NetPipe: A Network-based Pipeline for Discovery of Genes and Protein Complexes Regulating Meiotic Recombination Hotspots

Min Wu
School of Computer Engineering
Nanyang Technological University
Singapore, 639798
wumin@ntu.edu.sg

Chee Keong Kwoh
School of Computer Engineering
Nanyang Technological University
Singapore, 639798
asckkwoh@ntu.edu.sg

Xiaoli Li
Data Mining Department
Institute for Infocomm Research
1 Fusionopolis Way, #21-01
Connexis, Singapore, 138632
xlli@i2r.a-star.edu.sg

Jie Zheng*
School of Computer Engineering
Nanyang Technological University
Singapore, 639798
*zhengjie@ntu.edu.sg

ABSTRACT

The regulatory mechanism of recombination is one of the most fundamental problems in genomics, with wide applications in genome wide association studies (GWAS), birth-defect diseases, molecular evolution, cancer research, etc. Recombination events cluster into short genomic regions called "recombination hotspots" in mammalian genomes. Recently, a zinc finger protein PRDM9 was reported to regulate recombination hotspots in human and mouse genomes. In addition, a 13-mer motif contained in the binding sites of PRDM9 is also enriched in human hotspots. However, this 13-mer motif only covers a fraction of hotspots, indicating that PRDM9 is not the only regulator of recombination hotspots. Therefore, discovery of other regulators of recombination hotspots becomes a current challenge.

Meanwhile, recombination is a complex process unlikely to be regulated by individual proteins. Rather, multiple proteins need to act in concert as a molecular machinery to carry out the process accurately and stably. As such, the extension of the prediction of individual proteins to protein complexes is also highly desired. In this paper, we propose a network-based pipeline named NetPipe to identify protein complexes associated with meiotic recombination hotspots. Previously, we associated proteins with recombination hotspots using the binding information between these proteins and hotspots. Here, we exploited protein-protein interaction (PPI) data to prioritize many more other proteins without such binding information. Furthermore, we detected protein complexes conserved between human and mouse that are associated with hotspots. Evaluation results show that the top genes ranked in PPI networks have significant relations to recombination related GO terms. In addition, individual genes in the multi-protein complexes detected by NetPipe are enriched with epigenetic functions, providing more insights into the epigenetic regulatory mechanisms of recombination hotspots.

Keywords

Recombination hotspots; epigenetics; network-based pipeline; random walk; PPI data; conserved protein complexes.

1. INTRODUCTION

Recombination is a process that homologous chromosomes exchange their arms and such crossover events tend to occur more frequently within some short regions called "recombination hotspots". Recombination is one of the most fundamental processes in molecular biology and the understanding of the mechanisms for recombination hotspots would thus shed light on various important aspects in molecular biology and medicine, such as genome instability, birth-defect diseases, disease gene mapping, molecular evolution and so on.

Recently, there has been much progress in the discovery of the mechanisms for meiotic recombination hotspots in mammalian genomes. For example, in 2010, a zinc finger protein PRDM9 was reported as a *trans*-regulator of recombination hotspots in human and mouse genomes in three Science papers [2, 11, 15]. PRDM9 binds to DNA and its binding site contains a 13-mer motif previously found to be enriched in human hotspots [12]. In [19], Smagulova et al. analyzed the molecular features of mouse recombination hotspots using Chip-Seq data and observed that a consensus motif enriched in mouse hotspots aligns with the predicted binding site of mouse PRDM9 significantly. Using an LD-based approach named LDsplit, Zheng et al. [31] identified HapMap SNPs (single nucleotide polymorphisms) as *cis*-regulators of recombination hotspots. In addition, the authors [31] also found an enriched 11-mer motif which closely matches the aforementioned 13-mer motif bound by PRDM9 and enriched in human recombination hotspots.

Although significant breakthroughs have been made in the understanding of the regulatory mechanisms of meiotic recombination hotspots, they are mainly focused on the well-known protein PRDM9. However, it is estimated that PRDM9 can explain only 18% of variations in human recombination phenotype [2]. Meanwhile, the 13-mer motif contained in the binding sites of PRDM9 covers only 41% of human hotspots [12]. Therefore, PRDM9 is unlikely to be the only *trans*-regulator of recombination hotspots. To perform its functions, PRDM9 must interact with other proteins to form a protein complex or regulatory pathway. Hence, it is highly motivated to discover other genes and their high-level organizations, such as protein complexes or regulatory pathways, which are also associated with recombination hotspots. Furthermore, the function of PRDM9 for regulating recombination hotspots is well conserved among human, chimpanzee and mouse. It would also be an interesting question in comparative genomics whether there are some other genes and their pathways or complexes whose functional roles in regulating recombination hotspots are also conserved among species.

