Predicting DNA Sequence Motifs of Recombination Hotspots by Integrative Visualization and Analysis

Peng Yang[1], Min Wu[1], Chee Keong Kwoh[1], Pavel P. Khil[2], R. Daniel Camerini-Otero[2], Teresa M. Przytycka[3], and Jie Zheng[1,*]

[1]Bioinformatics Research Centre (BIRC), School of Computer Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798

[2]NIDDK, National Institutes of Health, 5 Memorial Drive, Bethesda, Maryland 20892, USA

[3]NCBI, NLM, National Institutes of Health, 8600 Rockville Pike, Bethesda, Maryland 20894, USA

[*]E-mail: zhengjie@ntu.edu.sg

Summary

Meiotic recombination hotspots play important roles in life sciences, but the regulatory mechanism remain unclear. To predict DNA sequence motifs that regulate recombination hotspots, we designed an open source software tool called "LDsplit" that detects SNPs (single nucleotide polymorphisms) associated with meiotic recombination hotspots. The association is measured by the difference of historical recombination rates at a hotspot between two sub-populations with different alleles of a candidate SNP. Providing user-friendly graphic interface and integrative visualization, LDsplit provides insight into the relation of recombination hotspots with proximal DNA sequence patterns.

Availability: LDsplit package (with Java source code, test data and user manual) is freely available for download (under GPL license) at web site http://www.ntu.edu.sg/home/zhengjie/software/LDsplit.htm

1 Introduction

Meiotic recombination is one of the most fundamental cellular processes. It plays important roles in chromosome integrity, genetic inheritance and evolution, *etc*. Recombination events tend to cluster into short genomic regions called "hotspots". The distribution of hotspots offers insight into linkage disequilibrium (LD) block structure, which provides a foundation for genome wide association studies (GWAS). Meiotic recombination events are required for correct segregation of homologous chromosomes during meiosis (*i.e.* generation of sex cells), thus abnormality in recombination procedure will cause aneuploidy and birth defect diseases such as Down syndrome. Despite its importance, however, the regulatory mechanisms of recombination hotspots remain unclear (*e.g.* why hotspots occur at some particular positions).

Recombination hotspots are inheritable and there exist variations in recombination rates among individuals. Thus, they can be studied as phenotypes being associated with genotypes. However, current technologies that measure recombination rates is either limited to small regions (*e.g.* sperm typing) or with low resolution (*e.g.* pedigree based methods). By contrast, LD-based methods inferring historical recombination rate from SNP data, *e.g.* LDhat [1], can detect recombination hotspots with high resolution and on a genome-wide scale. However, the LD-based methods cannot directly measure recombination rates of individuals.

The recent discovery of PRDM9 as a major *trans*-regulator of meiotic recombination hotspots provides a promising direction to elucidate the regulatory mechanisms [2] [3] [4]. PRDM9 is a zinc finger protein that binds to DNA sequences, initiating double strand breaks (DSB). Importantly, the 13-mer binding motif of PRDM9 in human genome, CCNCCNTNNCCNC, is highly enriched in hotspots [5]. This motif was extended from two motifs, CCTCCCT and CCCCACCCC, which were previously found to be enriched in human hotspots [6]. These studies implied that the binding affinity of proteins on these DNA motifs have a significant impact on the activities of hotspots. Therefore, a DNA mutation that breaks the form of the motifs will reduce the protein-DNA binding affinity, and thus suppress the recombination hotspots. This prediction is consistent with a sperm typing experiment on the DNA2 hotspot in MHC (major histocompatibility complex) class II region in human chromosome 6 [7]. In this example, the FG11 SNP is at the third position (T base) of the aforementioned CC**T**CCCT motif within the DNA2 hotspot. Observed by sperm typing, males with the T allele at FG11 SNP, which keeps the original 7-mer motif has a recombination rate 20-fold higher than males with the mutated A allele.

