Bioinformatics, 38(10), 2022, 2683–2691
https://doi.org/10.1093/bioinformatics/btac234
Advance Access Publication Date: 12 April 2022
Original Paper

Genome analysis

Deep learning identifies and quantifies recombination hotspot determinants

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

[1]Department of Computer Science and Engineering (CSE), The Chinese University of Hong Kong (CUHK), 999077, 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, Thuwal 23955-6900, Kingdom of Saudi Arabia, [4]The CUHK Shenzhen Research Institute, Shenzhen 518057, China and [5]Max Planck Institute for Molecular Genetics, Berlin 14195, Germany

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Tobias Marschall
Received on October 5, 2021; revised on March 8, 2022; editorial decision on March 31, 2022; accepted on April 8, 2022

Abstract

Motivation: Recombination is one of the essential genetic processes for sexually reproducing organisms, which can happen more frequently in some regions, called recombination hotspots. Although several factors, such as PRDM9 binding motifs, are known to be related to the hotspots, their contributions to the recombination hotspots have not been quantified, and other determinants are yet to be elucidated. Here, we propose a computational method, RHSNet, based on deep learning and signal processing, to identify and quantify the hotspot determinants in a purely data-driven manner, utilizing datasets from various studies, populations, sexes and species.

Results: RHSNet can significantly outperform other sequence-based methods on multiple datasets across different species, sexes and studies. In addition to being able to identify hotspot regions and the well-known determinants accurately, more importantly, RHSNet can quantify the determinants that contribute significantly to the recombination hotspot formation in the relation between PRDM9 binding motif, histone modification and GC content. Further cross-sex, cross-population and cross-species studies suggest that the proposed method has the generalization power and potential to identify and quantify the evolutionary determinant motifs.

Availability and implementation: https://github.com/frankchen121212/RHSNet.

Contact: liyu@cse.cuhk.edu.hk or xin.gao@kaust.edu.sa

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Recombination is an essential and fundamental genetic process in meiosis, which introduces new combinations of alleles and generates haplotypic diversity in sexually reproducing organisms, driving evolution and biodiversity (Baudat *et al.*, 2013; Halldorsson *et al.*, 2019; Spence and Song, 2019). Although the molecular mechanism of this process has not been fully uncovered (Baudat *et al.*, 2013), it is believed that in many species, including humans and mice, the event begins with the binding of DNA by the histone methyltransferase PRDM9 (Baudat *et al.*, 2010). The double-strand break (DSB) machinery, including the meiotic topoisomerase-like protein SPO11 (Paiano *et al.*, 2020), is then recruited by an unknown mechanism, forming DSBs. Specialized pathways repair these breaks, with the majority leading to non-crossovers while the minority developing crossovers (COs) (Mancera *et al.*, 2008). Although PRDM9 binds ubiquitously throughout the genome, the distribution of COs is non-random, clustered in narrow regions (see Supplementary Fig. S1), called recombination hotspots (Baudat *et al.*, 2010, 2013). Despite the unclear reason for forming hotspots, the following factors are suggested to be related to the locations of hotspots (Halldorsson *et al.*, 2019; Hinch *et al.*, 2019). The DNA binding domain of PRDM9 influences sequence specificity and the formation of DSBs (Baudat *et al.*, 2010; Myers *et al.*, 2008, 2010; Parvanov *et al.*, 2010); histone modifications can influence the chromatinic local structure and thus affect CO formation (Jin *et al.*, 2021; Lange *et al.*, 2016; Spence and Song, 2019); recombination occurs more frequently in GC-rich regions (Bhérer *et al.*, 2017; Halldorsson *et al.*, 2019). Yet, more factors influencing the recombination events and hotspot formation,

and the molecular mechanism behind it are yet to be discovered (Baudat *et al.*, 2013; Bell *et al.*, 2020).

Several studies (Auton *et al.*, 2015; Bell *et al.*, 2020; Brick *et al.*, 2018; Coop *et al.*, 2008; Halldorsson *et al.*, 2019; Hinch *et al.*, 2019; Kong *et al.*, 2010; Mancera *et al.*, 2008; Myers *et al.*, 2008) have been conducted to demystify this essential genetic process, resulting in a large amount of data with different properties. The availability of such datasets enlightens the possibility of investigating this problem from a different angle, i.e. in a data-driven manner. Despite the extensive biological experiments and the development of computational tools to construct genetic maps (Bruen *et al.*, 2006; Spence and Song, 2019) and perform binary classification (Brown and Lunter, 2019; Chen *et al.*, 2013; Liu *et al.*, 2017) of hotspot and coldspot sequences, computationally, researchers have not investigated the determinants of recombination hotspots systematically and quantitatively. Based on the accumulated datasets from the previous studies, it is very promising to develop computational methods to perform cross-study, cross-sex, cross-population and cross-species investigation, potentially providing more insights into this crucial biological process.

Deep learning has been proven to be a successful approach for performing classifications (Eraslan *et al.*, 2019; LeCun *et al.*, 2015). However, directly applying deep-learning models to this problem may cause difficulties in studying the recombination hotspot determinants quantitatively due to the complexities and interpretability issue of the model (Zou *et al.*, 2019). To analyse the accumulated recombination data and facilitate the study of the yet unclear recombination process, we propose a novel transparent computational method, RHSNet, which combines the strength of deep learning (LeCun *et al.*, 2015), activation backpropagation (Shrikumar *et al.*, 2017) and signal processing (Kay, 1993), to systematically identify and quantify the recombination hotspot determinants taking advantage of data from multiple previous studies crossing different populations (Bell *et al.*, 2020; Halldorsson *et al.*, 2019; Lange *et al.*, 2016; Spence and Song, 2019), sexes (Brick *et al.*, 2018) and species (Lange *et al.*, 2016; Mancera *et al.*, 2008). In addition to predicting hotspot sequences accurately and identifying the well-known determinants, our method can quantify the relative contribution of each determinant, showing their differences in different sexes and species, as well as their evolution across different populations.

2 Materials and methods

2.1 An overview of RNSNet

Our method leverages the strength of deep learning, activation backpropagation and signal processing to first predict the recombination hotspot sequences, then quantify the contribution of the input information, and finally extract determinants, such as the PRDM9 binding motif. As shown in Figure 1A, during the prediction, the input sequences of various lengths, depending on the data, go through a specific deep-learning model, which consists of two 1-D convolutional layers as the sequence feature extractor, a Gated Recurrent Unit (GRU) for capturing long-range information and a multi-head attention layer for detecting interactions within the sequence (see Supplementary Fig. S2), to output useful information from the raw sequences. Because histone modifications are also shown to be crucial to recombination (Lange *et al.*, 2016; Spence and Song, 2019), we use another deep-learning module to process the information, including H3K4me3 and H3K36me3 from testis and ovary. Then, the feature vectors extracted from sequence information and ChIP-seq information are normalized before combination to predict whether the input sequence is a hotspot sequence. In addition, we are further interested in identifying and quantifying the recombination hotspot determinants. One previous study (Brown and Lunter, 2019) extracts motifs by considering only the activation of the first layer in the deep-learning model. But this method omits the complexity of the downstream layers and has difficulty in quantifying the motif's contribution based on the entire model. To resolve the issue, we utilize an activation backpropagation method (Shrikumar *et al.*, 2017). The prediction of a specific sequence is backpropagated through the entire network back to the original inputs to assign contribution scores to the motifs. Note that we consider the entire deep-learning model and compute the score in a purely data-driven manner. However, extracting determinant information remains a problem because the computed scores can be noisy, with peaks having various lengths along the sequence, as shown in Figure 1B. We resolve this problem using signal processing techniques. We apply a low-pass filter onto the contribution score array. Then, we extract the significant motifs between two valleys with a peak. The user has the freedom to choose the low-pass filter, either obtaining long determinants or short ones with high confidence. Based on the outputs of RHSNet, we further perform comprehensive quantitative analysis, as shown in Figure 1C, which will be discussed in detail below.