In our previous study [25], we proposed an implementation, called the Odds-Ratio method, to predict other regulators of recombination hotspots from DNA-binding proteins based on their binding preference to hotspots against coldspots. Our Odds-Ratio method reported a list of candidate *trans*-regulators (including PRDM9) of mouse hotspots and these candidates are enriched with functions of histone modifications, highlighting the epigenetic mechanisms of recombination hotspots. However, the Odds-Ratio method requires the binding motifs of transcription factors (TF) to be known. Given that there are only a limited number of known TFs with binding motifs, Odds-Ratio method is thus less effective to search for more novel genes associated with recombination hotspots.

To address the above issues, this paper proposes a network-based pipeline called NetPipe to identify genes and protein complexes associated with recombination hotspots. NetPipe consists of three stages. First, for each input DNA-binding protein, we estimate a Hotspot-Binding (HB) profile showing its binding preference to hotspots. We then construct a HB network based on the similarity of HB profiles between genes (thereafter, we use terms "gene" and "protein" interchangeably), where a gene is a node and an edge connects two genes with similar HB profiles. We subsequently prioritize conserved genes between human and mouse associated with recombination hotspots by aligning their HB networks. Second, using genes prioritized by HB network alignment as seeds, we apply the Random Walk with Restart algorithm (RWR) to propagate the influences of these seeds to other proteins in protein-protein interaction (PPI) networks. As such, many proteins without known binding motifs will also be assigned a score showing their relationships with recombination hotspots. Third, we construct sub-PPI networks induced by top genes ranked by RWR for both human and mouse and detect conserved protein complexes in those sub-PPI networks, which may perform functions related to recombination hotspots.

In order to evaluate the results of NetPipe, we utilized various kinds of GO term analysis. First, the GO term enrichment analysis (as in our previous study [25]) shows that epigenetic functions are enriched in the seeds prioritized by our HB network alignment. This result is similar to our Odds-Ratio method, demonstrating that our HB network alignment is an effective alternative approach to identifying *trans*-regulators from TFs. Second, we calculated the semantic similarity between identified genes to existing recombination related GO terms (i.e., "DNA recombination" (GO:0006310) and "Meiosis" (GO:0007126)). Genes top-ranked by RWR are demonstrated to have high similarities with these recombination related GO terms. This shows RWR is a credible complement to the existing methods since it enables the detection of those novel genes without HB profiles. Last, different from most existing methods which only explore the individual genes, in this paper we also carried out analysis at protein-complex level which can capture the underlying modularity and functional organization among multiple proteins. Our GO term analysis for conserved complexes based on p-values [8] shows that epigenetic functions are enriched in those complexes, providing more confidence in the epigenetic mechanisms for recombination hotspots.

2. METHODS

In this section, we will introduce the three main steps of our NetPipe in more details.

2.1 HB Network Construction and Alignment

In our previous study [25], we calculated the binding sites of TFs, based on their binding motifs, to DNA sequences in mouse genome using the FIMO software [6]. Those preferring to bind to hotspots rather than coldspots will be predicted as *trans*-regulators of mouse hotspots. In this work, we first collected the hotspot-binding profiles (HB profiles) for those TFs. In particular, we divide the whole genome into λ bins with fixed length (e.g., 5M bases) and the HB profile of a TF g is represented as a $\lambda-$dimension vector, $HB(g) = (b_1, b_2, \cdots, b_\lambda)$, where b_i is the number of hotspots in the i^{th} bin that g binds to. Subsequently, we can build a HB network for TFs, where a node is a TF and an edge between two TFs indicates they have similar HB profiles. The similarity between two HB profiles is measured by Pearson correlation coefficient. Two TFs will be connected in the HB network when the similarity between their HB profiles is larger than a pre-defined threshold (e.g., 0.7 is used in this paper).

We constructed two HB networks for human and mouse, respectively. Then, they are aligned by the network comparison toolkit named NCT [18] to align them. The cross-species alignment of HB networks can detect evolutionarily conserved network motifs associated with recombination hotspots, which should be more significant than signals from single-species analysis. The procedure of NCT for network alignment is as follows. First, to detect the conserved patterns (paths or cliques) between two species, NCT will first build an orthology graph (also called network alignment graph), in which each node represents a pair of proteins with high sequence similarity (homologous or orthology proteins) and each edge represents a conserved interaction between the corresponding protein pairs in both species. Second, a subgraph in the orthology graph can have a likelihood ratio score that indicates its propensity to be conserved [17, 18]. Last, candidate subgraphs with high scores are predicted as conserved patterns by an exhaustive searching heuristic.

It is observed that proteins involved in multiple modules tend to be more biologically important [13]. Therefore, for those TFs in HB networks, we evaluate their relevance to recombination hotspots based on their frequency in the conserved clusters collected by NCT. More specifically, for a TF g, its relevance score $R(g)$ to recombination hotspots is finally measured by its frequency in the conserved clusters, i.e., the number of conserved clusters involving g, normalized by the maximum frequency over all the genes. We use the relevance scores to rank candidate genes related to recombination hotspots.

