LDsplit: A Java Program for Association Studies of Meiotic Recombination Hotspots Using SNP Data

*Peng Yang, Jing Guo and Jie Zheng**

Bioinformatics Research Centre (BIRC)
School of Computer Engineering
Nanyang Technological University, Singapore

* Email: zhengjie@ntu.edu.sg
URL: http://www.ntu.edu.sg/home/zhengjie/software/LDsplit.htm

Last updated: 21 October, 2014

Contents

1. Introduction

LDsplit is an open source Java software tool for detecting sequence polymorphisms associated with the regulation of recombination hotspots. The algorithm of LDsplit was originally proposed and published in [1], in which LDsplit was implement as Perl scripts. Using LDhat [2] as a working horse, LDsplit estimates the recombination profiles of two alleles for each candidate SNPs, and use the difference of hotspot strengths between two alleles to measure the significance of hotspot-SNP association. Running on HapMap SNP data, LDsplit was able to correctly predict association the FG11 SNP with the DNA2 hotspot observed by sperm typing experiments [3]. Moreover, from the proximal regions of SNPs discovered by LDsplit in human chromsome 6, we identified an 11-mer motif, GGNGGNAGGGG, which closely matches the 13-mer motif CCNCCNTNNCCNC bound to by PRDM9, the recently discovered *trans*-regulator of recombination hotspots [4].

This manual describes an upgraded version of the LDsplit written in Java, with a graphical user interface (GUI) and interactive data visualization, and gives instructions for installation and usage. It can run under both Microsoft Windows and Linux. First of all, LDsplit provides a GUI for LDhat, which is a command-line C++ program which requires computational sophistication under Linux or MS DOS. In comparison, using the Java version of LDsplit, user can specify parameters of LDhat using a user-friendly graphical interface, and browse the output recombination profiles and LD patterns. Moreover, it visualizes the running of LDsplit with a progress bar, and allows user to study the hotspot-SNP association. For example, user can browse the recombination profiles of sub-populations with certain alleles at each candidate SNP in a window of genomic region, and can export the hotspot-SNP association *p*-value and related information (*e.g.* SNP positions) to a text file for further analyses.

2. Installation

First, we assume that the user has installed JDK and Java SE on his or her computer (JDK and Java SE is available at http://java.com/en/download/inc/windows_upgrade_ie.jsp). The executable file and Java source code of LDsplit software can be downloaded for free at http://www.ntu.edu.sg/home/zhengjie/software/LDsplit.htm. The current LDsplit package includes the following files. First, there is one executable jar file, namely “LDsplit.jar”. User can double click on this file to launch LDsplit. Second, there are files required for running LDhat under Windows and Linux, namely “Lookup”, “Win32” and “Linux”. These files were downloaded from http://www.stats.ox.ac.uk/~mcvean/LDhat/instructions.html. Note that the “LDsplit.jar” file must be in the same directory as “Lookup”, “Win32” and “Linux” folders, because files in these folders are required for LDsplit.jar to run. However, users can create a shortcut of “LDsplit.jar” and move it to any location for convenient access. Third, there are “toy data” for tests, including raw input SNP data in the same format as LDhat, *i.e.* SNP sequence data (called *sites* file) and SNP positions data (called *locs* file), as shown in Figure 4(a) and Figure 4(b); these files can be extracted from the haplotype files of Hapmap http://hapmap.ncbi.nlm.nih.gov/) by cutting a window of chromosomes and processed into LDhat format using "inputfiles generation" function. There are also intermediate files recording the results of LDsplit (namely “result data”) in format of Java object serialization, which can be loaded into LDsplit for visualization and analysis.

To start with, user need first download the LDsplit package in zip file from the LDsplit webpage at http://www.ntu.edu.sg/home/zhengjie/software/LDsplit.htm, and then decompress the zip file into a folder under Windows. Since LDsplit runs as a Java program, please make sure that JRE have been installed and can be accessed in your computer. Thanks to the portability of Java, LDsplit can run under both Windows and Linux operating systems.

To start running LDsplit under Windows, users only need to double click the "LDsplit.jar" file and a main frame window will appear (see Figure 5). Note that LDsplit.jar cannot be moved outside the folder because its execution depends on other files such as the executables of LDhat. However, users can create a shortcut for "LDsplit.jar" and move the shortcut to any location (e.g. Desktop) for convenience of access.

