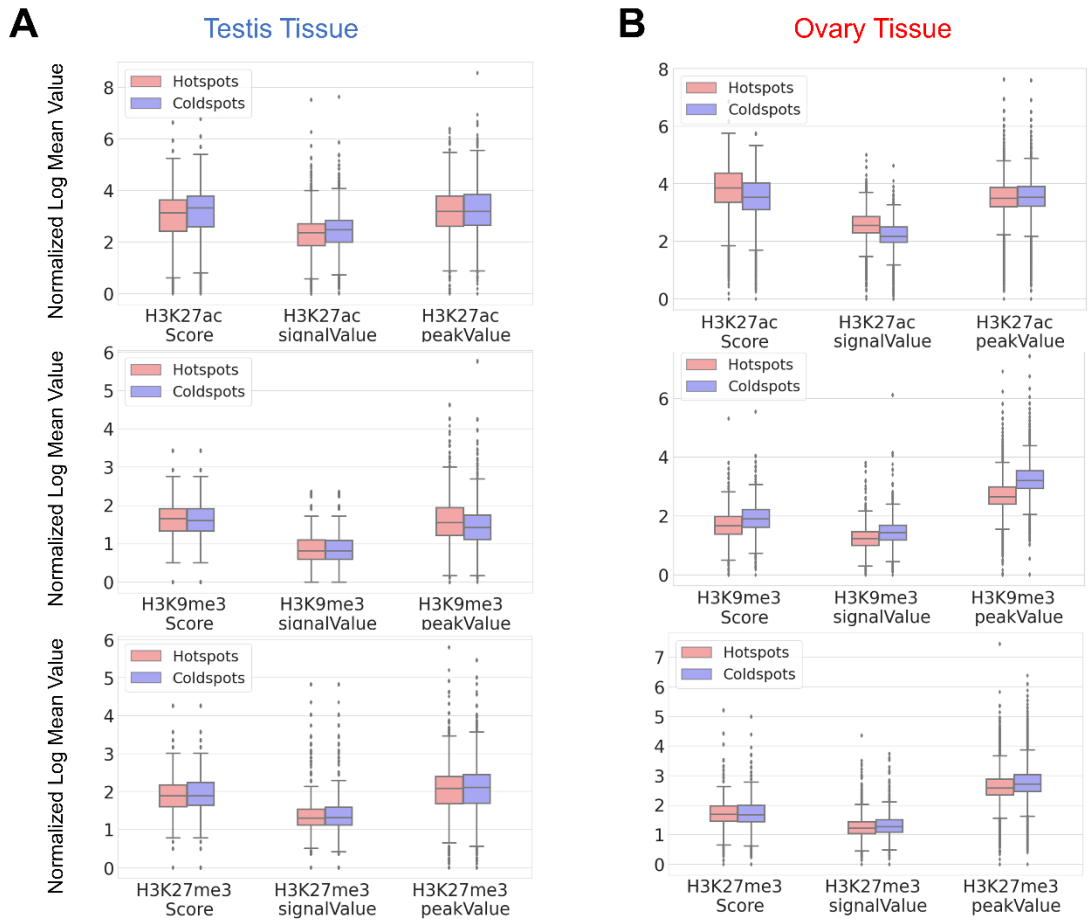
Supplementary Figure S6. Statistical comparison of GC content across different species.



Supplementary Figure S7. Statistical comparison of GC content across different sexes. The result is calculated from Icelandic(Halldorsson, et al., 2019) Human genetic maps.

Interestingly, the GC content of the coldspots (40.72%~41.88%), where the recombination rates are the lowest in the genome, is not lower than that of the entire genome, although it is lower than that of the hotspots. For some datasets, such as HapMap II, the result is expected because the coldspot set was constructed to match the GC content of the hotspot one. However, for the other datasets, the coldest coldspots show similar GC content, which suggests that GC content itself may not be the causation of hotspots. Instead, the higher GC content in hotspots may be the consequence of the determinant motifs, which are GC-rich, such as the PRDM9 binding motif. In the HapMap II dataset and the Icelandic dataset, where the resolution is high enough (up to 642bp), the determinant motifs identified by our model have much higher GC content than that of the overall hotspot regions. Furthermore, as we increase the filter factor, which forces our method to output shorter motifs with higher enrichment factors, the GC content increases further in these motifs (up to 65.3%). The separated paternal map and maternal map show a similar trend as the average map in the Icelandic dataset (see Supplementary Figure S7), although the signal is reduced because fewer people are included in each map. Intuitively, the results are consistent with the previous discoveries, as the PRDM9 motif, which is GC-rich, is the most popular motif in the hotspot region.

b. The histone modification feature distribution comparison on ovary tissue between hotspots and coldspots on maternal recombination map



Supplementary Figure S8. The normalized log mean value comparing the different distributions between hotspots and coldspots from the ChIP-seq features (signal Value, peak Value, Score) extracted from the ovary and testis tissues of Homo sapiens female and male adults. Here, we show the comparison over three different histone modifications: H3K27ac, H3K27me3, and H3K9me3.