2.2 Random Walk in PPI Networks

TFs closely related to recombination hotspots (or so-called *trans*-regulators) can be predicted by the above HB network alignment. However, the power of this method, as well as our previous Odds-Ratio method, would be limited due to the small number of TFs with known binding motifs. Out of tens of thousands of known human and mouse genes, there are only 158 binding motifs for human and 148 for mouse in two well-known databases (i.e., JASPAR [16] and TRANSFAC [10]), respectively. Meanwhile, a large amount of protein-protein interaction (PPI) data are available and they are often modeled as graphs, where nodes are proteins and edges are interactions between proteins, for predicting novel protein interactions [30], protein functions [4], protein complexes [8], disease genes [9] etc. In this work, we combine PPI data to evaluate the relevance of genes (proteins) to recombination hotspots, by a Random Walk with Restart algorithm (RWR) [7].

RWR simulates a random walker, which starts on a set of seed nodes and moves to their neighbors randomly at each step. Therefore, RWR propagates the influence from the seed nodes to the remaining nodes in the PPI network and can be used to measure the proximity of other nodes to the seed nodes. Let p_0 be the initial vector showing the relevance of seeds to recombination hotspots (i.e., assigned by the HB network alignment method) and p_t be a vector in which the i-th element shows the relevance of node i at step t. The relevance vector at step $t + 1$ is then calculated as

$$p_{t+1} = (1 - \gamma) \times W \times p_t + \gamma \times p_0, \qquad (1)$$

where W is the transition matrix of the PPI network and each element W_{ij} is the transition probability from node i to node j. The parameter $\gamma \in (0, 1)$ is the restart probability. At each step, the random walker may return to seed nodes with probability γ. We generally use the normalized adjacency matrix as the transition matrix. $p_0(i)$ is assigned with the relevance score $R(g_i)$ output from the HB network alignment in previous subsection if the i^{th} node g_i is a seed, and 0 otherwise. $p_\infty(i)$ is the final relevance of node i to recombination hotspots. We can obtain the relevance vector at the steady state (p_∞) efficiently by performing iterations until the difference between p_{t+1} and p_t is below a threshold, for example, 10^{-10} [9].

Based on the RWR algorithm in PPI networks, genes that are highly interactive with the seed genes will accumulate more influence pumped from the seeds. Hence, we can consider them as novel genes related to recombination hotspots even if they do not have HB profiles.

2.3 Detection of Protein-Complexes Conserved in PPI Networks

After prioritizing genes associated with recombination hotspots, we construct sub-networks for human and mouse, respectively, which are induced by those top-ranked genes (e.g., top 200 genes [21]). Furthermore, we can detect protein complexes highly related to recombination hotspots, which are conserved in both human and mouse. Generally, if two complexes from two species share main components and they are similar enough, we can consider them as conserved although they may have additional distinct members. The NCT algorithm is supposed to be applied here to detect conserved complexes. However, a conserved protein complex collected by NCT often means that two species have exactly the same complex if there is no further post-processing on it (Figure 3 shows an example output of NCT). To address this issue, we thus proposed a simple yet efficient approach to detect the conserved complexes with the following two steps.

First, we used the COACH algorithm [26] to detect protein complexes in human and mouse PPI sub-networks, respectively. Let $H = \{H_1, \cdots, H_m\}$ and $M = \{M_1, \cdots, M_n\}$ be the set of protein complexes predicted by COACH from human and mouse sub-networks respectively. Then, we can build a bipartite graph $G = (H, M, E, w)$, where H and M represent two sets of super-vertices (i.e., each predicted protein complex is considered as a super-node in the bipartite graph G) and the edge weights are defined using the neighborhood affinity (NA) score [26, 3] in Equation 2. Here, $|H_i \cap M_j|$ is the number of ortholog pairs between H_i and M_j. In previous studies [26, 3], two protein complexes with many common proteins, which have a NA score larger than or equal to a threshold (generally set as 0.25), will be considered as the same protein complex. Similarly, a pair of super-nodes (i.e., protein complexes) in our bipartite graph G with an edge weight larger than or equal to the threshold will be considered as a pair of conserved complexes and all the edges with weights lower than the threshold will be removed from G.

$$w(H_i, M_j) = \frac{|H_i \cap M_j|^2}{|H_i| \times |M_j|}. \qquad (2)$$

Second, we will detect conserved protein complexes by matching H and M in G. A matching P in G is a subset of E, where each vertex is involved in no more than one edge in P. The maximum weighted matching P^* has the maximum sum of weights of edges. Here, a modified augmenting-path algorithm [24] is employed to solve the maximum weighted matching problem and finally our conserved protein complexes are those pairs in the maximum weighted matching P^*.

3. RESULTS

Before showing the results of NetPipe, we briefly introduce the data used in our experiments. Recombination hotspots of mouse were downloaded from [19]. Recombination hotspots of human were collected from HapMap genetic map estimated by the LDhat package [1]. There are 9,874 and 39,551 hotspots in mouse and human genomes respectively. DNA sequences for mouse (version: MGSCv37) and human (version: GRCh37) were downloaded from NCBI.