Under Linux, users can run LDsplit either from GUI or from a command-line shell. In the former case, user can right click the icon of "LDsplit.jar" file, and select "open with Java Runtime", and a main frame of LDsplit will appear. Under command-line environment, users first move to the LDsplit directory and type the command "java –jar LDsplit.jar", and the GUI of LDsplit will appear. The rest steps are the same as under Windows.

We have tested LDsplit under Windows XP, Window 7, and Linux Ubuntu 8. For other Linux versions, users may need to re-compile the LDhat source code to generate executable files for their specific Linux platform. To do that, users can first download the source code of LDhat from http://www.stats.ox.ac.uk/~mcvean/LDhat/instructions.html, and follow the instructions for compilation in the manual of LDhat. After the executable files of LDhat are generated, please move them to the "Linux" folder of LDsplit.

3. Generation of input files

The calculation of recombination profiles needs two input SNP data files in the format of LDhat: *sites* file and *locs* file (details see section 4). The function of "Inputfiles generation" is to facilitate users to create *site* file and *loc* file in LDhat format. In Figure 1, it shows the main interface of this new function. Users only need to upload the querying raw phased haplotype data and define regions they want to generate two inputfiles avoiding the complex procedure of creating LDhat files.

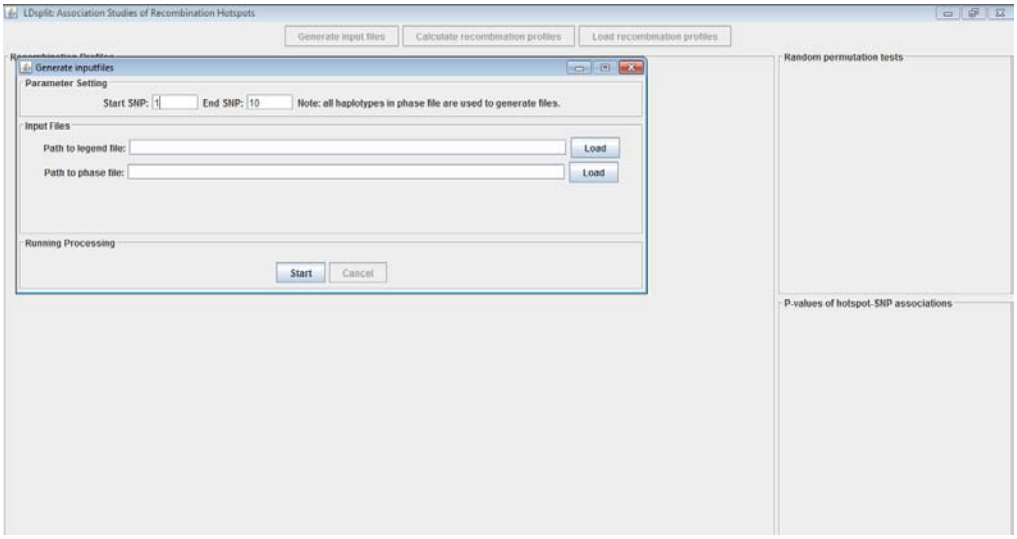

Figure 1. A screenshot of the full interface of "input files generation" function.

3.1 Upload sources files

Firstly, users should download the phased haplotype data from HapMap website: hapmap.ncbi.nlm.nih.gov/downloads/phasing/2006-07_phaseII/phased. Two types of files, phase file and legend file are required (Figure 2). In "toy raw input data" folder, two sample files, "sample.phase" and "sample.legend" are provided for illustration.

Figure 2. The panel for input files.

The phase file contains haplotypes of certain race and chromosome. Each line represents a single haplotype which is composed of 1 or 0 separated by space. The number of 0 and 1 indicates different allele types in specific SNP. The details of allele types and SNP information are recorded in legend file. The legend file consists of SNP information including SNP rc number, physical position and allele types.

Since all haplotypes in phase file are used to generate input files by default, users just need to remove unwanted haplotypes from phase file before uploaded. Besides, users could create the uploaded files using their own data and make sure that they are in HapMap format.

3.2 Region definition

The haplotypes may contain millions SNPs. The parameter setting panel provides users an interface to select a segment of whole sequence (Figure 3). They should define the start SNP and the end SNP of the querying region before *site* file and *loc* file are generated. The end SNP should be smaller than the maximum number of SNPs in individual haplotypes.