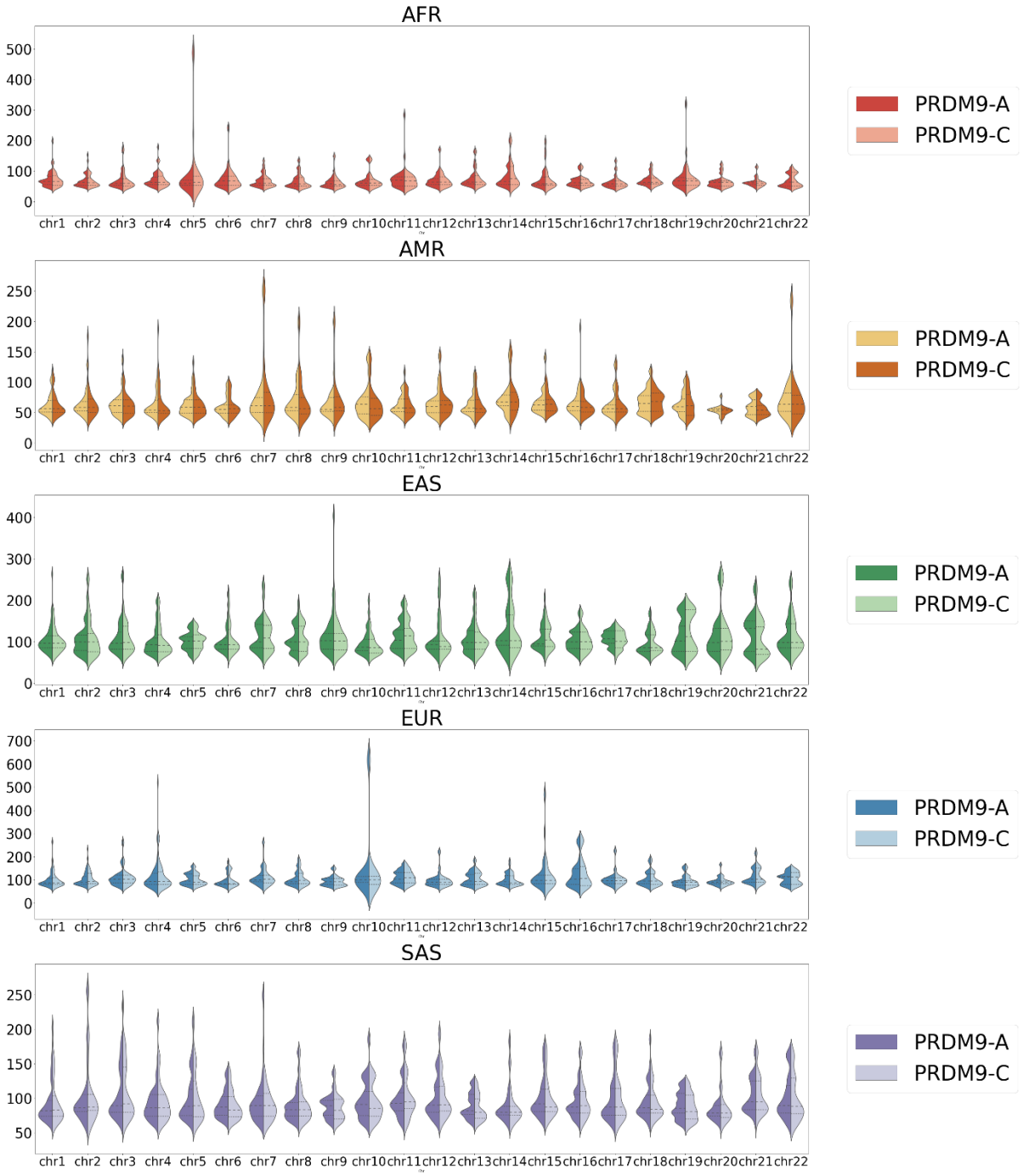
c. PRDM9 alleles identification across different datasets

Study	PRDM9 allele	hit@20	hit@50	hit@100
Icelandic	PRDM9-A	9	35	83
	PRDM9-C	0	6	10
HapMap II	PRDM9-A	5	16	49
	PRDM9-C	3	7	18
Sperm	PRDM9-A	19	35	69
	PRDM9-C	1	5	6

Supplementary Table S6. The hit@20/50/100 results of the RHSNet's identification across different studies.