2.2 Dataset construction

In our study, we use datasets from a number of projects, including the Icelandic (Halldorsson *et al.*, 2019) dataset, the HapMap II (Frazer *et al.*, 2007) dataset, the Sperm (Bell *et al.*, 2020) dataset, the 1000 Genomes Project (Auton *et al.*, 2015) dataset, the Mice (Lange *et al.*, 2016) dataset and the Yeast (Mancera *et al.*, 2008) dataset. The Icelandic (Halldorsson *et al.*, 2019) dataset, provided by 1 476 140 COs from 56 321 paternal meiosis and 3 055 395 COs from 70 086 maternal meiosis, has a 642 bp resolution (655 bp for the paternal part) generated from Icelandic pedigrees on the GRCh38 human reference genome (Harrow *et al.*, 2006), from which we select 20 000 hotspots with an average recombination rate of 51.07 cM/Mb and 20 000 coldspots with an average recombination rate of $1.78 \times e^{-10}$ cM/Mb (resolution from 500 to 1000 bp) for cross-validation. Based on the fact that, in the sex-average map, the average length of those hotspots is relatively shorter (averaging 526 bps) than that of coldspots (averaging 3071 bps), we sort the recombination rates of all the possible sequences and select the lowest 20 000 coldspots with a proper resolution to construct the negative samples.

Similar to Myers *et al.* (2008), we acquired the HapMap II (Frazer *et al.*, 2007) dataset from Brown and Lunter (2019), in which the data are segmented into hotspots (average rate 10.5 cM/Mb) and coldspots (average rate below 0.5 cM/Mb) regions from a hidden Markov model with emission probabilities defined as p(observedrate—hot) and p(observedrate—non-hot). Sperm (Bell *et al.*, 2020) dataset is built with 31 228 sperm cells from 20 sperm donors, among which 813 122 COs from 787 aneuploid chromosomes are identified. The recombination rates vary in 20 sperm donors ranging from 22.2 to 28.1 COs per cell. A fine-scale genetic map is generated by us from those 813 122 COs events by stepping along the genome at 500 kb intervals, dividing the number of COs that occurs up to each point and by the total number of cells. We further select 5000 hotspots with an average recombination rate of 19.96 cM/Mb and 5000 coldspots with an average recombination rate of $1 \times e^{-20}$ cM/Mb for cross-validation. For detailed description about the Mice dataset (Lange *et al.*, 2016), PRDM9 lacking Yeast (Mancera *et al.*, 2008) (*Saccharomyces cerevisiae*) dataset, and 1000 Genomes Dataset (Auton *et al.*, 2015) could be found in the Supplementary Section S1.

2.3 Low-pass filter-based signal extraction from guided backpropagation

The previous method, known as DeepLIFT (Shrikumar *et al.*, 2017), successfully assigns contribution scores of input DNA sequences. By calculating the gradient of each activated neuron through guided backpropagation, contribution scores can be computed efficiently in a single backward pass based on the reference sequence with (A, C, G, T) distributed with probabilities as 0.3, 0.2, 0.2 and 0.3 (Supplementary Section S2), respectively. With the reference sequence $r_1^0, r_2^0, r_2^0 \ldots$ as input, the reference activation y^0 could be computed as:

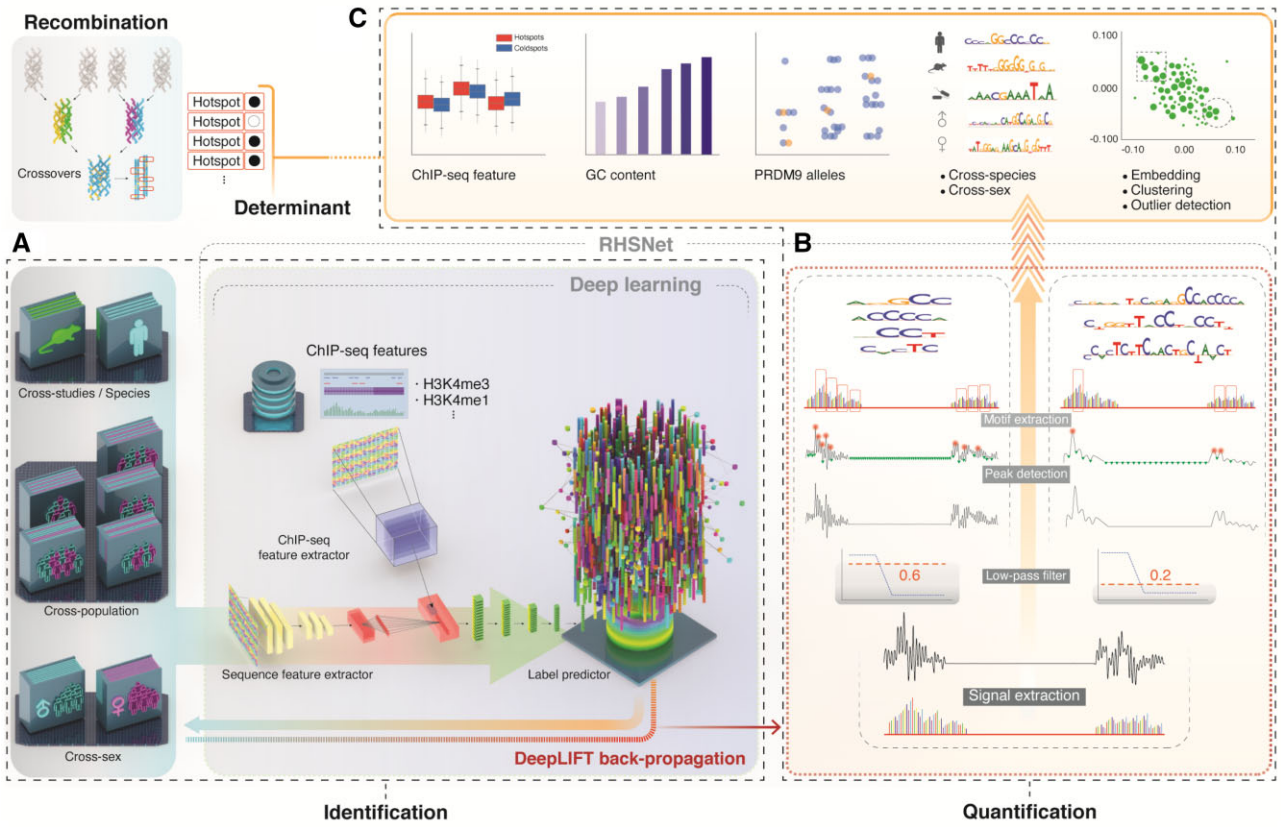

Fig. 1. Overview of the proposed framework, RHSNet, along with the proposed filter-based motif extraction approach. (**A**) The deep-learning algorithm of RHSNet accurately identifies recombination hotspots from different studies/species/populations/sexes, considering the ChIP-seq information. The model consists of two 1-D convolutional layers as the sequence feature extractor, a GRU for capturing long-range information and a multi-head attention layer for detecting interactions within the sequence. (**B**) RHSNet has a low-pass filter-based motif extractor that can quantify the contribution of hotspot-determinant motifs with flexible lengths (4–30 bp). We propose such a method based on gradient backpropagation and signal processing. (**C**) Comprehensive analysis across different studies/species/populations/sexes, based on RHSNet, provides more insights into the biological process and suggests the effectiveness of the proposed method

$$y^0 = f(r_1^0, r_2^0, r_2^0 \ldots). \tag{1}$$

The key idea of important score extraction is through guided backpropagation. For each neuron y, $\triangle y^+$ and $\triangle y^-$ are defined as having positive and negative component of $\triangle y$:

$$\triangle y = \triangle y^+ + \triangle y^-. \tag{2}$$

Now, given an input neuron x and the target neuron t, there is a difference of $\triangle t$ from the reference neuron r. The multiplier $m_{\triangle x \triangle t}$ could be defined as the contribution of $\triangle x$ to $\triangle t$ divided by $\triangle x$:

$$m_{\triangle x \triangle t} = \frac{C_{\triangle x \triangle t}}{\triangle x}. \tag{3}$$