To collect the HB profiles, the binding motifs of TFs were downloaded from JASPAR and TRANSFAC databases. After processing, we obtained 158 human binding motifs and 148 mouse binding motifs respectively. Human PPI data were downloaded from the BioGRID database [20], consisting of 11,120 proteins and 55,014 interactions among these proteins, while mouse PPI data were downloaded from [29], with 10,348 proteins and 63,882 interactions. Lastly, the GO data for various GO term analysis were downloaded from http://www.geneontology.org.

3.1 HB profiles for TF proteins

In our experiments, we set the bin size as 5Mb to divide both human and mouse chromosomes. Correspondingly, we obtained 35 bins and 27 bins for human chromosomes 6 and 11 respectively. Figure 1 shows the distribution of hotspots over all the bins in human chromosomes 6 and 11. We can easily observe that the distributions of the hotspots have dips in the middle of the chromosomes. This is also observed in the other chromosomes (data not shown). The observation is consistent with the fact that recombination hotspots occur more frequently in telomeres than centromeres [14]. Our HB profiles for genes, based on the above binning of chromosomes, would thus be promising to capture some biological insights.

Figure 2 shows the HB profiles for the PRDM9 gene in human chromosomes 6 and 11. We can find that the HB profiles of PRDM9 have similar overall trends as the distributions of hotspots, while also look different in some specific bins. In this work, we mainly compared the HB profiles among different genes and then built the HB networks based on the profile similarities. We will compare HB profiles for various genes with the background hotspot distributions for future studies. For example, if a protein has a HB profile significantly different from the background hotspot distributions, it may thus be inferred to perform functions related to recombination hotspots.

3.2 Genes prioritized by HB network alignment

After collecting HB profiles for TFs and building HB netowrks, NCT will generate clusters conserved in HB networks. Figure 3 shows a conserved cluster predicted by NCT. As previously introduced, NCT will assign a score to each output conserved cluster. The cluster here in Figure 3 with 15 TFs is the one with the highest score.

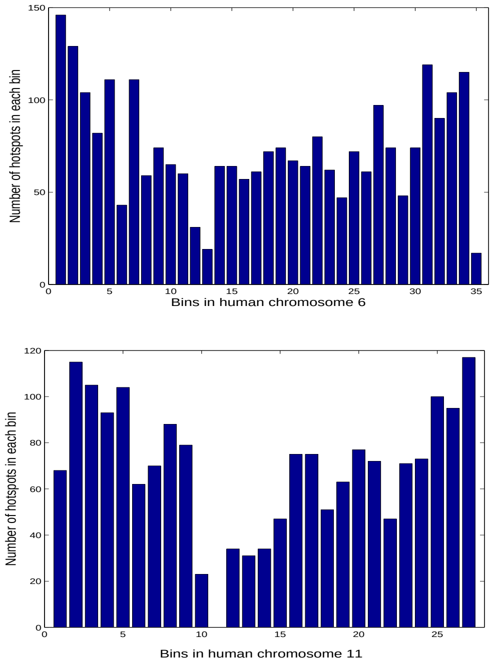


Figure 1: The number of hotspots in each bin of human chromosomes 6 and 11.

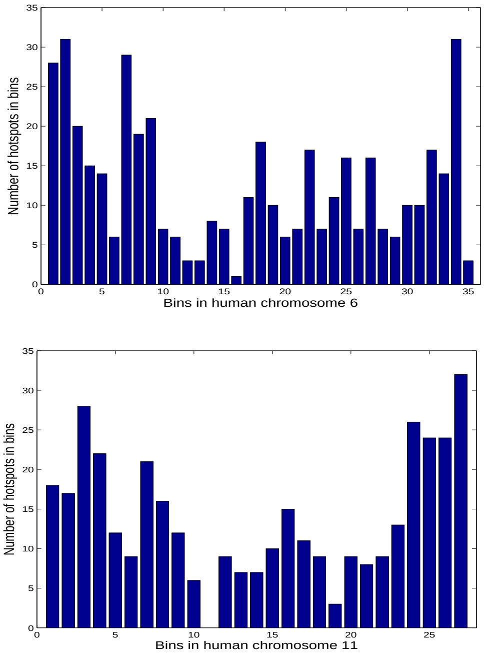


Figure 2: The number of hotspots binding to PRDM9 in each bin of human chromosomes 6 and 11.