Population	PRDM9 allele	hit@20	hit@50	hit@100
African	PRDM9-A	20	58	124
	PRDM9-C	0	0	8
American	PRDM9-A	19	25	44
	PRDM9-C	0	4	12
East Asian	PRDM9-A	0	2	7
	PRDM9-C	0	2	3
European	PRDM9-A	14	24	48
	PRDM9-C	0	3	4
South Asian	PRDM9-A	30	71	90
	PRDM9-C	0	2	2

Supplementary Table S7. The hit@20/50/100 results of the RHSNet's identification across different populations.

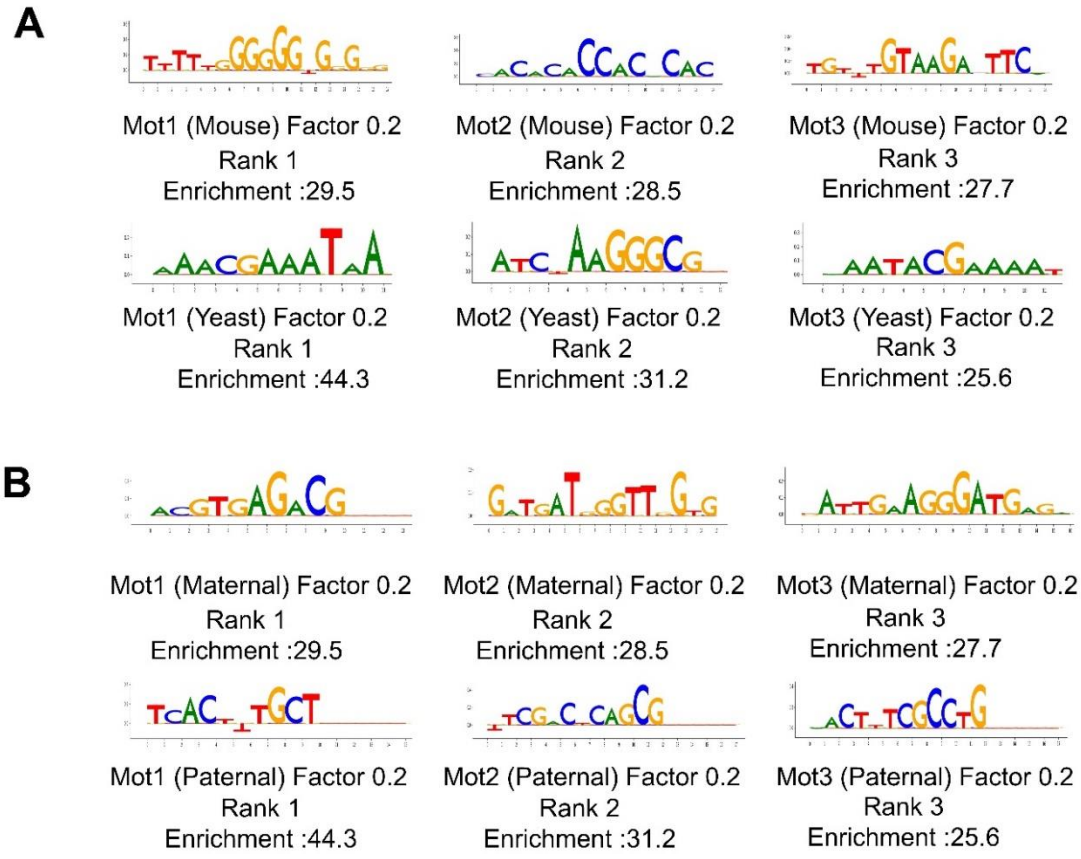


Supplementary Figure S9. Across five different populations from the 1000 Genome Project(Genomes Project, et al., 2015), we draw the recombination rate distribution of PRDM9-A/C alleles over 22 autosomes on the ground truth genetic maps.