Also, given $C_{\triangle x_i \triangle y_i}$ along with $C_{\triangle y_i \triangle t}$, we can show that the definition of $m_{\triangle x \triangle t}$ according to the chain rule would satisfy summation-to-delta:

$$\triangle t = \sum_i C_{\triangle x_i \triangle t}. \tag{4}$$

Notably, using the chain rule in which the input layers have one-hot encoded sequence $s_1, s_2, \ldots s_n$, hidden layers $y_1, y_2, \ldots y_n$ and the target output t, based on Equation (4), we can have:

$$m_{\triangle s_i \triangle t} = \sum_j m_{\triangle s_i \triangle y_j} m_{\triangle y_j \triangle t}. \tag{5}$$

We can compute the multipliers for each input sequence s_i efficiently via backpropagation. Inspired by DeepLIFT, a more advanced TF-MoDISco (Avsec *et al.*, 2019) introduced for transcription factor prediction was proposed. However, DeepLIFT and TF-MoDISco share a common disadvantage of having a strong assumption on the discovered motif length based on previously known probabilistic motif models. Such a strong assumption is further enhanced when adjusting the sliding window size similar to the expected length of the core motif and its flanks, which can be unknown for innovative motif discovery. Finding longer motifs is crucial in the recombination hotspot prediction task because each PRDM9 zinc finger is 28 amino acids long and is usually decoded within an 84 bp repeating tandem (Baudat *et al.*, 2010). Also, the PRDM9c (ZF8–13) motifs are usually 21 bp long, because they are accompanied by 5′ (five prime) and 3′ (three prime), making the traditional sliding window-based method less efficient.

A sliding window-based approach with a fixed window size is adopted by Brown and Lunter (2019) to find activated motifs from the networks' first layer, which has a strong assumption on the discovered motif length. Through backpropagation, RHSNet converts the discrete contribution scores into a continuous digital signal for the possible discovery of interesting motifs with variant lengths. In our motif extraction algorithm, we utilize low-pass filters with different kinds of factors and a peak detection algorithm, from which we could easily control the length range of the detected motifs by controlling the low-pass factor.

The contribution scores generated from backpropagation are firstly transformed into one-dimensional signals (see Fig. 3A) and fed into the low-pass filter. Opposite from high-pass filters, the low-pass filter allows low-frequency signals to pass, here, we show the transition function of an analogue low-pass filter with one order:

$$H(s) = \frac{1}{\frac{s}{W_n} + 1}, W_n = 2 \times \frac{f_c}{f_s}, \tag{6}$$

where f_c is the critical frequency and f_s is the sampling frequency. In practice, a low-pass filter with higher order would usually have a

better filtering performance. Therefore, an eighth-order low-pass filter is implemented in the programme. The W_n scalar can be flexibly chosen from [0.1,0.2,0.4] when users are looking for motifs with approximate length of 50, 10 and 5 bp, respectively. Detailed explanation can be found in Supplementary Section S2. The low-pass filter provides a smooth form of signals by eliminating short-term fluctuations and retaining long-term development trends, in which longer motifs enriched by a relatively high-frequency signal are reserved. By detecting each peak with its nearby valley, we could easily extract the motif in the middle. We choose 0.06 as the prominence parameter for the peaks. Also, the valleys are defined as the peaks of the reversed signal where the interval of each valley width is set to 1.

2.4 Enrichment factor definition

It is widely acknowledged that the identification of the DNA motifs is a vital task for recombination hotspot identification (Myers *et al.*, 2008). We define the motif enrichment factor by the ratio of the selected motif's contribution score over the average contribution score of the entire input sequence through backpropagation. Different from the recombination rate, which is an absolute value, the enrichment factor is more of a relative index, which indicates how strong the enriched motif signal is among the entire input sequence. That is, the larger the enrichment factor, the higher chance that such a cropped motif plays a more important role in the recombination events.

3 Results

3.1 Overall recombination hotspot prediction performance

Although our method is not designed specifically to perform binary recombination hotspot prediction, it can achieve superior performance on the task compared with existing methods. PseDNC (Chen *et al.*, 2013) was first proposed and tested on *S.cerevisiae* (yeast) chromosome. It needs to transform DNA sequences into a novel feature vector, namely pseudo amino acid, and then feed it into traditional classifiers, such as Support Vector Machine. Equivariant CNN (Brown and Lunter, 2019) as proposed as the first neural-net-based method towards this problem and was tested on HapMap II (Frazer *et al.*, 2007) data, and shows its strength as a motif finder. Here, we report the performance of the classification module in RHSNet (the Identification module in Fig. 1A). We evaluate the proposed deep-learning model's performance on different datasets [HapMap II (Frazer *et al.*, 2007), Icelandic (Halldorsson *et al.*, 2019) and Sperm (Bell *et al.*, 2020)], different sexes (Halldorsson *et al.*, 2019), different populations (Frazer *et al.*, 2007), different species (Lange *et al.*, 2016; Mancera *et al.*, 2008) and across different evaluation criteria, whose results are shown in Figure 2. With the same input data and evaluation criteria, our deep-learning model in RHSNet is constantly better than the competing models, including a simple 2-layer CNN, Equivariant CNN (Brown and Lunter, 2019) and PseDNC (Chen *et al.*, 2013), across different conditions in terms of $F1$ score (see Fig. 2A), except for the paternal Icelandic dataset. Meanwhile, on the sex-specific dataset, for which we can extract ChIP-seq information from the corresponding ovary and testis tissues, we utilize six histone modifications (H3K4me1, H3K4me3, H3K27ac, H3K9me3, H3K36me3 and H3K27me3) from testis and ovary tissues for the paternal and maternal datasets, respectively. Adding such information into our model (RHSNet-chip) can further boost the deep-learning model's performance, which is consistent with the previous research (Lange *et al.*, 2016; Spence and Song, 2019). To further test the robustness of our model, we evaluate the model's performance on hotspot regions with various lengths (see Fig. 2B, on the Icelandic dataset), using a different evaluation criterion, Matthews Correlation Coefficients (MCC). As illustrated in Figure 2B, our method is consistently better and shows more stable and robust performance than the baseline methods across different resolutions. As the resolution goes down and the prediction becomes less demanding, all the models' performance improves, although the training dataset size decreases. To further validate the generalization ability of RHSNet without ChIP-seq information, we test it against a dataset from different species. On the Mice dataset (Lange *et al.*, 2016), evaluated with Area Under the Receiver Operating Characteristic curve (AUROC), RHSNet can significantly improve over the existing methods (see Fig. 2C). We have also compared RHSNet with the non-default parameters for PseDNC and Equivariant CNN. The statistical results about RHSNet's identification performance can be referred to Supplementary Tables S4 and S5 and Supplementary Figures S3–S5.

