Prediction of Trans-regulators of Recombination Hotspots in Mouse Genome

Min Wu
School of Computer Engineering
Nanyang Technological University
Singapore
wumin@ntu.edu.sg

Chee-Keong Kwoh
School of Computer Engineering
Nanyang Technological University
Singapore
asckkwoh@ntu.edu.sg

Teresa M Przytycka
Computational Biology Branch
NCBI, NLM, National Institutes of Health
Bethesda, MD 20894, USA
przytyck@ncbi.nlm.nih.gov

Jing Li
EECS Department
Case Western Reserve University
Cleveland, Ohio 44106, USA
jingli@cwru.edu

Jie Zheng*
School of Computer Engineering
Nanyang Technological University
Singapore
*Corresponding author: zhengjie@ntu.edu.sg

***Abstract*—The regulatory mechanism of recombination is a fundamental problem in genomics, with wide applications in genome wide association studies, birth-defect diseases, molecular evolution, cancer research, etc. In mammalian genomes, recombination events cluster into short genomic regions called "recombination hotspots". Recently, a 13-mer motif enriched in hotspots is identified as a candidate cis-regulatory element of human recombination hotspots; moreover, a zinc finger protein, PRDM9, binds to this motif and is associated with variation of recombination phenotype in human and mouse genomes, thus is a trans-acting regulator of recombination hotspots. However, this pair of cis and trans-regulators covers only a fraction of hotspots, thus other regulators of recombination hotspots remain to be discovered. In this paper, we propose an approach to predicting additional trans-regulators from DNA-binding proteins by comparing their enrichment of binding sites in hotspots. Applying this approach on newly mapped mouse hotspots genome-wide, we confirmed that PRDM9 is a major trans-regulator of hotspots. In addition, a list of top candidate trans-regulators of mouse hotspots is reported. Using GO analysis we observed that the top genes are enriched with function of histone modification, highlighting the epigenetic regulatory mechanisms of recombination hotspots.**

***Index Terms*—Recombination hotspots, PRDM9, trans-regulation, motif-finding, transcription factor binding site, epigenetics and histone modification**

I. INTRODUCTION

Recombination is one of the most fundamental processes in molecular biology, and is under intense research in genomics. In many species, recombination events are clustered into narrow genomic regions (usually a few kb long) called "recombination hotspots". During meiosis, recombination events are required to ensure correct segregation of homologous chromosomes, and thus abnormality or absence of meiotic recombination can lead to aneuploidy disorders such as Down syndrome. In addition to mutations, recombination is an important evolutionary force that shapes the linkage disequilibrium (LD) patterns in human genetic variation; as a result, hotspots tend to overlap with boundaries of haplotype blocks, which is a key observation underlying genome-wide association studies (GWAS) and the HapMap project [6]. Therefore, an increased understanding of the mechanism of recombination hotspots would shed light on various important aspects in molecular biology and medicine, such as genome instability, disease gene mapping, molecular evolution, etc. Despite the importance of recombination hotspots, however, many questions remain open, such as the regulatory mechanisms of the locations and activities of hotspots.

Recently, breakthroughs have been made to discover the regulatory mechanisms of meiotic recombination hotspots in mammalian gnomes. In 2010, three Science papers [16], [4], [18] reported the identification of PRDM9 gene as a trans-regulator of recombination hotspots in human and mouse genomes. PRDM9 is a zinc finger protein that binds to DNA, and its binding site contains a 13-mer motif previously found to be enriched in human hotspots [17]. Using an LD-based approach named LDsplit, Zheng et al. [26] identified HapMap SNPs (single nucleotide polymorphisms) in human chromosome 6 that are associated with recombination hotspots, and confirmed the sperm typing experimental result on DNA2 hotspot [11]. Importantly, proximal to the SNPs identified by LDsplit, Zheng et al. found an enriched 11-mer motif which partially matches the aforementioned 13-mer motif in the binding site of PRDM9 [26]. Using Chip-Seq data, Smagulova et al. [23] analyzed the molecular features of mouse recombination hotspots, and observed that a consensus motif enriched in mouse hotspots aligns with the predicted binding site of mouse PRDM9 significantly. These exciting discoveries are promising to integrate previously separate observations into one picture.