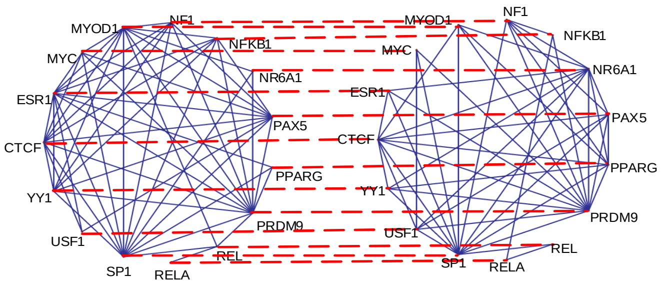
TFs are now ranked with respect to their relevance scores that are computed based on all the conserved clusters collected by NCT. In our experiments, we selected 15 TFs with the highest relevance scores as "seeds" for the subsequent RWR algorithm in PPI networks. These seeds are SP1, PRDM9, PAX5, ESR1, CTCF, NF1, NR6A1, MYOD1, YY1, USF1, PPARG, NFKB1, MYC, RELA and REL in the decreasing order of their relevance scores. Note that 13 out of these 15 seeds are in the cluster in Figure 3. With these seeds,we will next show the results in each step of NetPipe.

First, we utilized the GO term analysis as in our previous study [25] for these seeds. Table 1 shows top-10 GO terms enriched in these human seeds (results for mouse seeds are similar and thus are not shown here). The *gap* score of a GO term (in the 4^{th} column of Table 1) shows the enrichment of this GO term in a given set of genes. In Table 1, top three terms are quite interesting, namely GO:0007283 (spermatogenesis), GO:0007276 (gamete generation) and GO:0019953 (sexual reproduction). As we know, meiotic recombination hotspots play key roles in sexual reproduction. Our seeds enriched with functions highly related to "sexual reproduction", may perform their functions in the regulation of recombination hotspots. In addition, other top ranked terms are all epigenetic functions, indicating the conserved epigenetic mechanism for recombination hotspots across human and mouse species.

Table 1: GO terms enriched in human seeds with top-10 *gap* scores

Rank	GO terms	GO term descriptions	*gap*
1	GO:0007283	spermatogenesis	0.722
2	GO:0007276	gamete generation	0.487
3	GO:0019953	sexual reproduction	0.322
4	GO:0051573	negative regulation of H3-K9 methylation	0.302
5	GO:0016573	histone acetylation	0.284
6	GO:0051574	positive regulation of H3-K9 methylation	0.284
7	GO:0051571	positive regulation of H3-K4 methylation	0.277
8	GO:0031060	regulation of histone methylation	0.268
9	GO:0016568	chromatin modification	0.263
10	GO:0006338	chromatin remodeling	0.256

Second, we generated random seeds from all the input DNA-binding TFs as the input of RWR in our NetPipe and then collected their results, i.e., genes ranked by RWR algorithm and conserved protein complexes. In contrast to the enrichment of epigenetic terms in Table 1, there are no epigenetic functions enriched in the random seeds as shown in Table 3. It suggests that epigenetic functions are enriched in our seeds while not enriched in the whole set of TFs.



(A) Human

(B) Mouse

Figure 3: A conserved cluster predicted by NCT from HB networks.

Third, we computed the semantic similarity between our seeds and two manually-collected terms (i.e., "DNA recombination" (GO:0006310) and "Meiosis" (GO:0007126)) in Equation 3. These two terms are highly related to meiotic recombination hotspots. In Equation 3, $T(g)$ is the set of GO terms annotating a gene g, $S(t,g)$ is the similarity between a GO term t and a gene g and $S(t,V)$ is the similarity between t and a gene set V. In addition, $sim(t,t')$ is the semantic similarity between GO terms t and t' and we calculated it using the method in [23].

$$\begin{aligned} S(t,g) &= \max_{t'\in T(g)} sim(t,t') \\ S(t,V) &= \frac{1}{|V|}\sum_{g\in V} S(t,g). \end{aligned} \tag{3}$$

Table 2 shows the semantic similarity for random seeds, all the TFs with binding motifs and the whole set of human genes for comparison (the results for mouse are similar and they are not shown here). It is observed that the seeds from HB network alignment have much higher similarity to these two terms than other sets of genes, indicating that the HB network alignment method is indeed of help for selecting genes associated with recombination hotspots.

Table 2: Semantic similarity between human seeds and two recombination related terms

	DNA recombination	Meiosis	Average
Seeds	0.568	0.322	0.445
Random seeds	0.519	0.215	0.367
158 TFs	0.518	0.215	0.367
All human genes	0.272	0.161	0.216

Last, the above results and analysis show that the seeds prioritized by our HB network alignment method are very promising and biologically significant. In fact, our HB network alignment method can identify some genes highly related to recombination hotspots, which cannot be detected by our previous Odds-Ratio method [25]. For example, the gene YY1 has a low odds ratio score in [25], while it can be identified by our HB network alignment method. It is a core component of the chromatin remodeling INO80 complex which is involved in transcriptional regulation, DNA replication and DNA repair. It is annotated with the terms GO:0006310 (DNA recombination) and GO:0000724 (double-strand break repair via homologous recombination) [28] and is involved in recombination events by binding to DNA recombination intermediate structures [27].