Based on the above observations, an approach to the identification of regulatory motifs for recombination hotspots can be designed as follows. If we can find a SNP with alleles corresponding to significantly different recombination rates of a hotspot, then this SNP is likely to be in a regulatory motif, in which the SNP allele with a lower recombination rate breaks the motif form. Note that it is also possible that a DNA mutation created a new hotspot (*i.e.* increase the recombination rate) in evolutionary history. While sperm typing method can confirm such relation, it is limited to very short DNA regions. It is highly desirable yet challenging to identify SNP-hotspot pairs like the case of FG11-DNA2 on a large scale. To meet this challenge, we developed an algorithm (see Figure 1) in our previous work [8], using which we were able to confirm the sperm typing case of FG11-DNA2 and also identified a sequence motif closely matching the aforementioned 13-mer binding motif of PRDM9 [2]. In this paper, to apply the algorithm to the large-scale discovery of *cis*-regulatory motif of recombination hotspots, we developed the algorithm into an open source software tool named LDsplit. Importantly, we added a user-friendly graphic interface (GUI) as well as other features to allow integrative and interactive analysis of recombination hotspots.

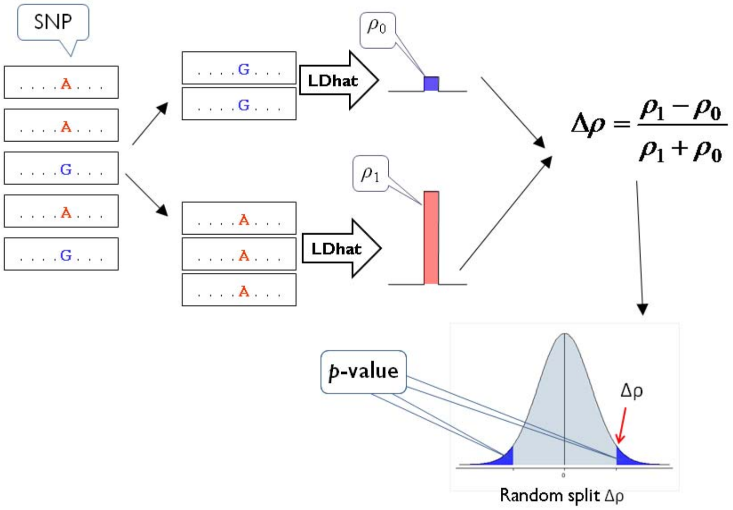


Figure 1: Workflow chart of the LDsplit algorithm.

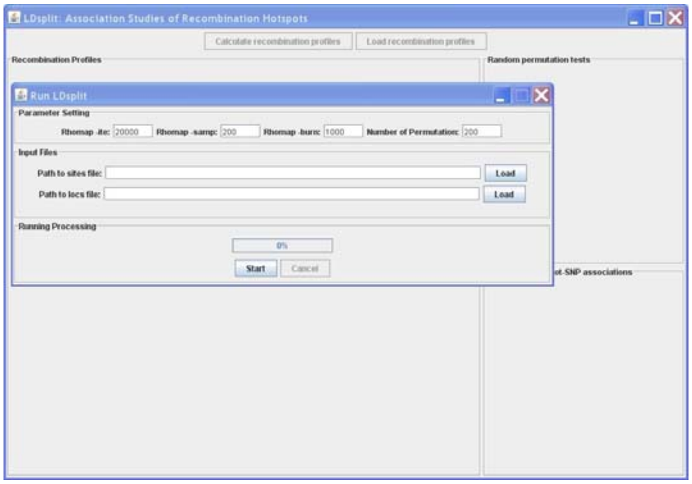
2 Methods

The rationale of LDsplit is that, if a SNP is associated with a recombination hotspot, then between two alleles of the SNP the strengths of the hotspot are likely to be different. It is based on the assumption that historical recombination rate estimated by LD-based computational methods can approximate the extant recombination rate accurately. Based on this rationale, the LDsplit algorithm is designed as follows (Figure 1). For each candidate SNP, LDsplit first divides the population of chromosomes into two subpopulations by SNP alleles (*i.e.* all chromosomes in one subpopulation have the same allele); then it calls LDhat to estimate the recombination rates for each subpopulation, and calculates the normalized difference ($\Delta\rho$) of hotspot strengths between the SNP alleles as $(\rho_0 - \rho_1)/(\rho_0 + \rho_1)$, where ρ_0, ρ_1 denote the strengths of the hotspot in two different subpopulations; the *p*-value of association is estimated by comparing the observed $\Delta\rho$ with the null distribution of random $\Delta\rho$ simulated by permutation tests (*i.e.* each time randomly split the population into two pseudo-populations to calculate a random $\Delta\rho$ value).