It has been observed that, despite over 99% sequence identity between the human and chimpanzee genomes, the positions of recombination hotspots are rarely conserved between the two species [25]. This puzzle has been partially answered by Myers et al. [16], who found that, as PRDM9 evolves rapidly, its binding sites are very different between human and chimpanzee. The “hotspot paradox” states that due to biased gene conversion a hotspot tends to kill itself, nevertheless, there remain many hotspots in extant genomes [20]. This paradox may be explained by the rapid evolution of PRDM9 as well, i.e. many new hotspots can be generated in a short time by a few mutations in the zinc finger binding array of PRDM9. It is believed that epigenetic mechanisms play key roles in the regulation of meiotic recombination. PRDM9 is a transcription factor with epigenetic functions (e.g. histone H3K4 trimethyltransferase activity). Importantly, PRDM9 is uniquely expressed in early meiosis and its deficiency is associated with sterility, which coincides with the association of meiotic recombination hotspots with birth-defect diseases. However, it is estimated that PRDM9 can explain only 18% of variations in human recombination phenotype [4], and the 13-mer motif covers only 41% of human hotspots [17]. Therefore, PRDM9 is unlikely to be the only trans-regulator of recombination hotspots. To carry out recombination accurately, it must function in concert with other proteins to form a regulatory pathway. Hence, it is highly motivated to discover other genes and regulatory pathways regulating recombination hotspots.

The approaches to the discovery of PRDM9 and recent related works on recombination hotspots [12] typically search for enriched motifs in hotspots, and then search for proteins that may bind to the motifs. Although successful in the discovery of PRDM9, this approach has a few limitations. First, unsupervised motif-finding is a notoriously difficult problem, and motifs found in this way tend to be short due to the limited power of motif-finding algorithms and large amounts of sequence data. Second, it may be difficult to infer the protein that binds to a short enriched motif, either because the enrichment of the motif is not due to the binding of a trans-regulator of hotspots, or because many proteins bind to the same motif. Last but not least, the procedure of identifying PRDM9 is a manual process that requires biochemical and genetic knowledge rather than an automatic discovery in large scale. The emergence of high-throughput genomic data of more human populations and other species calls for an efficient automatic procedure for discovering trans-regulators like PRDM9. It is our goal to develop such a method of genome-wide discovery of trans-regulators of recombination hotspots.

In this paper, we propose an approach to discovering trans-regulatory proteins similar to PRDM9 of recombination hotspots in mouse genome. Instead of starting from short sequence motifs enriched in hotspots, we scan the binding sites of DNA-binding proteins (e.g. transcription factors) across DNA sequences of hotspots and coldspots. As shown in Figure 1, the statistical score based on the enrichment of target binding sites is designed to predict the likelihood of each protein to initiate recombination. Moreover, a novel method is designed to identify the Gene Ontology (GO) terms that are shared by candidate trans-regulators. Applying this pipeline of knowledge discovery on a genome-wide map of mouse hotspots recently published [23], we first confirmed that PRDM9 is a major trans-regulator of mouse hotspots. Second, we identified a list of top candidate trans-regulators of mouse hotspots. Interestingly, our GO analysis shows that the candidate regulators predicted as such are enriched with the function of histone modification, highlighting the epigenetic regulatory mechanisms known to be key in recombination hotspots. Our method can be used for the automatic discovery of trans-regulators in addition to PRDM9 on new genetic data of humans and other species. The results in this paper not only confirm the discovery of PRDM9 gene, but also provide new candidate proteins to guide further experimental studies of recombination hotspots.

II. Methods

In this section, we introduce our method for predicting tran-regulators as shown in Figure 1. Firstly, we collect mouse TFs and their binding sequences. Secondly, we define odds ratio for these TFs showing their preference for binding in hotspots by analyzing FIMO search outputs. This odds ratio is then utilized to show the significance of each TF’s regulatory function for recombination. Lastly, we perform GO term validation and analysis of hotspot coverage for our candidate trans-regulators.