3.2 PRDM9 binding motif, GC content and histone modification affect recombination hotspots

The existing research has shown that PRDM9 binding motif (Baudat *et al.*, 2010; Myers *et al.*, 2008, 2010; Parvanov *et al.*, 2010), histone modification (Lange *et al.*, 2016; Spence and Song, 2019) and GC content (Bhérer *et al.*, 2017; Halldorsson *et al.*, 2019) influence the recombination hotspot. We use RHSNet to analyse how the above factors are related to the recombination hotspots in different datasets. As we have discussed, in our method, with different low-pass filter factors, we can extract motifs with different lengths and different enrichment factors (see Figs 1 and 3A). In Figure 3B, we compare the top-ranked motifs regarding the enrichment factor from different datasets [HapMap II (Frazer *et al.*, 2007), Icelandic (Halldorsson *et al.*, 2019) and Sperm (Bell *et al.*, 2020)] with different filter factors against the PRDM9 binding motif4: CCNCCNTNNCCNC and SPO11-oligo (Lange *et al.*, 2016). Clearly, in the Icelandic dataset, the top-ranked motifs are highly correlated (top10: 87.5% $\pm$ 1.01; top100: 64.1% $\pm$ 1.54) with the canonical PRDM9 motif regarding the pairwise sequence alignment matching score. Although, compared with Icelandic, the PRDM9 pattern is less significantly enriched in the top-ranked motifs from the HapMap II dataset, we still obtain 53.91% GC content in the top 100 motifs (see Fig. 3D), with the PRDM9 binding pattern appearing in these motifs. In contrast, the top-ranked motifs from the Sperm datasets are different from the ones from the other two datasets, though the PRDM9 pattern still appears. Unlike the HapMap II (Frazer *et al.*, 2007) and Icelandic (Halldorsson *et al.*, 2019) datasets, the Sperm (Bell *et al.*, 2020) dataset focuses more on comparison across individuals' cells rather than aggregating them, and is resolved to much larger regions, with the median resolution as 240 kilo-base pairs (kb), among which 9746 (1.2%) are inferred within 10 kb. Consequently, we inevitably involve noisy sequences in the training dataset, which reduces the sensitivity of our method and also leads to lower prediction confidence compared to the other datasets when the filter factor is 0.4. GC content is shown to be positively correlated with the recombination rate (Bhérer *et al.*, 2017; Halldorsson *et al.*, 2019). As shown in Figure 3D and Supplementary Figure S6, in all the datasets, the GC content of the hotspots is indeed higher than that of the entire genome (HapMap II: 44.64% versus 39.26%, Icelandic: 42.81% versus 39.26%, Sperm: 48.16% versus 39.26%). The comparison of GC content in hotspots across different resolutions (see Fig. 3E) also suggests that the determinants in the central hotspot area are GC-richer than the marginal, detailed description can be found in Supplementary Section S4. Histone modifications usually accompany the recombination event, and in the PRDM9 knockout organism, DSB is directed by histone modifications (Brick *et al.*, 2012) (see Fig. 4A). We quantify the contribution of different histone modifications to recombination hotspot formation using activation backpropagation. In Figure 4B, on the Icelandic dataset, we compare the distribution of different features from three modifications between hotspots and coldspots, including H3K4me1, H3K4me3 and H3K36me3. The distribution differences suggest that the histone modification patterns are different in the two kinds of regions. We use the Icelandic maternal data and the histone modifications from the ovary to further investigate such features. In Figure 4C, we show the feature correlations among six histone modifications across hotspots. H3K4me1is correlated with H3K36me3, which is similar to that in the functional elements in the genome (Auboeuf *et al.*, 2002). To illustrate the contribution of histone modification features, in Figure 4D, we visualize the 2D vector embedding of recombination

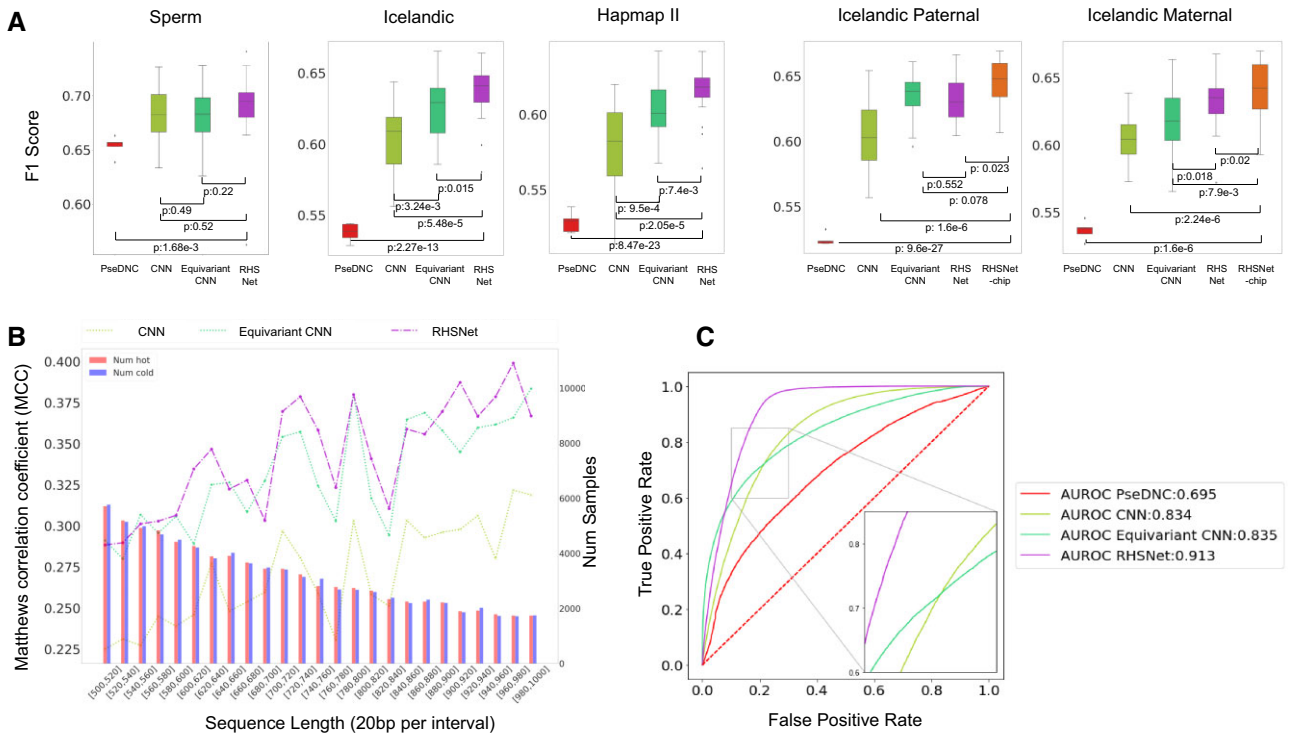

Fig. 2. Performance of RHSNet across different studies/sexes/species. Notice that, here, RHSNet refers to the Identification module of RHSNet in Figure 1. (**A**) The performance of RHSNet on datasets from different studies and different populations. We used 5-fold cross-validation to evaluate the performance of different methods including PseDNC (Chen *et al.*, 2013) and Equivariant CNN (Brown and Lunter, 2019) with their optimal parameters. The box plots show the *F*1-score distributions from four trials of the 5-fold cross-validation. *P*-values calculated from the two-tailed Student's *t*-test indicate the significance of the improvement. RHSNet-chip refers to RHSNet accompanied with ChIP-seq information in sex-specific maps. We used several histone modifications from testis for the paternal maps and ovary for the paternal map. (**B**) Robustness testing with MCC on the Human Icelandic dataset of RHSNet over different input lengths ranging from 500 to 1000 bp. The number of hotspot and coldspot sequences within each interval of 20 bp is showed as a histogram at the bottom. RHSNet's performance is robust across different resolutions. (**C**) On the Mice dataset, RHSNet also shows significant performance improvement on predicting the hotspot sequences in terms of the AUROC score. This result suggests the generalization ability of our method on different datasets, even across different species