The implementation of LDsplit consists of two stages. First, LDhat is run to calculate the recombination profiles for subpopulations split by SNP alleles and pseudo-population in permutation tests. Second, we estimate the *p*-values of hotspot-SNP associations and visualize the recombination profiles of sub-populations. Next, we will introduce our implementation of LDsplit in more details. Users can also look for these details in the user manual, which is available on the website of LDsplit.

2.1 Calculating recombination profiles

We directly apply LDhat to calculate recombination profiles for a window consisting of sequences of SNPs (or haplotypes). There are three types of recombination profiles: (1) the profile of the whole input population of haplotypes; (2) profiles of sub-populations of haplotypes each corresponding to an allele of a candidate SNP (i.e. for each SNP, it splits the population into two sub-populations according to the two alleles of the SNP); (3) profiles of pseudo-populations from a *random* split of the input population. Meanwhile, the input SNP data in this stage consist of two text files, namely *sites* file (consisting of haplotypes in FASTA format) and *locs* file (consisting of physical locations of the SNPs on the chromosomes) as shown in Figure 2. Both of these two files can be extracted from HapMap SNP data by cutting a window of haplotypes. In addition, the user set the parameters involved in this stage (guides provided in the user manual). Because LDhat is computationally costly, this stage usually takes a long time (*e.g.* a few hours for 180 haplotypes each of 200 SNPs). The output of this stage can be exported to hard disk as a Java serialization file, which can be loaded back into LDsplit for analysis in the next stage.


LDsplit: Association Studies of Recombination Hotspots
Calculate recombination profiles
Load recombination profiles
Recombination Profiles
Random permutation tests
ot-SNP associations
Run LDsplit
Load
Load
Start
Cancel

Parameter Setting

Rhomap -ite: 20000 Rhomap -samp: 200 Rhomap -burn: 1000 Number of Permutation: 200

Input Files

Path to sites file:

Path to locs file:

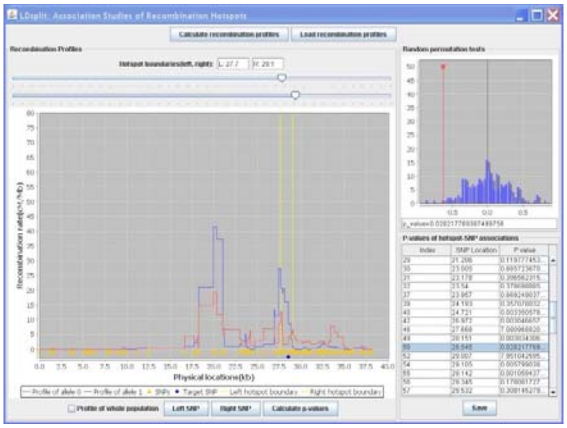
Running Processing

0%

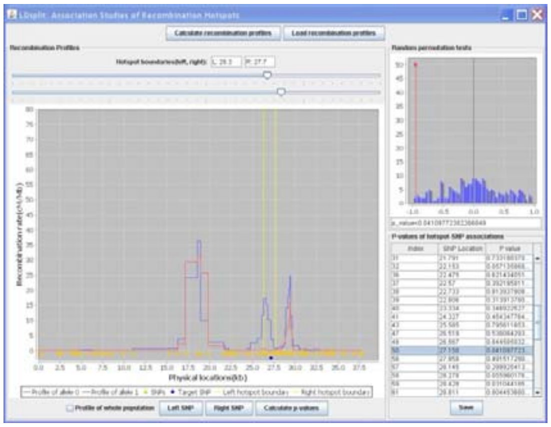
Figure 2. The panel for importing input data and setting parameters to run LDsplit