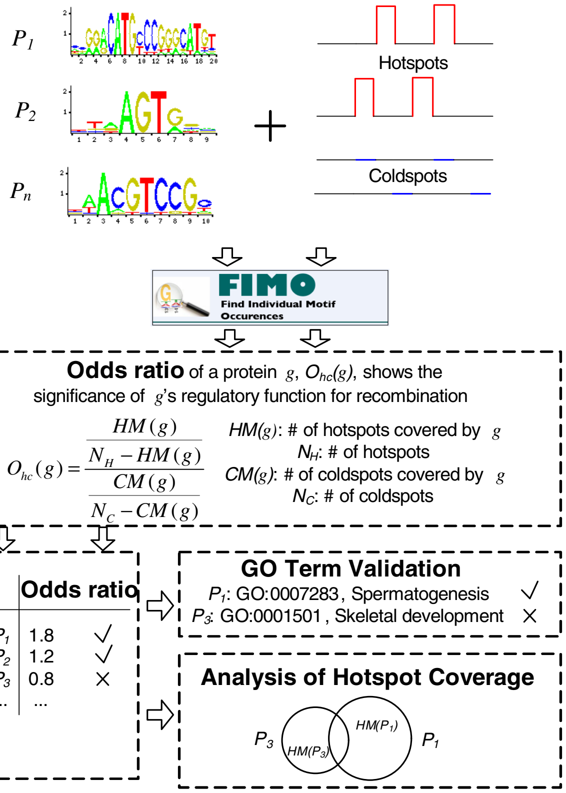
P_1
P_2
P_n
Hotspots
Coldspots
FIMO
Find Individual Motif Occurences
Odds ratio of a protein g, $O_{hc}(g)$, shows the significance of g's regulatory function for recombination
$O_{hc}(g)=\frac{\frac{HM(g)}{N_H-HM(g)}}{\frac{CM(g)}{N_C-CM(g)}}$
$HM(g)$: # of hotspots covered by g
N_H: # of hotspots
$CM(g)$: # of coldspots covered by g
N_C: # of coldspots
Odds ratio
P_1 1.8 ✓
P_2 1.2 ✓
P_3 0.8 ✗
... ...
GO Term Validation
P_1: GO:0007283, Spermatogenesis ✓
P_3: GO:0001501 , Skeletal development ✗
Analysis of Hotspot Coverage
P_3
$HM(P_3)$
$HM(P_1)$
P_1

Fig. 1. The flowchart of our method for predicting tran-regulators of recombination hotspots.

A. Derive consensus motif

Persikov et al. [19] proposed a method to predict the binding between a DNA motif and a given protein using support vector machines (SVM). Myers et al. [16] first applied this method to generate some approximate motifs as candidates. Each candidate motif will be assigned a score showing its propensity to bind with the given protein. Since the potential interactions between zinc fingers were not taken into consideration in [19], Myers et al. then continued to maximize the score of a candidate binding motif by successively changing single bases within it. Final binding motifs were then obtained when no score increases for them. The predicted mouse PRDM9 binding sequence and degeneracy are shown in Figure S5 in the supporting online material of [16]. We thus take this predicted PRDM9 binding sequence as our consensus motif.

Transcription factor (TF) binding affinities are typically modeled as position frequency matrices. JASPAR database (http://jaspar.genereg.net) [21] provides open-access for matrix profiles describing the DNA-binding patterns of TFs. The current release of Jaspar database holds 457 non-redundant, curated matrix profiles. For example, there are 53 for mouse, 75 for human and 117 for yeast. These 53 matrix profiles for mouse TFs are used in this paper to predict trans-regulators of mouse hotspots.

B. Protein binding preference in hotspots

For above binding motifs, we employ the software tool FIMO [10] to scan for their occurrences in both hotspots and coldspots. FIMO takes two files as inputs, namely, a file containing one or more query motifs and another file as the sequence database. Particularly, each query motif is represented as a position-specific frequency matrix and the sequence database consists of known hotspots and our generated coldspots (see the Results section for more details about coldspot generation). FIMO computes a log-likelihood ratio score for each position of the given sequence database and converts this score to p-value and q-value to show the statistical significance of this position. Finally, FIMO outputs a ranked list of motif occurrences, each of them associated with a log-likelihood ratio score, p-value and q-value.

Using the numbers of motif occurrences in hotspots and coldspots, as output by FIMO search, we measure the preference of a protein to bind in hotspots with the odds ratio $O_{hc} = (HM/HN)/(CM/CN)$. Here, HM is the number of hotspots with at least one motif occurrence (i.e. a hit of FIMO search), HN is the number of hotspots without any hit (i.e. $HN = N_H - HM$, N_H is the number of hotspots as shown in Figure 1), CM is the number of coldspots with at least one hit, and CN is the number of coldspots without any hit (i.e. $CN = N_C - CM$, N_C is the number of coldspots). This odds ratio measures the relative risk associated with the presence of a binding motif in hotspots compared to coldspots. Hereafter, we will use the odds ratio O_{hc} to measure the likelihood that a protein is a trans-regulator of recombination hotspots.