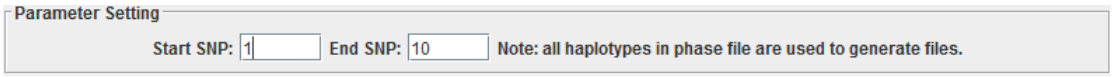

Figure 3. The panel for region selection.

3.3 Save outputs

After users successfully submit the generation task, two files named “input.site” and “input.loc” will be generated in LDhat format in a few seconds. A inquiring window will pop up to ask whether they want to save result files. Select "Yes" and choose a location to save files. Choose "No" to end the task without saving outputs.

4. Calculating Recombination Profiles

In this section, we describe how to use 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 sub-populations from a *random* split of the input population. Since LDhat is computationally costly, this stage of calculation is time consuming. When the calculation finishes, these recombination profiles can be saved to a file by Java serialization, which can be loaded back into LDsplit later for visualization and analysis (to be described in Section 4).

4.1 Input SNP data

User need to prepare two raw input SNP data files in the format of LDhat: (1) *sites* file: A text file consisting of haplotypes in FASTA format, see Figure 4(a); (2) *locs* file: the physical locations of the SNPs on the chromosome, see Figure 4(b). These data represent the genotype within a genomic region (or a window) on a population of chromosomes. They can be generated by "inputfiles generation" function. Users can also try a few example input data in the “toy raw input data” folder contained in the LDsplit package.

As shown in Figure 4(a), in a sites file, the first line consists of three fields: the number of sequences, the number of SNPs in the alignment, and a flag (1 or 2) which is 1 when the SNPs are haplotype (or phased) and is 2 when the data are genotypes (or unphased). The SNP sequence needs to be arranged in one line with its corresponding annotated ID above it (as in FASTA format) and the numbers of SNPs in all sequences must be equal. The number of sequences in each file is at least **20** by default. As shown in Figure 4(b), in a locs file, the first line contains the number of SNPs, the total physical length of the genomic region in kb, and a flag (L or C) where “L” indicates the model fitted is crossing-over and “C” indicates the model is gene conversion. Please see the user manual of LDhat (http://www.stats.ox.ac.uk/~mcvean/LDhat/manual.pdf) for more details of the format.

Figure 4(a). Examples of *sites* files with input SNP sequences in binary numbers or DNA bases.

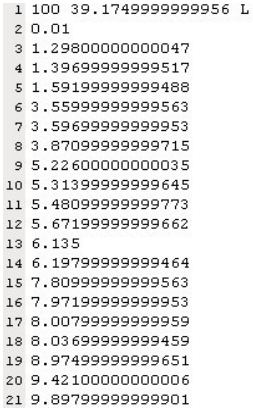

Figure 4(b). An example *locs* file indicating physical locations of SNPs in kb.

To load the sites and locs files to LDsplit, user can launch (by double clicking the “LDsplit.jar” file) the Main Frame of LDsplit, and by clicking the button labeled “Calculate recombination profiles” at the top of the Main Frame, a panel of “Run LDsplit” will appear (see Figure 5). On this panel, after clicking the “Load” buttons of sites file and locs file respectively, users can navigate to the locations of the two files on hard disk.