e. The most important motifs detected in different species and sexes

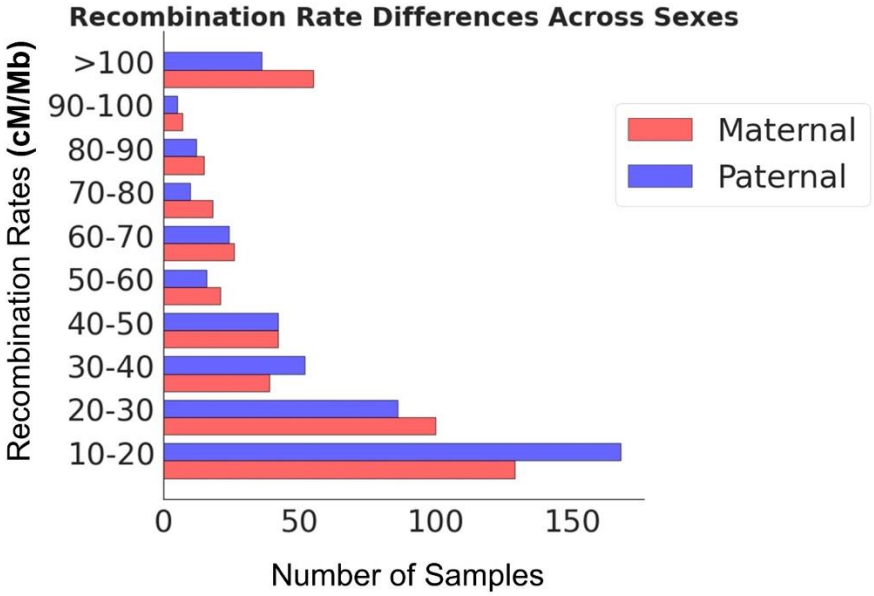
In the Icelandic population, within the central region of the embeddings, the motifs in both the maternal population and the paternal population are PRDM9-related motifs (see **Figure 6E**). However, regarding the outliner motifs, they are very different. For males, the motifs tend to be short and strong, while the motifs are long and diverse in the females.

By calculating the difference between two embedding distributions, we quantify the difference between two species via Maximum Mean Discrepancy(Tolstikhin, et al., 2016) (MMD) (see **Supplementary Section 2: Methods**). For example, using our method, the evolutionary distance between Human and Mice (0.0221, $p = 0.9774$) is much smaller than that between Human and Yeast (0.3729, $p = 8.5 * 10^{-3}$).



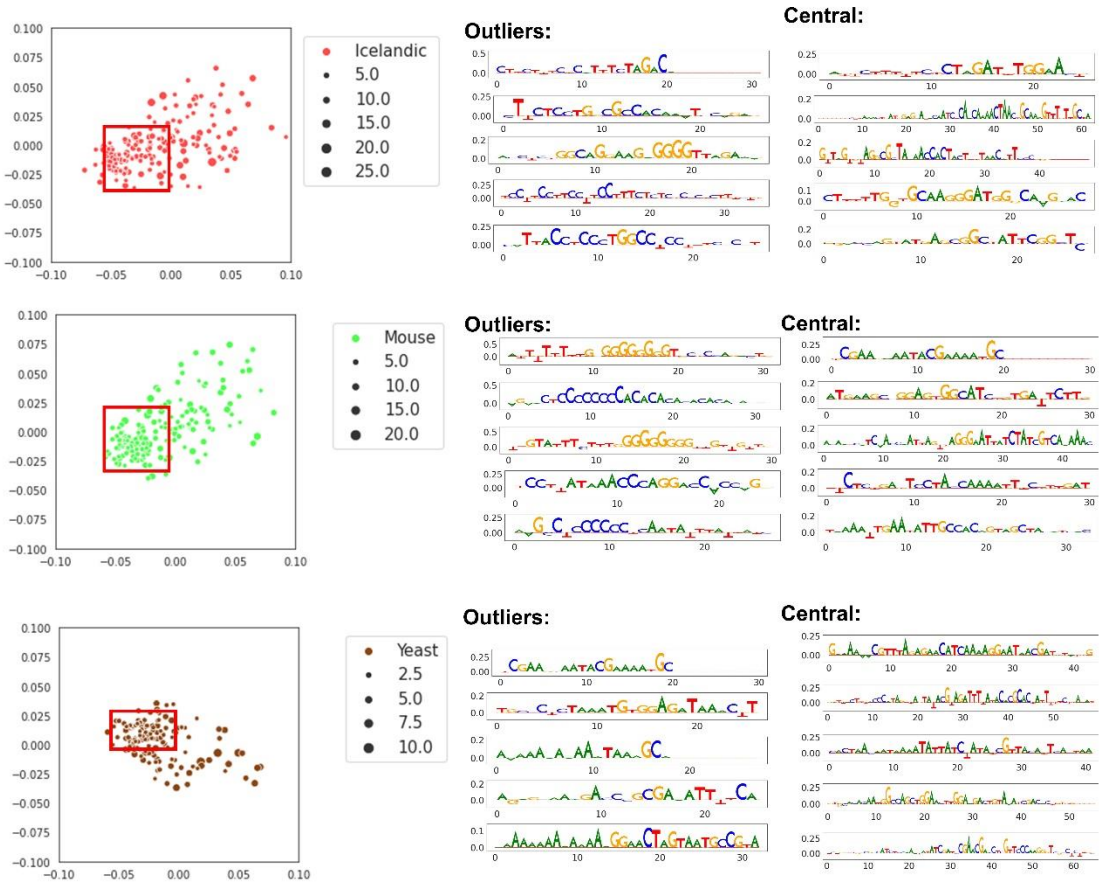
Supplementary Figure S10. (A) The top 3 detected motifs from the Mouse(Lange, et al., 2016) and Yeast(Mancera, et al., 2008) datasets.
(B) The top 3 detected motifs from the paternal and maternal genetic maps, respectively.