C. Finding associated GO terms

Given a gene g, $T(g)$ is the set of GO terms annotating this gene. We define the similarity between a term t and a gene g, $S(t,g)$, in equation 1 and subsequently define the similarity between t and a set of genes G, $S(t,G)$, in equation 2.

$$S(t,g) \quad = \frac{1}{|T(g)|} \sum_{t' \in T(g)} sim(t,t') \qquad (1)$$

$$S(t,G) \quad = \frac{1}{|G|} \sum_{g \in G} S(t,g) \qquad (2)$$

Here, $sim(t,t')$ in equation 1 is the semantic similarity between GO terms t and t' and we applied the method in [24] to calculate $sim(t,t')$.

Let HG and LG denote the sets of genes with high odds ratio and low odds ratio respectively. The scores $S(t,HG)$ and $S(t,LG)$ can be utilized to show t's enrichment in HG and LG respectively. Therefore, their gap with respect to the term t, $gap(t)$ in equation 3, can be used to discriminate t's enrichment in HG and LG. For example, a large gap indicates that t is enriched in the genes with high ratio while not enriched in those with low ratio.

$$gap(t) = \frac{S(t,HG) - S(t,LG)}{\max(S(t,HG), S(t,LG))} \qquad (3)$$

III. Results

A. Experimental data

We downloaded the mouse recombination hotspots from [23]. There are 9874 hotspots in all and the average hotspot width is 3414.08 bases. According to the hotspot boundaries, we extracted their DNA sequences from mouse genome and mouse DNA sequences (version: MGSCv37) were downloaded from NCBI. In addition, the GO data for GO term analysis were downloaded from [1].

Then, as statistical control we selected coldspots that have the following properties. First, coldspots have the same widths as hotspots and each chromosome has the same number of coldspots and hotspots. Second, each coldspot is at least 50kb far away from other hotspots. Third, any two coldspots do not have common sequences, i.e. they are non-overlapping.

B. Enrichment of PRDM9 binding in mouse genome

First, we applied our method on the PRDM9 protein, which has been recently discovered as a trans-acting regulator of meiotic recombination hotspots, and is under intense research. The binding sequence of mouse PRDM9 was a 33-mer obtained from [16]. A matrix representing the degeneracy of mouse PRDM9 binding motif was fed into FIMO to search in the DNA sequences of hotspots and coldspots. To get more reliable estimation on coldspots, we randomly selected and searched by FIMO on coldspots for 20 times and then counted the average numbers of motif occurrences over the 20 runs.

As shown in Table 1, the binding sites of PRDM9 are more enriched in hotspots than coldspots, as demonstrated by the odds ratio O_{hc} = 1.63 with p-value less than 10^{-4} using χ^2 test with Yates' correction.

TABLE I
Number of occurrences of mouse PRDM9 motif (FIMO, q-value < 0.05) on hotspots and coldspots.

	# hits	# regions with hit(s)	# regions without hit
Hotspots	4954	1405	8469
Coldspots	3603.55	912	8962

C. Other TFs from JASPAR with binding sites enriched in hotspots

Encouraged by the positive results on PRDM9 obtained using our approach, we analyzed other mouse transcription factors (TFs) from JASPAR database [21], in hope of identifying proteins with enriched binding sites in hotspots vs. coldspots. From JASPAR database, we downloaded the degeneracy matrices of 53 TFs, which are input to FIMO to search for hits in mouse hotspots and coldspots.

TABLE II
The numbers of FIMO hits of mouse TFs from JASPAR database (with top 12 O_{hc}, ranked by the p-values of the odds ratios).