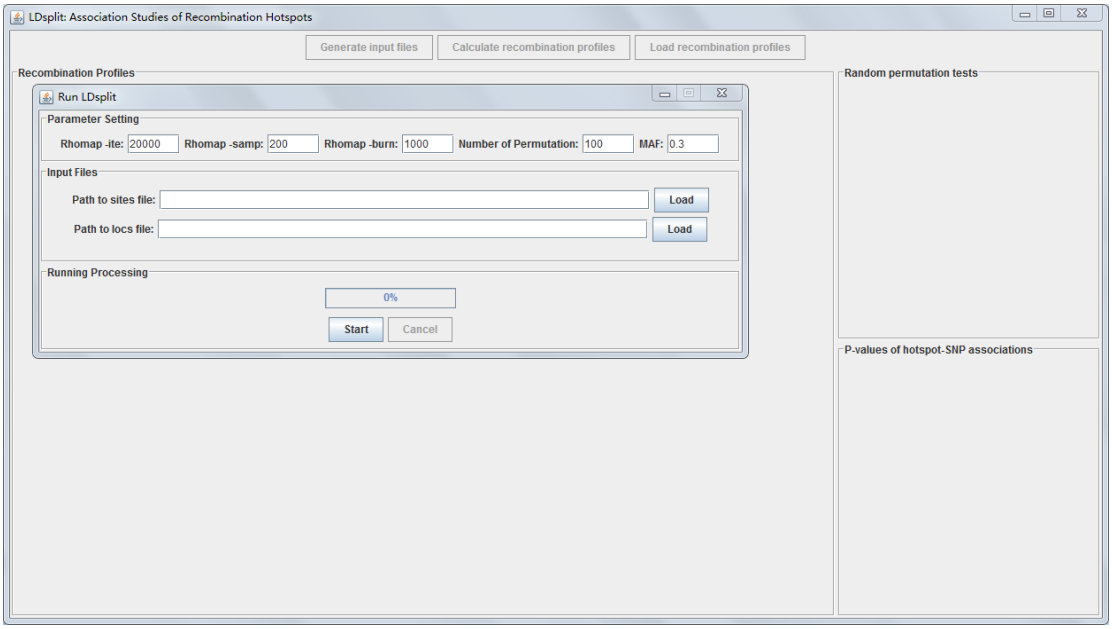

Figure 5. The panel for importing input data and setting parameters.

4.2 Setting LDsplit parameters

On the "Run LDsplit" panel, five parameters can be specified:

- *Rhomap –its*: Number of iterations for the Markov Chain Monte Carlo simulation
- *Rhomap –samp*: Number of iterations between successive samples from the chain
- *Rhomap –burn*: Number of iterations to run chain for as burn-in period
- *Number of Permutations*: Number of permutations (random splits) for *p*-value calculation in LDsplit
- *MAF*: Threshold for every proximal SNP with minor allele frequency (MAF). The MAF is set to 0.3, because SNPs with quite small MAF will provide small samples of haplotypes for which an LD-based method may lead a biased estimation of the recombination rate.

For the first three parameters, user can see the user manual of LDhat for more details. In Table 1 we suggest ranges of values for these parameters according to our experiments. In particular, based on our experience, we suggest that the value of "Rhomap –samp" be set to around 5% of the value of "Rhomap -its".

Table 1. Parameters of LDsplit and suggested ranges of values

Parameter	**Default value**	**Value limitation**	**Description**
Rhomap –its (***i***)	20000	$i \geq 5000$	Number of iterations for rjMCMC
Rhomap –samp (***s***)	200	$s > 100$, $s < 10\% \times i$	Number of iterations between successive samples from chain
Rhomap –burn (***b***)	1000	$b > 0$	Number of iteration to run chain as burn-in period
Number of permutations (***p***)	200	$p > 50$, $p < 1000$	Number of permutations
MAF	0.3	$m>0$, $m<1$	Minor allele frequency

After loading sites and locs file and setting the parameters, user can click the "start" button to run the LDsplit program. The progress bar will be updated periodically to indicate the estimated percentage of total work that has been finished so far. As LDhat is computationally intensive for estimating recombination rate, the progress bar may appear "frozen" during the first couple of minutes, but then its progress will be seen; thus, user's patience would be very much appreciated. If a user wants a quick start to get some rough results, he/she may first set the parameters to relatively small values, and then increase the parameter values to get more precise results later.

4.3 Save recombination profiles to a file

When LDsplit finishes running (*i.e.* the progress bar reaches 100%), a dialogue box will pop-up asking if user wants to save result to a file. To save it, user can click "yes" and specify the file name and target location on hard disk. As mentioned previously, the result file is in the format of Java serialization of objects in the LDsplit program, and thus only LDsplit can recognize its format. To load the result file back, click the "Load recombination profiles" button at the top of LDsplit main frame (Figure 3).

4.4 Cleaning junk files

If running process is cancelled by user (by clicking the "Cancel" button before the progress bar reaches 100%), intermediate sites and locs files and folders containing intermediate data will be left in the same directory as the executable "LDsplit.jar" file. Users need to manually delete these junk files and folders. However, if LDsplit completes the running process without interruption (i.e. the progress bar reaches 100%), no intermediate files and folders will be left.