2.2 Deriving hotspot-SNP associations

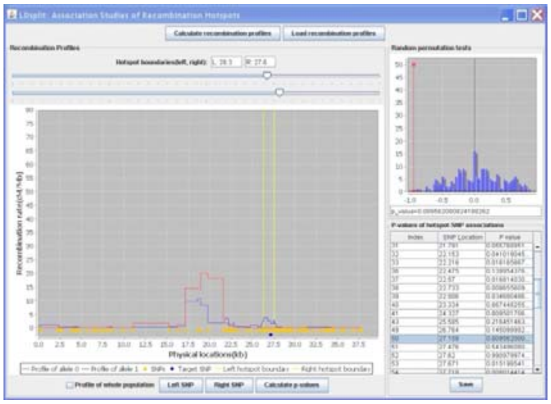
In this stage, using the data output in the first stage, LDsplit visualizes the recombination profiles of the whole and sub-populations of chromosomes for exploratory analysis. To estimate hotspot-SNP associations, users first choose the boundaries of a hotspot using two sliders. For the chosen hotspot, LDsplit calculates the *p*-value of its association with every proximal SNP with minor allele frequency (MAF) no less than a threshold (*e.g.* 30%) in the window. For every candidate SNP, users can browse the recombination profiles of its two alleles (shown as blue and red lines in Figure 3). The physical positions of SNPs are shown as yellow dots below the recombination profiles, and user can click buttons (labelled “Left SNP” and “Right SNP”) to navigate among candidate SNPs (MAF ≥ 30%) and inspect their allele-specific recombination profiles. Moreover, the histogram of random $\Delta\rho$ values from the permutations tests and the observed $\Delta\rho$ value for a SNP chosen by users will be displayed in a window on the top-right panel, illustrating how *p*-values are calculated. A table on the bottom-right panel displays the *p*-values and positions of all candidate SNPs, and user can save it to a text file for further analysis (Figure 3). To identify sequence motifs associated with recombination hotspots, one can search for DNA sequence patterns enriched in proximal regions of the identified SNP.


LDsplit: Association Studies of Recombination Hotspots
Calculate recombination profiles
Load recombination profiles
Recombination Profiles
Hotspot boundaries(left, right):
Random permutation tests
Recombination rate(cM/Mb)
Physical locations(kb)
Profile of allele 0
Profile of allele 1
SNPs
Target SNP
Left hotspot boundary
Right hotspot boundary
P-values of hotspot-SNP associations
Index
SNP Location
P value
Profile of whole population
Left SNP
Right SNP
Calculate p-values
Save

(a) African population (Yoruba in Ibadan, Nigeria, or YRI)


LDsplit: Association Studies of Recombination Hotspots
Calculate recombination profiles
Load recombination profiles
Recombination Profiles
Hotspot boundaries(left, right):
Random permutation tests
Recombination rate(cM/Mb)
Physical locations(kb)
P-values of hotspot-SNP associations
Index
SNP Location
P value
Profile of whole population
Left SNP
Right SNP
Calculate p-values
Save

(b) Asian population (Chinese and Japanese, or CHB+JPT)


LDsplit: Association Studies of Recombination Hotspots
Calculate recombination profiles
Load recombination profiles
Recombination Profiles
Hotspot boundaries(left, right):
Random permutation tests
Recombination rate(cM/Mb)
Physical locations(kb)
P-values of hotspot-SNP associations
Index
SNP Location
P value
Profile of whole population
Left SNP
Right SNP
Calculate p-values
Save

(c) European population (Utah Residents with Northern and Western European Ancestry, or CEU)

Figure 3. Screen shots of LDsplit in analysis of FG11-DNA2 case using HapMap SNP data of three populations: (a) African, (b) Asian, and (c) European. In each screen shot, the DNA2 hotspot is within the two yellow vertical bars, and the FG11 SNP is marked by the blue dot below the orange dots which represent SNPs. The blue and red profiles correspond to the T allele (hot) and C allele (cold) at the FG11 SNP. On the top right is shown the histogram of random *$\Delta\rho$* from 200 permutations, and the red vertical bar marks the observed *$\Delta\rho$* between FG11 alleles. On the bottom right is the table of SNPs and their *p*-values of association with the DNA2 hotspot.

3 Experimental Results

The accuracy and efficacy of LDsplit algorithm have been tested on real and simulated data and achieved successful results in our previous work [8]. For instance, from the proximal regions of SNPs discovered by LDsplit in human chromosome 6, we identified an 11-mer motif that closely matches the 13-mer binding motif of PRDM9, a recently discovered major *trans*-regulator of recombination hotspots in human and mouse genomes [2].

In this newer version, LDsplit was implemented in Java language with a user-friendly interface as shown in Figure 2 and Figure 3. This greatly facilitates the analysis of data, providing users with an integrative view of genomic context of hotspots (e.g. flanking DNA sequences). We tested the newly implemented LDsplit on HapMap SNP data to predict the association of the FG11 SNP with the DNA2 hotspot, previously reported by sperm typing experiment [7]. As shown the Figure 3, the T allele of the FG11 SNP, which corresponds to increased recombination rate in the sperm typing result, has a higher recombination rate than the cold C allele. Moreover, in all 3 populations, FG11 has significant association with the DNA2 hotspot (*p*-values in African, Asian, and European populations are 0.02822, 0.0411 and 0.00956 respectively, all less than the *p*-value threshold of 0.05). This result demonstrates the efficacy of LDsplit to correctly identify SNP-hotspot pairs previously only observed in sperm typing experiments. Note that, while sperm typing is limited to a few short regions due to high cost and technical limitations, the computational method of LDsplit can be applied genome-wide to thousands of hotspots in human as well as other species.