HG Genes	HM	HN	CM	CN	O_{hc}	p-value
KLF4	886	8988	643.15	9230.85	1.415	< 0.0001
ZFX	437	9437	329	9545	1.343	< 0.0001
CTCF	1002	8872	769.9	9104.1	1.336	< 0.0001
RXRA	1015	8859	819.55	9054.45	1.266	< 0.0001
ESRRB	792	9082	663.4	9210.6	1.211	0.0002
GABPA	110	9764	67.05	9806.95	1.648	0.0006
MYCN	197	9677	137.45	9736.55	1.440	0.0006
SPZ1	322	9552	249.55	9624.45	1.300	0.0013
MYC	129	9745	91	9783	1.423	0.0061
PAX5	194	9680	151.35	9722.65	1.287	0.0113
EGR1	85	9789	61.55	9812.45	1.384	0.0343
T	189	9685	155.95	9718.05	1.216	0.0411

Table II shows the numbers of FIMO hits for 12 mouse TFs whose odds ratio scores O_{hc} are larger than 1.2. The last column shows the p-values [1] of the odd ratios using χ^2 test with Yates' correction. All the odds ratios in this table have p-values less than 0.05, indicating that they all are statistically significant. In addition, out of all 21 TFs with odds ratios larger than 1, 17 TFs have an statistically significant odds ratio with p-values less than 0.05. For later GO analysis, the set HG consists of 12 TFs in this table and LG consists of 24 TFs with O_{hc} smaller than 0.9.

Note that the odds ratio of PRDM9 is 1.63 as shown in Table I and higher than other mouse TFs (except for GABPA, which however covers much less hotspots than PRDM9), which confirms the recent discovery that PRDM9 is a major trans-regulator of recombination hotspots [4], [16], [18].

D. GO term analysis

Our GO analysis shows that genes with high preference of binding on hotspots are enriched with epigenetic functions. As shown in Table III, out of the top 10 GO terms measured by the gap scores, 8 GO terms are directly related with epigenetic regulation (e.g. histone modification, DNA methylation, chromatin modification). The top 3 are all histone modification (methylation and acetylation). Interestingly, the top 13^{th} term (GO:0007283) is spermatogenesis, which suggests that predicted candidate genes are related with the generation of male gamete, confirming the key role of meiotic recombination hotspots in sexual reproduction. Another GO term of interest (GO: 0032204, not shown in the Table III) is ranked 16, namely regulation of telomere maintenance, which suggests a link with chromosome organization. Note that as the gap score becomes smaller down the list, the proportion of epigenetic terms becomes lower, and none of the 10 GO terms with the lowest gap scores is related with epigenetics. Therefore, the gap scores of epigenetic GO terms are associated with the odds ratios of candidate genes indicating preference of binding on hotspots. The functional connection with epigenetics observed here is consistent with the discovery of PRDM9, which is itself a histone methyltransferase. Indeed, much attention has been paid to epigenetic regulatory mechanisms of recombination hotspots (see the review [3] and references therein). Our approach and results in this paper would bring additional insight into the epigenetic control of recombination hotspots.

We next introduce in details several genes with high preference of binding on hotspots which are also annotated with some of these top-ranked GO terms. First, ZFX is a zinc finger X-chromosomal protein and it is annotated with the term GO:0007283 (spermatogenesis). It is reported that ZFX mutation results in small animal size and reduced germ cell number in male and female mice. Second, MYC with the term GO:0032204 (regulation of telomere maintenance) is believed to regulate expression of 15% of all genes through binding on Enhancer Box sequences (E-boxes) and recruiting histone acetyltransferases (HATs). In addition to its role as a classical transcription factor, MYC also functions to regulate global chromatin structure by regulating histone acetylation. Third, CTCF with both the terms GO:0010216 (maintenance of DNA methylation) and GO:0006306 (DNA methylation) is a sequence-specific DNA-binding transcriptional regulator, insulator, and organizer of higher-order chromatin structure. It contains 11 C_2H_2-type zinc fingers and is involved in promoter activation or repression, hormone-responsive gene silencing, methylation-dependent chromatin insulation, and genomic imprinting. Interestingly, the KLF4 gene, which has top odds ratio score and p-value in Table II, does not show epigenetic functions. It is annotated with two GO terms, namely GO:0006355 (regulation of transcription, DNA-dependent) and GO:0045892 (negative regulation of transcription, DNA-dependent) which are ranked 7 and 15 in Table III. It might imply that recombination hotspots are regulated in concert by both epigenetic and DNA-dependent mechanisms.

E. Analysis of hotspot coverage

In this subsection, we aim to analyze how well those top-ranked genes cover current known hotspots.