5. Hotspot-SNP Associations

After the time consuming calculation of recombination profiles is finished, exploratory analyses may be conducted as follows. First, user can browse the recombination profiles of the whole input population of chromosomes. This is actually a visualization interface for LDhat, which is not provided in the original LDhat package. Second, two sliders allow user to specify boundaries of a hotspot according to the recombination profile interactively. In calculating recombination profiles, we did not fix on any hotspot, because a user can specify hotspot boundaries only after seeing the recombination profiles; moreover, a window of genomic region may contain multiple hotspots, but the recombination profile data (saved in a file) can be re-used for different areas of the window. Third, the association of a user-specified hotspot with each candidate SNP can be calculated as the difference of hotspot strengths between two alleles of the SNP, and the *p*-value will be calculated based on the simulation of null distribution by random permutations of chromosomes between the two alleles. Moreover, user can navigate across different SNPs in the window, and browse their associations (*i.e.* hotspot differences between alleles and *p*-values) with the hotspot.

5.1 Loading recombination profiles

To load the file of recombination profiles, user can click the “Load recombination profiles” button on the Main Frame of LDsplit, and navigate to directory of the file (Figure 6). Note that any result file previously exported from LDsplit can be loaded back to be analyzed. For example, the LDsplit package contains a few such files in the “toy result file” folder, which were saved by the authors of this manual.

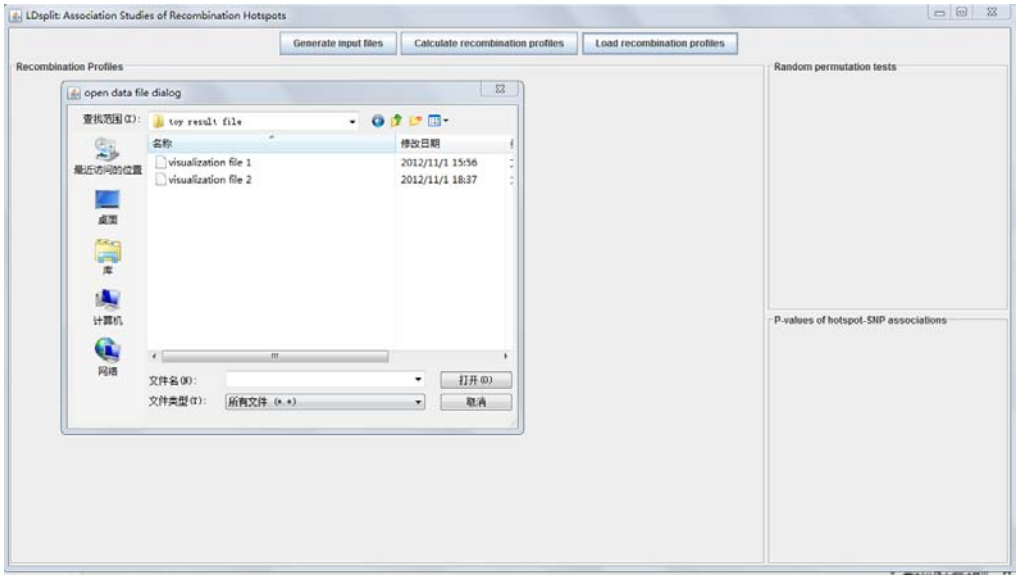

Figure 6. Loading recombination profiles from a file generated by LDsplit itself.

5.2 Plotting recombination profiles

After loading the recombination profile data from a file, the recombination profiles of two split subpopulations, corresponding to the two alleles of a candidate SNP, will be plotted as blue and red lines (Figure 7). In this chart, the *x*-axis represents physical positions along the input window of chromosome, and the *y*-axis is the recombination rate (in cM/Mb). Above the *x*-axis, the yellow triangle dots denote physical locations of input SNPs, and a blue diamond dot indicate the *target* SNP (*i.e.* the red and blue recombination profiles being plotted are from sub-populations corresponding to the alleles of this SNP).