3.3 Genes re-ranked by RWR

In the above subsection, we show that seeds are enriched with epigenetic terms and have high similarity with two recombination-related GO terms. Here, we focus on analyzing those novel non-seed genes top-ranked by the RWR algorithm. Table 4 shows top non-seed genes in the PPI network ranked by the RWR algorithm and their semantic similarity to terms "DNA recombination" (GO:0006310) and "Meiosis" (GO:0007126). We observed that those top-ranked human non-seeds have a higher similarity with these two terms than the seeds themselves (0.480 v.s. 0.445). Similarly, top-ranked human genes from random seeds also have a

Table 3: GO terms enriched in random human seeds with top-10 average *gap* scores (over 100 random sets of seeds)

Rank	GO terms	GO term descriptions	Average *gap*
1	GO:0015695	organic cation transport	0.0326
2	GO:0048241	epinephrine transport	0.032
3	GO:0055085	transmembrane transport	0.0276
4	GO:0010248	establishment and/or maintenance of transmembrane electrochemical gradient	0.0276
5	GO:0000301	retrograde transport, vesicle recycling within Golgi	0.026
6	GO:0006891	intra-Golgi vesicle-mediated transport	0.0258
7	GO:0015909	long-chain fatty acid transport	0.0255
8	GO:0042953	lipoprotein transport	0.0254
9	GO:0015908	fatty acid transport	0.0254
10	GO:0046323	glucose import	0.0251

Table 4: Top human and mouse genes ranked by RWR algorithm and their similarity to two recombination related terms.

	Human		Mouse	
Rank	Genes	Similarity	Genes	Similarity
1	UIMC1	0.464	CREBBP	0.504
2	UBC	0.647	SMAD3	0.438
3	EP300	0.463	EP300	0.303
4	HDAC1	0.446	RB1	0.439
5	SMARCA4	0.417	SMAD4	0.457
6	SMAD3	0.438	TBP	0.37
7	CREBBP	0.504	GTF2I	0.37
8	POLR2A	0.425	LOC637733	0
9	KPNA2	0.712	HDAC1	0.446
10	SMAD2	0.432	BRCA1	0.746
11	TP53	0.575	HSPA8	0.509
12	RUNX1	0.395	POLR2A	0.384
13	SMAD4	0.393	THRB	0.375
14	ID3	0.478	YWHAB	0.266
15	DAXX	0.41	ID2	0.361
	AVERAGE	0.480	AVERAGE	0.398

higher similarity than these random seeds (0.391 v.s. 0.367). This observation implies the usefulness of PPI data for us to find and study those individual proteins related to recombination hotspots. In addition, human genes generated by real seeds in Table 4 have significantly higher similarities to recombination-related GO terms than those by random seeds, once again demonstrating that our seeds generated by HB network alignment are biologically meaningful. We also observed that some genes (e.g., UBC and HDAC1) are both top-ranked using real seeds and random seeds. It is reasonable that some hub genes (i.e., with many interacting partners) will accumulate influence using either real seeds or random seeds. As such, how to estimate and normalize such bias of network properties (e.g., degree) would be investigated in future.

Here, we briefly show some proteins top-ranked by RWR, which may play important roles in recombination hotspots. Human UIMC1 is ranked as the top 1 non-seed gene and it is a component of BRCA1-A complex [22]. It has annotations including GO:0006302 (double-strand break repair), GO:0016568 (chromatin modification) and GO:0045739 (positive regulation of DNA repair). HDAC1 in both human and mouse is a component of the histone deacetylase complex and it is annotated with GO terms like GO:0006338 ("chromatin remodeling") and GO:0006476 ("protein deacetylation"). Interestingly, human KPNA2 is also captured by RWR algorithm. It was previously reported to be involved in recombination, with a GO annotation GO:0000018 ("regulation of DNA recombination") [5].

In summary, many important genes associated with recombination hotspots are identified by RWR in PPI networks. Currently, protein interaction data for various species are still incomplete and noisy. In BioGRID database, no interacting partners can be found for the PRDM9 protein in human or mouse. However, we believe that PPI data will provide more insights when they are further enriched.

3.4 Conserved complexes between human and mouse

We exploit the Gene Ontology (GO) to evaluate the functional enrichment of our conserved protein complexes based on p-values [8]. A predicted protein complex with low p-values indicates that it is enriched by proteins from the same functional group and it is thus statistically significant. In our experiments, 14 out of 15 conserved complexes have the lowest p-values among all the GO terms, which are smaller than 0.001. This result shows that those conserved complexes are indeed enriched by a common, specific function, demonstrating the ability of our NetPipe for predicting protein complexes.

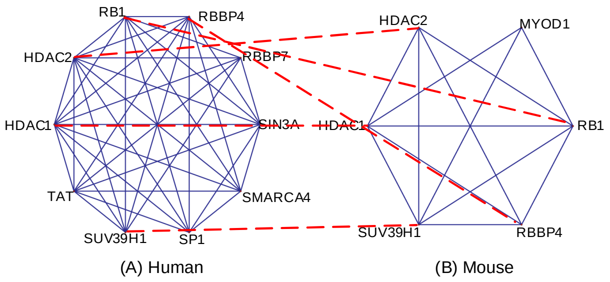


Figure 4: Conserved complexes predicted by our NetPipe.