[1] Since the above PRDM9 binding sequence is quite long (i.e. 33 bases), its occurrences in both hotspots and coldspots have low p-values and q-values. However, the binding sequences in JASPAR are shorter than PRDM9 binding sequence. As such, the p-values of their FIMO hits are higher than the 33-mer motif and most of their q-values are higher than 0.05. For fair comparison, we set the p-value threshold for JASPAR binding sequences as 3.73×10^{-6}, which is the highest p-value for all PRDM9 occurrences with q-value ≤ 0.05.

TABLE III

GO TERMS ENRICHED IN TFS WITH HIGH ODD RATIO WHILE NOT ENRICHED IN THOSE WITH LOW RATIO (WITH TOP 15 *gap* SCORES).

Rank	GO terms	GO term descriptions	*gap*
1	GO:0051573	negative regulation of histone H3-K9 methylation	0.197
2	GO:0031060	regulation of histone methylation	0.182
3	GO:0035065	regulation of histone acetylation	0.175
4	GO:0000122	negative regulation of transcription from RNA polymerase II promoter	0.174
5	GO:0010216	maintenance of DNA methylation	0.170
6	GO:0006306	DNA methylation	0.168
7	GO:0045892	negative regulation of transcription, DNA-dependent	0.160
8	GO:0051574	positive regulation of histone H3-K9 methylation	0.160
9	GO:0016568	chromatin modification	0.160
10	GO:0051571	positive regulation of histone H3-K4 methylation	0.157
11	GO:0006338	chromatin remodeling	0.153
12	GO:0006357	regulation of transcription from RNA polymerase II promoter	0.151
13	GO:0007283	spermatogenesis	0.146
14	GO:0016584	nucleosome positioning	0.142
15	GO:0006355	regulation of transcription, DNA-dependent	0.141

TABLE IV

HOTSPOT COVERAGE.

Genes in HG	HM	$\mid \cap HS$(PRDM9) $\mid$
T	189	24
PAX5	194	47
KLF4	886	177
GABPA	110	23
RXRA	1015	157
MYCN	197	44
SPZ1	322	63
CTCF	1002	163
ESRRB	792	110
ZFX	437	112
MYC	129	29
EGR1	85	21

Given a gene g, $HS(g)$ is the set of hotspots covered by g. As shown in Table I, PRDM9 covers 1405 hotspots, i.e. its HM number = $|HS$(PRDM9)$|$ = 1405. Table IV shows the number of hotspots covered by genes in HG (i.e. the set of genes with high odds ratios) and the number of common hotspots covered by PRDM9 and genes in HG. For example, the gene T covers 189 distinct hotspots and 24 of them are also covered by PRDM9.

For further analysis, we built a hotspot coverage graph $HC = (V, E, w)$ where nodes are PRDM9 and genes in HG and edges show the hotspot coverage similarity between nodes. In particular, $V = \{PRDM9\} \cup HG$ and each pair of nodes has a weight, indicating their hotspot coverage similarity, based on the meet/min coefficient [9] in equation 4.

$$w(g_i, g_j) = \frac{|HS(g_i) \cap HS(g_j)|}{\min\{|HS(g_i)|, |HS(g_i)|\}}, \forall g_i, g_j \in V. \qquad (4)$$

In this hotspot coverage graph, we divide genes into several clusters and genes in the same clusters will have higher hotspot coverage similarity. A simple solution for clustering [7] is as follows. We first set a threshold w_t and filter all the edges with weights lower than w_t. The remaining connected components are considered as gene clusters. In our experiments, we gradually increased the threshold w_t and obtained a hotspot coverage graph with 4 clusters as shown in Figure 2 when w_t = 0.16. Another solution is to apply hierarchical clustering algorithm. When we set the number of clusters we want to obtain as 4, the 4 clusters of hierarchical clustering are exactly the same as those in Figure 2. In the cluster with 10 red TFs including PRDM9 in Figure 2, all the other TFs have edges to PRDM9 with weights at least 0.16, indicating that they all share similar hotspot coverage patterns with PRDM9. Meanwhile, three singleton clusters consist of TFs with the lowest odds ratios as shown in Table II and they have no edges to PRDM9. These two observations confirm once again that PRDM9 is a major trans-regulator for mouse hotspots.