e. Recombination rate comparison between paternal and maternal motifs

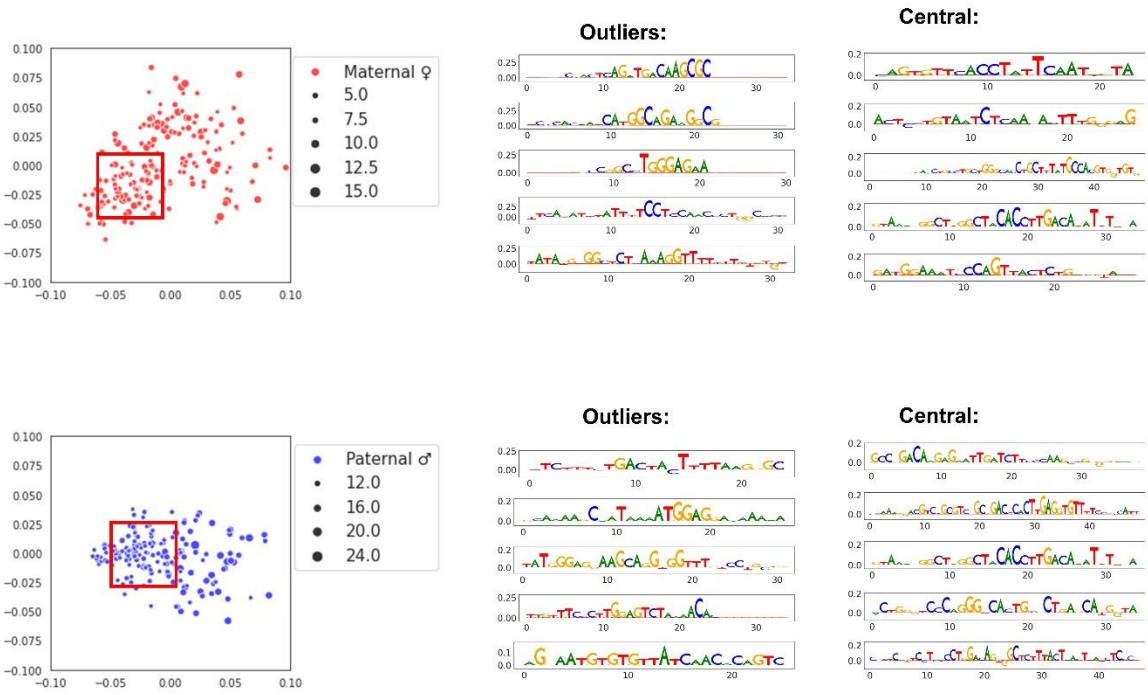


Supplementary Figure S11. Statistical comparison of the recombination rate between the detected paternal and maternal motifs.