hotspots and coldspots from the Principal Component Analysis (PCA) extracted from the last layer of CNN, RHSNet and RHSNet-chip. As shown in the figure, the proposed deep-learning model in RHSNet can learn different representations for hotspots and coldspots, and thus identify the hotspot regions. RHSNet-chip, incorporating the histone modification information, can further enlarge the difference in the learned representation between hotspots and coldspots. To further quantify the contribution of features from histone modifications, we use activation backpropagation across the entire network, visualizing their importance scores in Figure 4E. The results are consistent with previous studies (Brick *et al.*, 2012), with H3K4me3 and H3K36me3 being the two most essential modifications. Other related modifications, such as H3K4me1 (Yamada *et al.*, 2013) and H3K27ac (Chen *et al.*, 2020), are also captured by our method, although they are less studied for this problem.

3.3 RHSNet reveals the contribution of different PRDM9 alleles in different populations

Not only has PRDM9 been found to be the major determinant of the recombination hotspots in humans and mice (Baudat *et al.*, 2010), but different PRDM9 alleles are also believed to influence recombination hotspot activities in humans (Auton *et al.*, 2015; Berg *et al.*, 2010; Spence and Song, 2019). PRDM9-A is the most abundant allele in human populations (found in around 86% of European and around 50% of African populations), and PRDM9-C is the second most common one in African populations (12.8%) (Berg *et al.*, 2010). The two alleles have different binding preferences (see Supplementary Section S2). Despite the imperfect way of identifying the motifs, PRDM9-C binding motifs are found to potentially elevate the recombination rates in the African populations. On the dataset from Phase 3 of the 1000 Genomes Project (Auton *et al.*, 2015), in the African population, both detected PRDM9-A/C binding motifs show significantly higher recombination rates than the other populations (see Fig. 5A), which is consistent with the previous study (Spence and Song, 2019). Furthermore, among the top 100 motifs (low-pass filter: $Wn = 0.1$) for each population detected by our method, the ratio of PRDM9-A/C binding motifs in the African population (PRDM9-A ratio: 50.4%; PRDM9-C ratio: 33.3%) is much higher than that of the other populations (see Fig. 5B, Supplementary Tables S6 and S7 and Supplementary Fig. S9). On the other hand, as we define enrichment factor by considering the recombination rate of the entire region around the motif, the enrichment factor of PRDM9-C binding motifs in the African population is corrected to be on the same level as the other populations due to the higher overall recombination rate in the population (see Fig. 5C). Although using the absolute value of the recombination rate to show the contribution of recombination hotspot determinants is straightforward, our framework provides an alternative quantification method by utilizing the relative criterion, which may be more robust to the local region noise and population batch effect. The relation between the recombination rate and the enrichment factor of a motif is complex, which cannot be modelled with a linear regression (see Fig. 5D), with the Pearson correlation coefficient being -0.034 and the R^2 score being 1.79×10^{-3} on motifs extracted from five populations. However, together with the recombination rate, by quantifying the relative contribution, our method provides insights into the recombination hotspot determinants.

3.4 Generalization to PRDM9-lacking species and sensitivity to sex differences

The recombination event has been studied in a broad range of species (Lam and Keeney, 2015; Lange *et al.*, 2016; Pan *et al.*, 2011;

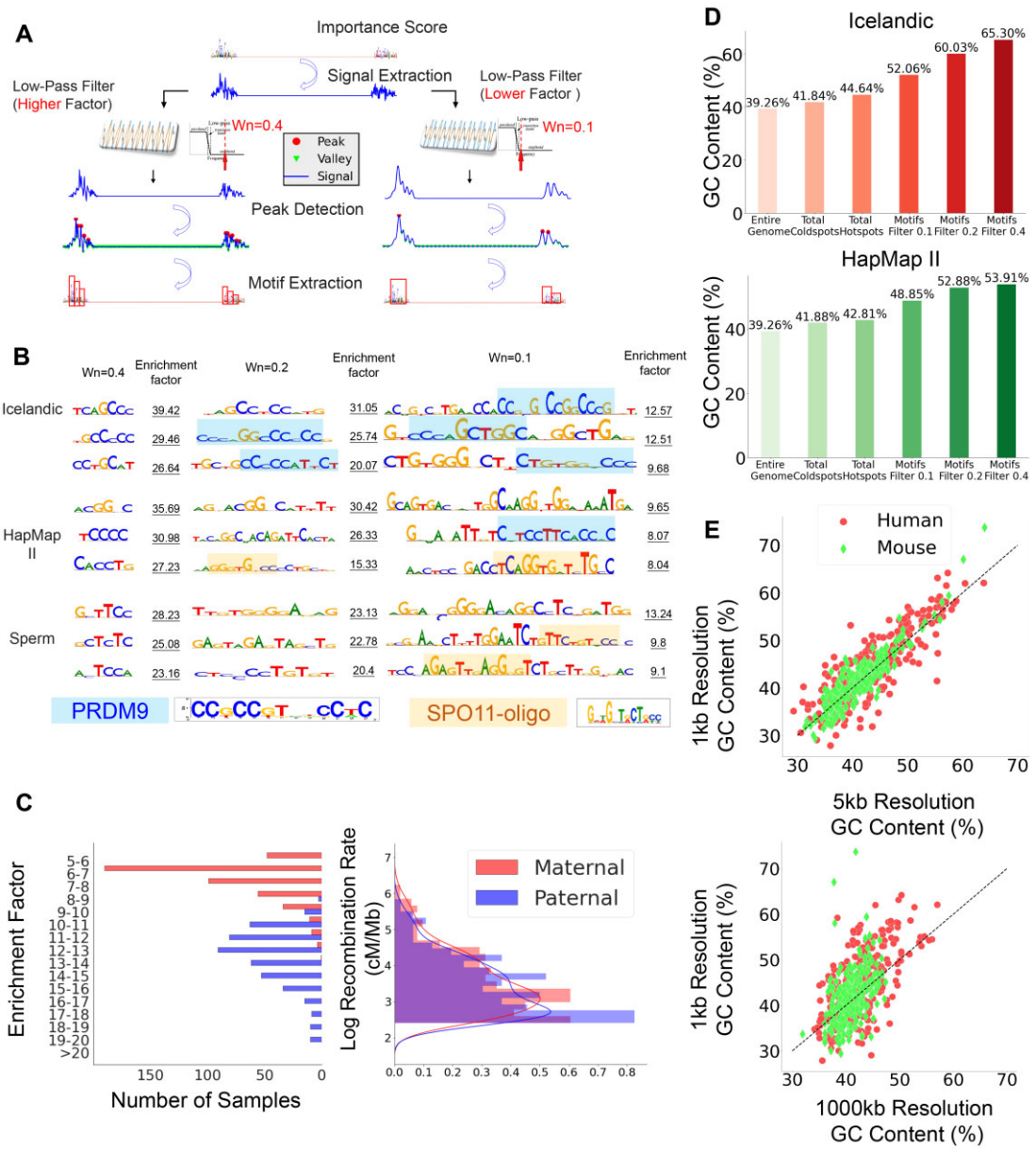

Fig. 3. RHSNet quantifies the contribution of PRDM9 binding motifs in variant lengths across different studies/populations/sexes. (**A**) RHSNet can extract motifs of variant lengths with different low-pass filter factors. (**B**) Assembled motif detection results are shown using different low-pass filters ranging from 0.1 to 0.4. PRDM9 binding motifs and 12-bp motif enriched in SPO11-oligo hotspots show high enrichment factors. (**C**) The logged recombination rates and the enrichment factor distribution across sexes within the detected motifs. The recombination rates are higher in the females, while the enrichment factors are higher in the males. The lower erosion rate of the hotspot motifs in males may make the event determinants more conservative. (**D**) GC content compared across different studies. We show the GC content among the entire genome, total hotspots, total cold spots and RHSNet's detected motifs. (**E**) GC content of recombination hotspots cropped in different resolutions (1, 5 and 1000 kb) across different species. The GC content is usually higher in the hotspots and the determinant motifs compared to the nearby regions