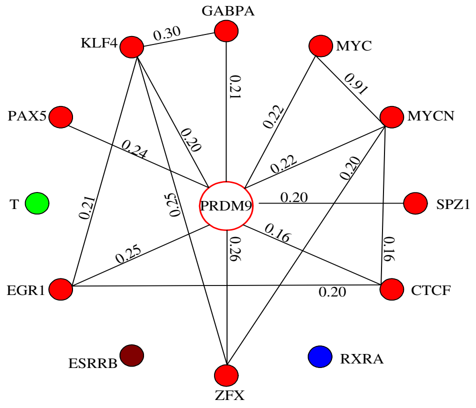

Fig. 2. Hotspot coverage graph. Each color stands for a gene cluster.

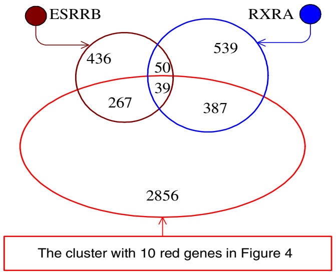

Fig. 3. Venn diagram of the hotspot coverage.

We also show the hotspot coverage of the above 4 clusters in Figure 3. Compared with other clusters, the cluster with only the gene T covers a much smaller number of hotspots and it is thus not shown in Figure 3. In this figure, the red cluster with 10 TFs covers 3549 hotspots which more than double the 1405 hotspots covered by PRDM9 alone and shares 39 hotspots with the other two clusters. In addition, the number of hotspots covered by all the 4 clusters in Figure 2 is 4679. The fact that many hotspots (i.e., 5195 out of 9874 known hotspots) are still not covered by PRDM9 or motifs in our study suggests that JASPAR database is far from complete (currently, there is even no matrix profile for PRDM9 in JASPAR) and we need to search for additional proteins and motifs in the future.

IV. Conclusion

In this paper, we proposed a new approach to discovering trans-regulators of recombination hotspots in mouse genome. Starting from experimentally identified or predicted binding sites of DNA binding proteins, we scan the DNA sequences of hotspots and coldspots for target binding occurrences of each protein. The relative enrichment of binding targets in hotspots is used to estimate the likelihood that a protein has regulatory effect on recombination hotspots. We increased the rigor by designing a GO analysis method to identify shared functions of candidate genes. Applying our method to newly mapped genome-wide mouse recombination hotspots, we confirmed the recent discovery that PRDM9 is a major trans-regulator of recombination hotspots. Further, we identified a list of additional proteins as candidate trans-regulators. GO analysis shows that the most prominent function shared by these candidate genes are histone modification, which confirms and provides new insights into the epigenetic mechanism of recombination hotspots. Thus the approach developed in this paper can be used to identify additional trans-regualtors of hotspots. The predicted proteins and their functional analysis can shed light on the pathways (rather than the single gene of PRDM9) regulating recombination, and be used to guide further experimental studies of recombination hotspots.

Currently, the number of proteins examined in this paper is small compared with the number of transcription factors (i.e. only 53 transcription factors in mouse genome from JASPAR database). In the future, more DNA-binding proteins from more sources (e.g. TRANSFAC [15], ZiFDB [8]) will be examined by our approach for more comprehensive results. Meanwhile, we searched for target binding sites of proteins using FIMO, which does not allow for insertion and deletions in motif matching. However, it is known that the DNA target sites of some proteins contain indels [22]. Therefore, more flexible motif finding algorithms that take into account special sequence patterns (e.g. nucleotide adjacent dependency [5]) may be used to address this problem. Although the recombination hotspots analyzed in this paper was obtained experimentally, our approach is not limited to this type of data and we can computationally estimate recombination rates from sequence polymorphism data in large scale, either based on LD structure [2], [13] or pedigree structure [14]. In addition, as our results of GO analysis suggested that epigenetic mechanism is shared by top candidate genes, we will follow up with studies of epigenetic interaction between histone and DNA as mediated by PRDM9 and other predicted proteins.

Acknowledgement

We would like to thank Pavel P. Khil from Genetics and Biochemistry Branch, NIDDK, NIH, USA for helpful discussions and providing motif information in [23]. This research has been supported by Startup Grant M58020030 from Nanyang Technological University, Singapore, and Tier 2 AcRF Grant MOE2008-T2-1-074 from Ministry of Education, Singapore. This work was also supported by the Intramural Research Program of NIH, National Library of Medicine, USA.

References

[1] Gene ontology database. http://www.geneontology.org.

[2] A. Auton and G. McVean. Recombination rate estimation in the presence of hotspots. *Genome Res*, (8):1219–1227, 2007.