f. Detailed motif embedding and clustering results

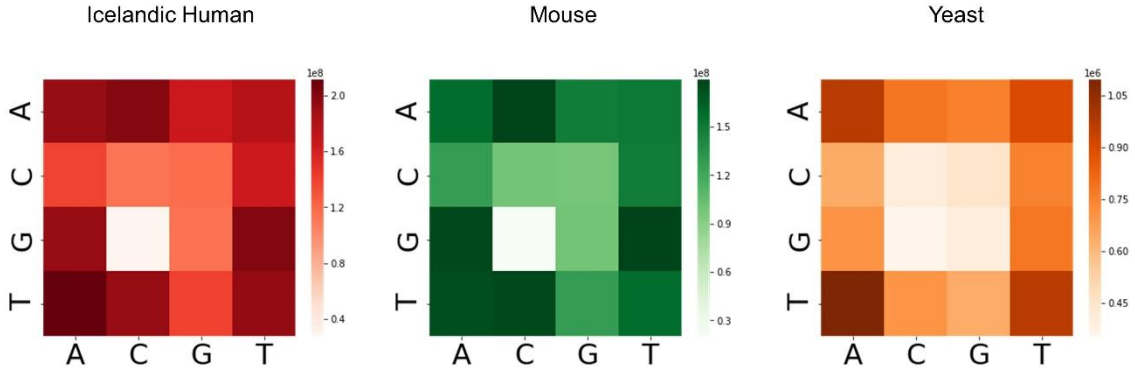


Supplementary Figure S12. Embedding vectors in 2D space across different species. The central region is bounded by a red bounding box, and the outliers are defined by comparing the local density of a sample to the local densities of its 50 neighbors. The top 5 ranked motifs are visualized for both central and outlier motifs.

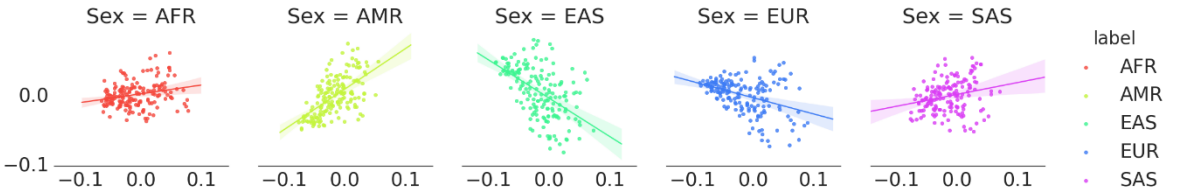


Supplementary Figure S13. Embedding vectors in 2D space across different sexes. The central region is bounded by a red bounding box, and the outliers are defined by comparing the local density of a sample to the local densities of its 50 neighbors. The top 5 ranked motifs are also visualized for both central and outlier motifs.

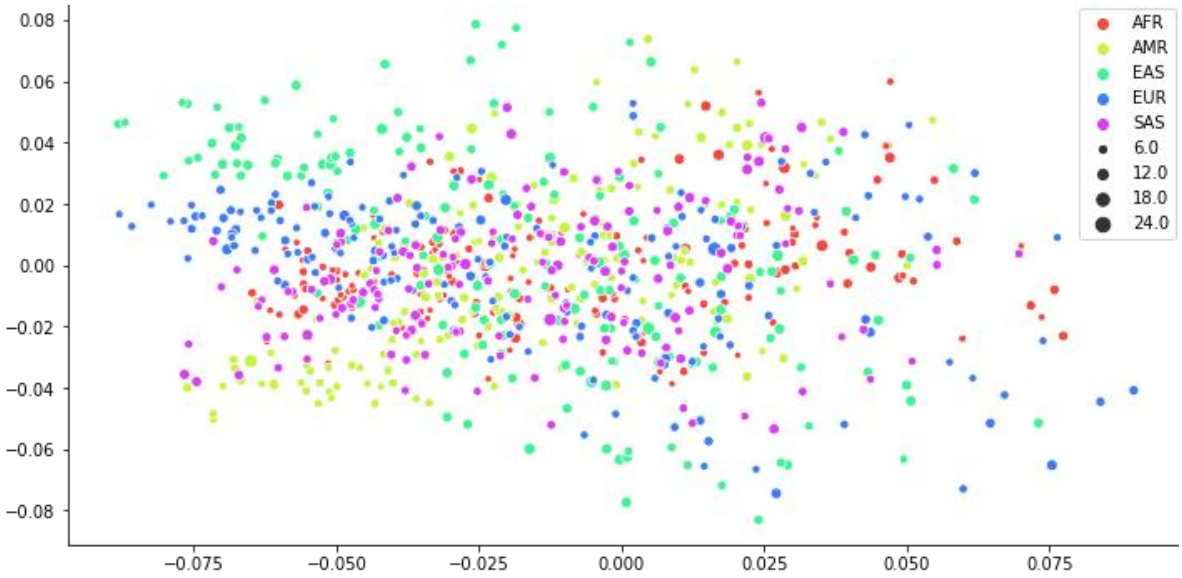
g. Motif visualization of the entire genome