Table 5: Top-5 GO terms for the conserved complex predicted by NetPipe as shown in Figure 4.

	Human		
Rank	P-value	GO term	Term description
1	6.53e-012	GO:0016580	Sin3 complex
2	5.65e-011	GO:0016581	NuRD complex
3	7.08e-009	GO:0035098	ESC/E(Z) complex
4	4.71e-008	GO:0000792	heterochromatin
5	5.38e-007	GO:0005654	nucleoplasm
	Mouse		
1	1.69e-009	GO:0016580	Sin3 complex
2	7.86e-009	GO:0016581	NuRD complex
3	4.47e-007	GO:0005654	nucleoplasm
4	2.99e-006	GO:0035098	ESC/E(Z) complex
5	9.98e-006	GO:0000792	heterochromatin

More specifically for the conserved complexes in human and mouse, we list top 5 GO terms with the lowest p-values for them. Figure 4 shows a conserved complex predicted by NetPipe. It is also very interesting that epigenetic functions are enriched in these conserved complexes predicted by our NetPipe. For instance, Table 5 shows top 5 "cellular component" GO terms of the conserved complexes in Figure 4. Here, the Sin3 complex (GO:0016580) is a transcriptional repressor of protein-coding genes, through the gene-specific deacetylation of histones. The NuRD complex (GO:0016581) has ATP-dependent chromatin remodeling activity in addition to histone deacetylase (HDAC) activity. The ESC/E(Z) complex (GO:0035098) methylates lysine-27 and lysine-9 residues of histone H3.

Next, we focus on 3 specific histone-related "cellular component" terms, namely, GO:0000118 (histone acetyltransferase complex), GO:0000123 (histone deacetylase complex) and GO:0035097 (histone methyltransferase complex). Table 6 shows the number of complexes whose top-5 GO terms include these 3 histone-related terms or their descendent terms. For example, Sin3 complex and NuRD complex are two main components of histone deacetylase complex. Their corresponding terms GO:0016580 and GO:0016581 are descendants of the term GO:0000123 (histone deacetylase complex) in Gene Ontology.

In this work, a conserved complex refers to a pair of human and mouse complexes, which may be slightly different in their protein components but still similar enough overall. NetPipe predicted 15 conserved complexes from real seeds (i.e., 15 pairs of human and mouse complexes), out of which 6 human complexes and 7 mouse complexes are enriched with the term GO:0000118 (histone acetyltransferase complex) and 4 pairs are both enriched with this term. However, the number of conserved complexes predicted from random seeds, whose top-5 GO terms include these histone-related terms, is much smaller as shown in Table 6. This indicates that epigenetic functions enriched in the conserved complexes detected NetPipe from real seeds is not merely by chance. In our previous study [25], we found that proteins preferring to bind to hotspots are enriched with epigenetic function and we thus shed light on the epigenetic mechanism of recombination hotspots. In this paper, protein complexes with epigenetic functions, which are formed by proteins highly related to recombination hotspots, would be a complementary evidence for the above hypothesis.

4. DISCUSSION AND CONCLUSIONS

In this paper, we proposed a network-based pipeline NetPipe to identify genes and protein complexes associated with recombination hotspots. By using protein interaction data, we can prioritize many more proteins without binding information, which can address the limitation that our previous Odds-Ratio method can only work for a small number of TFs with binding motifs available. Meanwhile, we also detected protein complexes conserved in human and mouse that are associated with hotspots. As far as we know, this is the first work to study the protein complexes conserved for recombination hotspots. Evaluation results show the effectiveness of our NetPipe. Novel genes ranked in PPI networks have high similarity to recombination related GO terms, showing PPI data are indeed a good source to select individual genes associated with recombination hotspots. For example, human protein KPNA2 is also captured by RWR algorithm. It was previously reported to be involved in recombination, with a GO annotation GO:0000018 (regulation of DNA recombination) [5]. In addition, individual genes and protein complexes detected by our NetPipe are enriched with epigenetic functions, providing more insights into the epigenetic regulatory mechanisms of recombination hotspots.

In the future, we will work on the following two directions to extend our current study. First, PRDM9 with no records in the current PPI databases would possibly be due to the incompleteness of the databases themselves. As such, we will take the reliability of PPI data into consideration, i.e., adding novel false negative interactions and eliminating false positives. It is reasonably expected that PPI data with higher quality will provide more accurate prioritization for genes associated with hotspots. Second, we will look for experimental evidence reported in literature or even wet-lab experiments to support our computational predictions. For example, a candidate gene in mouse can be knocked out and it will be verified to be highly associated with recombination hotspots if the recombination rates of many hotspots vary much after its knock-out.