Singhal *et al.*, 2015), including humans, mice, yeast, birds and pigs. In addition to the human data, our method can be further applied to other species, regardless of having the PRDM9 gene. Although PRDM9 is shown to be the major determinant of recombination hotspots in both humans and mice, the predicted binding motifs are different in the two species (Baudat *et al.*, 2010). We apply our method to the Mice data (Lange *et al.*, 2016) and identify the most significant determinant motifs. In addition to the GT-rich motifs, which are enriched in the SPO11-oligo hotspots and the usual mice PRDM9 binding sites (Lange *et al.*, 2016), surprisingly, we also identify an AC-rich motif (see Supplementary Fig. S10A). Although this motif has not been studied extensively in the mice-related literature, it is reported as a part of the binding motif for the PRDM99R zinc finger binding domain (Brick *et al.*, 2012; Lange *et al.*, 2016). Unlike apes and mice, birds and yeast lack a PRDM9 gene, leading to different recombination hotspot patterns in these species (Lam and Keeney, 2015; Pan *et al.*, 2011; Singhal *et al.*, 2015). On the Yeast dataset (Mancera *et al.*, 2008), the poly-(A) motif is identified as the most significant determinant in hotspots (see Supplementary Fig. S10A), which is completely different from mice and humans. However, the result is consistent with the previous study, which demonstrates that Poly-(A) motif occurs more frequently in the hotspots (Mancera *et al.*, 2008). Moreover, the motifs enriched in the yeast promoters (Badis *et al.*, 2008) are also

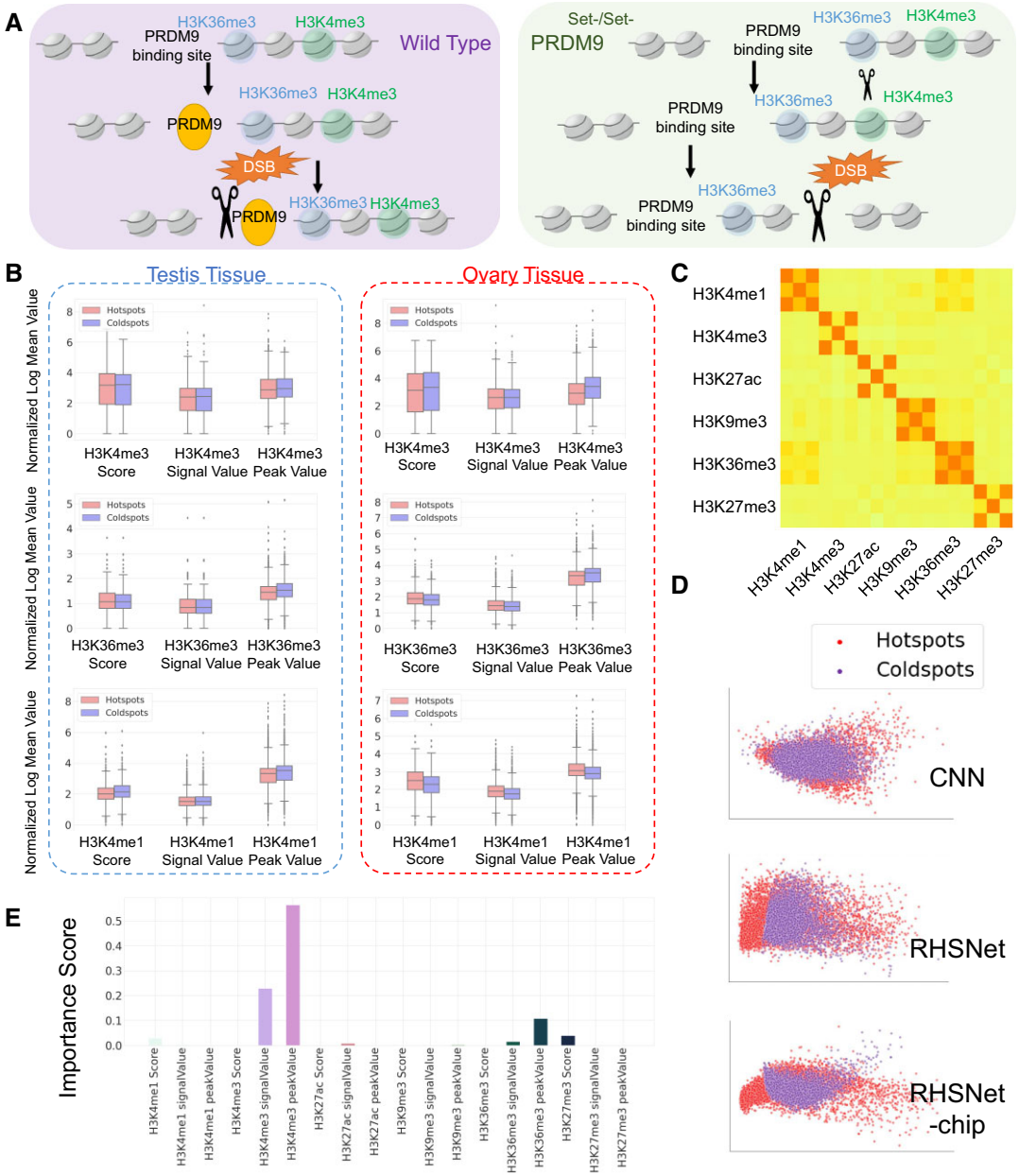

Fig. 4. Histone modification affects recombination hotspots formulation. (**A**) The DSB formation machinery (scissors) is directed to PRDM9 binding sites with functional PRDM9 protein. However, in the absence of PRDM9, DSB is directed to PRDM9-independent H3K4me3 marks. (**B**) On the maternal and paternal maps from the Icelandic dataset, we show the feature distribution comparison from recombination hotspots and coldspots over three histone modifications: H3K4me1, H3K4me3 and H3K36me3. (**C**) Heatmap of the 18 feature correlations within six histone modifications across hotspots. (**D**) For the ChIP-seq feature of female adult's ovary tissue, we show the 2D vector embedding of recombination hotspots and coldspots from the PCA extracted from the last layer of CNN, RHSNet and RHSNet-chip. The difference between hotspot features and coldspot features is more significant in the RHSNet-chip framework, demonstrating the importance of the ChIP-seq feature, although RHSNet alone is significant enough. (**E**) The importance score calculated from the contribution backpropagation in the ChIP-seq feature extraction branch quantifies the contribution of each histone feature to hotspot formulation. The activation of the H3K4me3 features and the H3K36me3 features suggests a higher contribution to the recombination event

predicted to be of vital importance to the recombination event (see Supplementary Fig. S10A), which supports the theory that, in the PRDM9-lacking species, hotspots are highly conserved due to the natural selection pressure (Lam and Keeney, 2015). Detailed description for RHSNet's sensitivity to sex differences can be found in Supplementary Section S4.