Supplementary Figure S14. Heatmaps of 2-mer distributions over the entire genome. Each grid represents the 2-mer appearance frequency (AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, TT) across GRCh38 human reference genome, GRCm38/mm10 reference genome, and S288C Yeast reference genome.



Supplementary Figure S15. The embedding results with regression models for five different populations.



Supplementary Figure S16. The overall embedding results within the same 2D space for five different populations.

References

Baudat, F., Imai, Y. and de Massy, B. Meiotic recombination in mammals: localization and regulation. *Nat Rev Genet* 2013;14(11):794-806.

Bell, A.D.*, et al.* Insights into variation in meiosis from 31,228 human sperm genomes. *Nature* 2020;583(7815):259-+.

Bherer, C., Campbell, C.L. and Auton, A. Refined genetic maps reveal sexual dimorphism in human meiotic recombination at multiple scales. *Nat Commun* 2017;8:14994.

Breunig, M.M.*, et al.* LOF: Identifying density-based local outliers. *Sigmod Rec* 2000;29(2):93-104.

Brick, K.*, et al.* Extensive sex differences at the initiation of genetic recombination. *Nature* 2018;561(7723):338-+.

Brown, R.C. and Lunter, G. An equivariant Bayesian convolutional network predicts recombination hotspots and accurately resolves binding motifs. *Bioinformatics* 2019;35(13):2177-2184.

Chen, W.*, et al.* iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. *Nucleic Acids Res* 2013;41(6):e68.

Chen, Y.*, et al.* Refined spatial temporal epigenomic profiling reveals intrinsic connection between PRDM9-mediated H3K4me3 and the fate of double-stranded breaks. *Cell Res* 2020;30(3):256-268.

Cherkassky, V. The nature of statistical learning theory~. *IEEE Trans Neural Netw* 1997;8(6):1564.

Dekker, J.*, et al.* The 4D nucleome project. *Nature* 2017;549(7671):219-226.

Frazer, K.A.*, et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007;449(7164):851-U853.

Genomes Project, C.*, et al.* A global reference for human genetic variation. *Nature* 2015;526(7571):68-74.

Halldorsson, B.V.*, et al.* The rate of meiotic gene conversion varies by sex and age. *Nat Genet* 2016;48(11):1377-1384.

Halldorsson, B.V.*, et al.* Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science* 2019;363(6425):eaau1043.

Halldorsson, B.V.*, et al.* Characterizing mutagenic effects of recombination through a sequence-level genetic map (vol 363, eaaw8705, 2019). *Science* 2019;363(6430):939-939.

Ji, Y.*, et al.* DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics* 2021.

Kong, A.*, et al.* Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* 2010;467(7319):1099-1103.

Lange, J.*, et al.* The Landscape of Mouse Meiotic Double-Strand Break Formation, Processing, and Repair. *Cell* 2016;167(3):695-+.

Mancera, E.*, et al.* High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature* 2008;454(7203):479-485.

Roadmap Epigenomics, C.*, et al.* Integrative analysis of 111 reference human epigenomes. *Nature* 2015;518(7539):317-330.

Tolstikhin, I., Sriperumbudur, B.K. and Scholkopf, B. Minimax Estimation of Maximum Mean Discrepancy with Radial Kernels. *Advances in Neural Information Processing Systems 29 (Nips 2016)* 2016;29.
van der Maaten, L. and Hinton, G. Visualizing Data using t-SNE. *J Mach Learn Res* 2008;9:2579-2605.
Yamada, S., Ohta, K. and Yamada, T. Acetylated Histone H3K9 is associated with meiotic recombination hotspots, and plays a role in recombination redundantly with other factors including the H3K4 methylase Set1 in fission yeast. *Nucleic Acids Research* 2013;41(6):3504-3517.