4 Discussions

In this paper we presented LDsplit, an open source Java program, for predicting *cis*-regulatory motifs of meiotic recombination hotspots through SNP analysis. Its graphical user interface facilitates integrated and interactive analysis of hotspots and proximal DNA sequences. It will be a useful tool for many researchers whose projects involve genetic recombination.

The discovery of PRDM9 as a major *trans*-regulator of meiotic recombination is a major recent breakthrough in the field of recombination and chromosome dynamics. However, the functions of PRDM9 and its exact roles in mediating recombination remain unclear. Moreover, it is unlikely that PRDM9 is the only *trans*-regulator of meiotic recombination. Thus it is still a challenge to uncover the whole regulatory pathway, including additional proteins and sequence motifs associated with recombination hotspots.

Through a population genetics approach, the LDsplit software has been used to confirm that the binding motif of PRDM9 is associated the variation of recombination hotspots among individuals [8]. Furthermore, it can be used to search for more motifs associated with hotspots on a genome-wide scale. Such a bioinformatics approach can be used to identify *cis*- and *trans*-regulators that interact with PRDM9 to form the regulatory machinery of recombination and genome integrity.

One implication of the discovery of PRDM9 is that epigenetic factors (*e.g.* DNA methylation and histone modification) play an important role in the regulation of recombination hotspots. Several recent studies confirmed this link [9] [10]. Nevertheless, the genetic factors (*e.g.* GC content, sequence motifs, the binding of transcription factors) are still important determinants of recombination rate. An important future work is to integrate genetic and epigenetic factors into one predictive model, and elucidate their complementary roles in the regulation of recombination dynamics. Thus, the interactions between genetic and epigenetic layers of regulation should also be investigated. To this end, new integrative bioinformatics approaches need to be developed.

Acknowledgements

This project is supported by the following grants: MOE AcRF Tier 1 Grant RG32/11, and Tier 2 AcRF Grant MOE2008-T2-1-074 from Ministry of Education, Singapore; Intramural Research Program of NIH (National Library of Medicine and NIDDK), USA.

References

[1] A. Auton and G. McVean, "Recombination rate estimation in the presence of hotspots," *Genome research,* vol. 17, pp. 1219-27, Aug 2007.

[2] 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,* vol. 327, pp. 876-9, Feb 12 2010.

[3] F. Baudat, J. Buard, C. Grey, A. Fledel-Alon, C. Ober, M. Przeworski, G. Coop, and B. de Massy, "PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice," *Science,* vol. 327, pp. 836-40, Feb 12 2010.

[4] E. D. Parvanov, P. M. Petkov, and K. Paigen, "Prdm9 controls activation of mammalian recombination hotspots," *Science,* vol. 327, p. 835, Feb 12 2010.

[5] 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,* vol. 40, pp. 1124-9, Sep 2008.

[6] S. Myers, L. Bottolo, C. Freeman, G. McVean, and P. Donnelly, "A fine-scale map of recombination rates and hotspots across the human genome," *Science,* vol. 310, pp. 321-4, Oct 14 2005.

[7] A. J. Jeffreys and R. Neumann, "Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot," *Nature genetics,* vol. 31, pp. 267-71, Jul 2002.

[8] 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,* vol. 11, p. R103, 2010.

[9] F. Smagulova, I. V. Gregoretti, K. Brick, P. Khil, R. D. Camerini-Otero, and G. V. Petukhova, "Genome-wide analysis reveals novel molecular features of mouse recombination hotspots," *Nature,* vol. 472, pp. 375-8, Apr 21 2011.

[10] M. Wu, C.-K. Kwoh, T. M. Przytycka, J. Li, and J. Zheng, "Prediction of Trans-regulators of Recombination Hotspots in Mouse Genome," presented at the IEEE Bioinformatics and Biomedicine, Atlanta, USA, 2011.