3.5 Recombination hotspot motif embedding for evolutionary determinants discovery

We extend our method to identify and quantify the recombination hotspot determinants more intuitively and systematically. Instead of only listing the detected determinants and their enrichment factors, we learn the representation of the motifs in a 2D space, visualizing and clustering them in such a space. To avoid black-box modelling, we also visualize the physical meaning of the motifs with heatmaps (see Fig. 6). Within the Icelandic dataset (see Fig. 6A), the standard deviation of the determinant motif embeddings in the maternal population is much larger than that in the paternal population (maternal: $0.0447 \pm 1.9 \times 10^{-2}$, $P = 3.11 \times 10^{-6}$; paternal: $0.0355 \pm 1.6 \times 10^{-2}$), which suggests that, in females, the recombination hotspot determinants are more diverse and less conservative than that in the males. This finding further supports our hypothesis that diverse factors contribute to the female-biased hotspots. For all

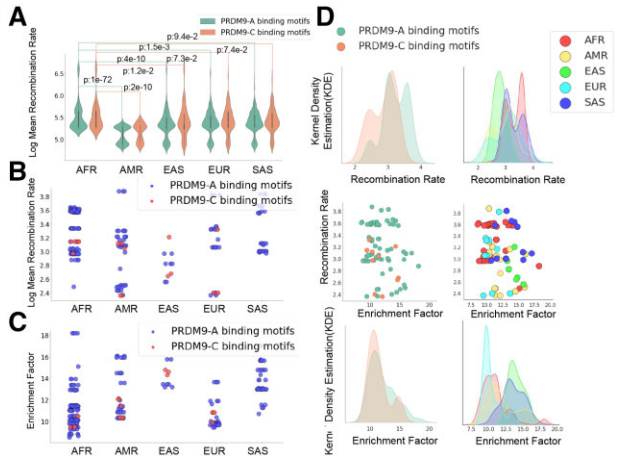

Fig. 5. Recombination rates and enrichment factors of PRDM9 alleles within different populations. (**A**) Recombination rates of ground truth hotspots at different PRDM9 binding motifs in five populations, normalized by log average recombination rate. The African population has significantly higher recombination rates in both alleles compared to the other populations. (**B**) Recombination rates of the RHSNet-identified PRDM9-A and PRDM9-C binding motifs. The recombination rate of PRDM9-A/C alleles in the African population is significantly higher than that in other populations. (**C**) Enrichment factors of the RHSNet-identified PRDM9-A and PRDM9-C binding motifs. The enrichment factor of PRDM9-C binding motifs in the African population is corrected to be on the same level as the other populations due to the higher overall recombination rate in the population. (**D**) Paired relation between enrichment factors and recombination rates among all the detected PRDM9-A/C alleles across different populations. We show the recombination rate distribution of all the PRDM9-A/C alleles together and within each population in the upper row. As shown in the middle row, the relation between enrichment factors and recombination rates is more complex than linear correlation. In the bottom row, we illustrate the enrichment factor distribution of all the PRDM9-A/C alleles together and within each population

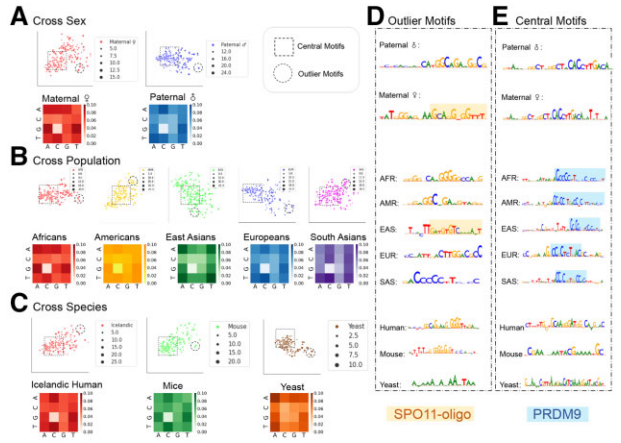

Fig. 6. Motif embedding and outlier detection for recombination hotspot determinants discovery. (**A**) Visualization of the motif embeddings over different sexes in the Icelandic dataset. The size of the scatter reflects the enrichment factor of that motif. Each heatmap below the cluster suggests the 2-mer appearance frequency in the detected motifs. The representative central and outlier motifs are shown in (**D**) and (**E**). (**B**) Visualization of the motif embeddings across different populations. The enrichment factors differ significantly between central motifs and outlier motifs across populations. (**C**) Visualization of the motif embeddings across different species. Poly-(A) motifs are the most enriched outlier motifs in Yeast hotspots. Also, the Human/Mice outlier shows evolutionary discovery in motifs correlated with SPO11 oligos

the visualizations of the motifs across different sexes (see Fig. 6A and Supplementary Fig. S13), different populations (see Fig. 6B and Supplementary Figs S15 and S16) and different species (see Fig. 6C and Supplementary Figs S12 and S14), the motifs within the central area of the embedding space tend to have a smaller enrichment factor value, represented by the size of the point, than the outlier motifs. Because the enrichment factor value shows the importance of the determinant, investigating the outlier motifs may identify the evolutionary important motifs of the population and species. Similar to the Icelandic data, within the 1000 Genomes Project dataset, the enrichment factor differences between central motifs and the outliers across different populations share a similar pattern (see Fig. 6B), where enrichment factors differ significantly between central motifs and outlier motifs across populations, especially in East Asians (AFR: $P = 2.85 \times 10^{-4}$; AMR: $P = 8.51 \times 10^{-2}$; EAS: $P = 1.12 \times 10^{-5}$; EUR: $P = 1.16 \times 10^{-2}$; SAS: $P = 0.23$). For the convenience of the study, we randomly select the most distinct outlier motifs in different populations across the embedding space and visualize them in Figure 6D. Interestingly, the motifs enriched in the SPO11-oligo hotspots (Lange *et al.*, 2016) show up in the East Asian population. Although the molecular studies are mainly performed on Mice (Lange *et al.*, 2016), and researchers have not performed such studies on different human populations systematically, our method provides the first quantitative depiction of the recombination hotspot determinant motifs across diverse populations. We further extend our analysis to different species (see Fig. 6C). A similar pattern appears. The central motifs, shared by different species, have smaller enrichment factors than the outlier motifs, which are likely to be species-specific (Icelandic human: $P = 1.57 \times 10^{-25}$; Mice: $P = 2.2 \times 10^{-8}$; Yeast: $P = 7.4 \times 10^{-5}$). For example, the poly-(A) motifs are the most important ones for yeast, which does not have the PRDM9 gene (Mancera *et al.*, 2008). On the other hand, our method provides a new way to define the evolutionary distance between different species (Shen *et al.*, 2020), using the embedding of the recombination hotspot determinants.

4 Discussion

Recombination is one of the most important processes in miosis for sexually reproducing organisms, which can produce genetic diversity for natural selection. Despite its important role in evolution, people know little about the entire process and its molecular mechanism. Although a large amount of data have been accumulated from various giant projects, such as HapMap (Frazer *et al.*, 2007) [3.1 million human single nucleotide polymorphisms (SNPs) genotyped in 270 individuals], Sperm (Bell *et al.*, 2020) (31 228 human gametes from 20 sperm donors) and 1000 Genomes Project (Auton *et al.*, 2015) (84.7 million SNPs of 2504 individuals), seldom have researchers developed methods to analyse data from different studies and even different species.