[3] P. Barthes, J. Buard, and B. de Massy. Epigenetic factors and regulation of meiotic recombination in mammals. *Epigenetics and Human Reproduction*, 2011.

[4] F. Baudat and *et al*. Prdm9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science*, 327(5967):836–840, 2010.

[5] F. Y. L. Chin, H. C. M. Leung, M. H. Siu, and S. M. Yiu. Optimal algorithm for finding dna motifs with nucleotide adjacent dependency. In *APBC*, pages 343–352, 2008.

[6] The International HapMap Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467:52–8, 2010.

[7] J. Dutkowski and J. Tiuryn. Identification of functional modules from conserved ancestral protein-protein interactions. *Bioinformatics*, 23(13):149–158, 2007.

[8] F. Fu and *et al.* Zinc finger database (zifdb): a repository for information on c2h2 zinc fingers and engineered zinc-finger arrays. *Nucleic Acids Research*, 37(Database-Issue):279–283, 2009.

[9] D. Goldberg and F. P. Roth. Assessing experimentally derived interactions in a small world. *PNAS*, 100(8):4372–4376, 2003.

[10] C. E. Grant, T. L. Bailey, and W. S. Noble. Fimo: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018, 2011.

[11] A. J. Jeffreys and R. Neumann. Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. *Nature Genetics*, 31:267–271, 2002.

[12] H. Jiang and *et al*. High recombination rates and hotspots in a plasmodium falciparum genetic cross. *Genome Biology*, 12(4):33, 2011.

[13] J. Li, M. Q. Zhang, and X. Zhang. A new method for detecting human recombination hotspots and its applications to the hapmap encode data. *American Journal of Human Genetics*, (4):628–639, 2006.

[14] X. Li, Y. Chen, and J. Li. Detecting genome-wide haplotype polymorphism by combined use of mendelian constraints and local population structure. In *Pacific Symposium on Biocomputing*, pages 348–358, 2010.

[15] V. Matys and *et al*. Transfac: transcriptional regulation, from patterns to profiles. *Nucleic Acids Research*, 31(1):374–378, 2003.

[16] S. Myers, R. Bowden, A. Tumian, R. E. Bontrop, C. Freeman, T. S. MacFie, G. McVean, and P. Donnelly. Drive against hotspot motifs in primates implicates the prdm9 gene in meiotic recombination. *Science*, 327(5967):876–879, 2010.

[17] S. Myers, C. Freeman, A. Auton, P. Donnelly, and G. McVean. A common sequence motif associated with recombination hot spots and genome instability in humans. *Nature Genetics*, 40(9):1124–1129, 2008.

[18] E. D. Parvanov, P. M. Petkov, and K. Paigen. Prdm9 controls activation of mammalian recombination hotspots. *Science*, 327(5967):835, 2010.

[19] A. V. Persikov, R. Osada, and M. Singh. Predicting dna recognition by cys$_2$his$_2$ zinc finger proteins. *Bioinformatics*, 25(1):22–29, 2009.

[20] M. Pineda-Krch and R. J. Redfield. Persistence and loss of meiotic recombination hotspots. *Genetics*, 169:2319–2333, 2005.

[21] E. Portales-Casamar, S. Thongjuea, and *et al*. Jaspar 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 38(Database-Issue):105–110, 2010.

[22] T. Riley, X. Yu, E. Sontag, and A. Levine. The p53hmm algorithm: using profile hidden markov models to detect p53-responsive genes. *BMC Bioinformatics*, 10, 2009.

[23] F. Smagulova, I. Gregoretti, K. Brick, P. Khil, R. Camerini-Otero, and G. Petukhova. Genome-wide analysis reveals novel molecular features of mouse recombination hotspots. *Nature*, 472(7343):375–378, 2011.

[24] J.Z. Wang, Z. Du, R. Payattakool, P. Yu, and C. Chen. A new method to measure the semantic similarity of go terms. *Bioinformatics*, 23(10):1274–1281, 2007.

[25] W. Winckler and *et al.* Comparison of fine-scale recombination rates in humans and chimpanzees. *Science*, 308:107–111, 2005.

[26] J. Zheng, P. P. Khil, R. D. Camerini-Otero, and T. M. Przytycka. Detecting sequence polymorphisms associated with meiotic recombination hotspots in the human genome. *Genome Biology*, 11(R103):1–15, 2010.