5. REFERENCES

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

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

[3] H. N. Chua, K. Ning, W.-K. Sung, H. W. Leong, and L. Wong. Using indirect protein-protein interactions for protein complex prediction. *J. Bioinformatics and Computational Biology*, 6(3):435–466, 2008.

[4] H. N. Chua, W.-K. Sung, and L. Wong. Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics*, 22(13):1623–1630, 2006.

[5] C. A. Cuomo, S. A. Kirch, J. Gyuris, R. Brent, and M. A. Oettinger. Rch1, a protein that specifically interacts with the rag-1 recombination-activating protein. *PNAS*, 91(13):6156–6160, 1994.

Table 6: Histone related terms enriched in conserved complexes.

	Real seeds			Randomly generated seeds		
# Complexes	15			14.0 (over 100 runs)		
	Human	Mouse	Conserved in Both	Human	Mouse	Conserved in Both
histone acetyltransferase	6	7	4	3.76	1.91	0.65
histone deacetylase	4	3	3	2.14	0.91	0.44
histone methyltransferase	2	1	1	0.1	0.1	0.05

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

[7] S. Köler, S. Bauer, D. Horn, and P. N. Robinson. Walking the interactome for prioritization of candidate disease genes. *American Journal of Human Genetics*, 82(4):949–958, 2008.

[8] X. L. Li, M. Wu, C. K. Kwoh, and S. K. Ng. Computational approaches for detecting protein complexes from protein interaction networks: A survey. *BMC Genomics*, 11(S1):S3, 2010.

[9] Y. Li and J. C. Patra. Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. *Bioinformatics*, 26(9):1219–1224, 2010.

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

[11] 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.

[12] 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.

[13] S. Navlakha and C. Kingsford. Exploring biological network dynamics with ensembles of graph partitions. In *Pacific Symposium on Biocomputing*, pages 166–177, 2010.

[14] K. Paigen and P. Petkov. Mammalian recombination hot spots: properties, control and evolution. *Nature Reviews Genetics*, 11:221–233, 2010.

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

[16] 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.

[17] R. Sharan, T. Ideker, B. P. Kelley, R. Shamir, and R. M. Karp. Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data. In *RECOMB*, pages 282–289, 2004.

[18] R. Sharan, S. Suthram, R. M. Kelley, T. Kuhn, S. McCuine, P. Uetz, T. Sittler, R. M. Karp, and T. Ideker. Conserved patterns of protein interaction in multiple species. *PNAS*, 102(6):1974–1979, 2005.

[19] 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.

[20] C. Stark, B.-J. Breitkreutz, A. Chatr-aryamontri, L. Boucher, R. Oughtred, M. S. Livstone, J. Nixon, K. V. Auken, X. Wang, X. Shi, T. Reguly, J. M. Rust, A. G. Winter, K. Dolinski, and M. Tyers. The biogrid interaction database: 2011 update. *Nucleic Acids Research*, 39(Database-Issue):698–704, 2011.

[21] O. Vanunu, O. Magger, E. Ruppin, T. Shlomi, and R. Sharan. Associating genes and protein complexes with disease via network propagation. *PLoS Computational Biology*, 6(1), 2010.

[22] B. Wang, S. Matsuoka, B. A. Ballif, D. Zhang, A. Smogorzewska, S. P. Gygi, and S. J. Elledge. Abraxas and rap80 form a brca1 protein complex required for the dna damage response. *Science*, 316(5828):1194–1198, 2007.

[23] J. 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.

[24] D. B. West. *Introduction to Graph Theory.* Prentice Hall, Upper Saddle River, NJ, USA, 2nd edition, 2001.

[25] M. Wu, C. K. Kwoh, T. M. Przytycka, J. Li, and J. Zheng. Prediction of trans-regulators of recombination hotspots in mouse genome. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 57–62, 2011.

[26] M. Wu, X. Li, C. K. Kwoh, and S.-K. Ng. A core-attachment based method to detect protein complexes in ppi networks. *BMC Bioinformatics*, 10, 2009.

[27] S. Wu, Y. C. Hu, H. Liu, and Y. Shi. Loss of yy1 impacts the heterochromatic state and meiotic double-strand breaks during mouse spermatogenesis. *Mol. Cell. Biol.*, 29(23):6245–56, 2009.

[28] S. Wu, Y. Shi, P. Mulligan, F. Gay, J. Landry, H. Liu, J. Lu, H. H. Qi, W. Wang, J. A. Nickoloff, C. Wu, and Y. Shi. A yy1-ino80 complex regulates genomic stability through homologous recombination-based repair. *Nat Struct Mol Biol*, 14(12):1165–1172, 2007.

[29] S. Yellaboina, D. Dudekula, and M. Ko. Prediction of evolutionarily conserved interologs in mus musculus. *BMC Genomics*, 9(1):465, 2008.

[30] H. Yu, A. Paccanaro, V. Trifonov, and M. Gerstein. Predicting interactions in protein networks by completing defective cliques. *Bioinformatics*, 22(7):823–829, 2006.

[31] 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.