Here, we propose a new computational method, RHSNet, which enjoys the strength of deep learning (LeCun *et al.*, 2015), activation backpropagation (Shrikumar *et al.*, 2017) and signal processing (Kay, 1993), to identify and quantify the recombination hotspot determinants. Although our method is not designed specifically for recombination hotspot region prediction, it can outperform almost all the previous methods in this task across different studies, different populations, different sexes and different species. More importantly, RHSNet can identify and quantify the determinants that contribute significantly to the recombination hotspot formation. In addition to quantifying the relation between PRDM9 binding motif (Baudat *et al.*, 2010; Myers *et al.*, 2008, 2010; Parvanov *et al.*, 2010), histone modification (Lange *et al.*, 2016; Spence and Song, 2019), GC content (Bhérer *et al.*, 2017; Halldorsson *et al.*, 2019) and recombination hotspots, it reveals the contribution of different PRDM9 alleles in different populations. Further studies on different species, including PRDM9-lacking species, and different sexes suggest the generalization power and sensitivity of the proposed method. The cross-sex, cross-population and cross-species studies show the potential of our method to identify the evolutionary determinants. Although RHSNet is purely data-driven and more work can be done to further improve it, including landscape prediction across the entire genome (Adrion *et al.*, 2020), using the gene annotation related to the location information (Yandell and Ence, 2012), chromatin accessibility information (Kumasaka *et al.*, 2016) as well as the conditional analysis (Wu *et al.*, 2018), it is potentially helpful

to assist researchers in illuminating the mechanisms underlying recombination and evolution.

Acknowledgements

Figure 1 is created by Heno Hwang, a scientific illustrator at King Abdullah University of Science and Technology (KAUST).

Data availability

Recombination hotspots and coldspots datasets for training and testing are acquired from multiple data sources across different studies, sexes, populations, and different species, and are now made available at: https://drive.google.com/drive/folders/1kNMrZgsceA_ZuwhEDjEeiA3eSBHix2Cv?usp=sharing.

Funding

This work was supported by the KAUST Office of Sponsored Research (OSR) under award numbers [BAS/1/1624-01, FCC/1/1976-23-01, FCC/1/1976-26-01, REI/1/0018-01-01, REI/1/4216-01-01, REI/1/4437-01-01, REI/1/4473-01-01, URF/1/4098-01-01, REI/1/4742-01-01].

Conflict of Interest: none declared.

References

Adrion,J.R. *et al.* (2020) Predicting the landscape of recombination using deep learning. *Mol. Biol. Evol.*, **37**, 1790–1808.

Auboeuf,D. *et al.* (2002) Coordinate regulation of transcription and splicing by steroid receptor coregulators. *Science*, **298**, 416–419.

Auton,A. *et al.*; The 1000 Genomes Project Consortium. (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.

Avsec,Ž. *et al.* (2019) Deep learning at base-resolution reveals motif syntax of the cis-regulatory code. *BioRxiv*, 737981. https://doi.org/10.1101/737981.

Badis,G. *et al.* (2008) A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters. *Mol. Cell.*, **32**, 878–887.

Baudat,F. *et al.* (2010) PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science*, **327**, 836–840.

Baudat,F. *et al.* (2013) Meiotic recombination in mammals: localization and regulation. *Nat. Rev. Genet.*, **14**, 794–806.

Bell,A.D. *et al.* (2020) Insights into variation in meiosis from 31,228 human sperm genomes. *Nature*, **583**, 259–264.

Berg,I.L. *et al.* (2010) PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. *Nat. Genet.*, **42**, 859–863.

Bhérer,C. *et al.* (2017) Refined genetic maps reveal sexual dimorphism in human meiotic recombination at multiple scales. *Nat. Commun.*, **8**, 14994–14999.

Brick,K. *et al.* (2012) Genetic recombination is directed away from functional genomic elements in mice. *Nature*, **485**, 642–645.

Brick,K. *et al.* (2018) Extensive sex differences at the initiation of genetic recombination. *Nature*, **561**, 338–342.

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

Bruen,T.C. *et al.* (2006) A simple and robust statistical test for detecting the presence of recombination. *Genetics*, **172**, 2665–2681.

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

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

Coop,G. *et al.* (2008) High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. *Science*, **319**, 1395–1398.

Eraslan,G. *et al.* (2019) Deep learning: new computational modelling techniques for genomics. *Nat. Rev. Genet.*, **20**, 389–403.

Frazer,K.A. *et al.*; International HapMap Consortium. (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–861.

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

Harrow,J. *et al.* (2006) GENCODE: producing a reference annotation for ENCODE. *Genome Biol.*, 7, S4.

Hinch,A.G. *et al.* (2019) Factors influencing meiotic recombination revealed by whole-genome sequencing of single sperm. *Science*, **363**, eaau8861.

Jin,X. *et al.* (2021) Genome-wide variability in recombination activity is associated with meiotic chromatin organization. *Genome Res.*, **31**, 1561–1572.

Kay,S.M. (1993) *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice-Hall, Inc.

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

Kumasaka,N. *et al.* (2016) Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nat. Genet.*, **48**, 206–213.

Lam,I. and Keeney,S. (2015) Nonparadoxical evolutionary stability of the recombination initiation landscape in yeast. *Science*, **350**, 932–937.

Lange,J. *et al.* (2016) The landscape of mouse meiotic double-strand break formation, processing, and repair. *Cell*, **167**, 695–708.

LeCun,Y. *et al.* (2015) Deep learning. *Nature*, **521**, 436–444.

Liu,B. *et al.* (2017) iRSpot-EL: identify recombination spots with an ensemble learning approach. *Bioinformatics*, **33**, 35–41.

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

Myers,S. *et al.* (2008) A common sequence motif associated with recombination hot spots and genome instability in humans. *Nat. Genet.*, **40**, 1124–1129.

Myers,S. *et al.* (2010) Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. *Science*, **327**, 876–879.

Paiano,J. *et al.* (2020) ATM and PRDM9 regulate SPO11-bound recombination intermediates during meiosis. *Nat. Commun.*, **11**, 1–15.

Pan,J. *et al.* (2011) A hierarchical combination of factors shapes the genome-wide topography of yeast meiotic recombination initiation. *Cell*, **144**, 719–731.

Parvanov,E.D. *et al.* (2010) PRDM9 controls activation of mammalian recombination hotspots. *Science*, **327**, 835–835.

Shen,X.-X. *et al.* (2020) An investigation of irreproducibility in maximum likelihood phylogenetic inference. *Nat. Commun.*, **11**, 1–14.

Shrikumar,A. *et al.* (2017) Learning important features through propagating activation differences. In: Proceedings of the 34th International Conference on Machine Learning, PMLR. pp. 3145–3153.

Singhal,S. *et al.* (2015) Stable recombination hotspots in birds. *Science*, **350**, 928–932.

Spence,J.P. and Song,Y.S. (2019) Inference and analysis of population-specific fine-scale recombination maps across 26 diverse human populations. *Sci. Adv.*, **5**, eaaw9206.

Wu,Y. *et al.* (2018) Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits. *Nat. Commun.*, **9**, 1–14.

Yamada,S. *et al.* (2013) 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 Res.*, **41**, 3504–3517.

Yandell,M. and Ence,D. (2012) A beginner's guide to eukaryotic genome annotation. *Nat. Rev. Genet.*, **13**, 329–342.

Zou,J. *et al.* (2019) A primer on deep learning in genomics. *Nat. Genet.*, **51**, 12–18.