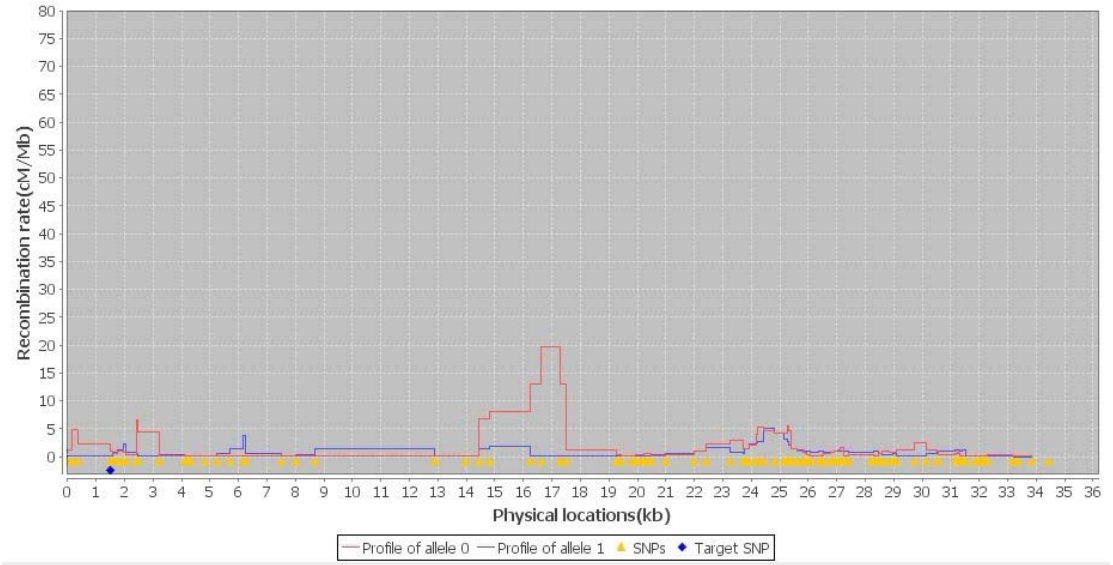

Figure 7. A screenshot of recombination profiles of subpopulations split at a SNP.

If user wants to view the recombination profile of the whole population of chromosomes, he/she can click and tick the square button left side of “Profile of both alleles” at the bottom of the Main Frame, and a green line representing the whole recombination profile will be plotted (Figure 8). The green line will disappear if user removes the tick by clicking the square button again.

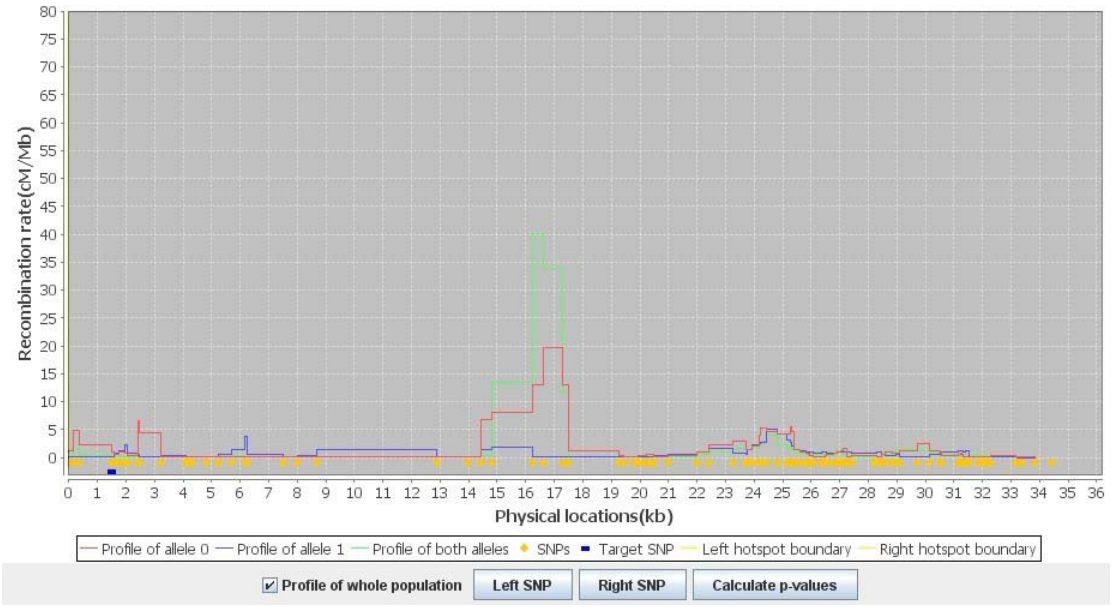

Figure 8. A screenshot of recombination profile of the whole population (the green line).

There are two buttons labeled “Left SNP” and “Right SNP”. By clicking one of the two buttons user can switch target SNP to the neighbor SNP on the left (or the right) side of the current target SNP, and the plot of recombination profiles will be updated accordingly. Note that some SNPs (plotted as yellow dots) do not qualify as candidate split SNP because their MAFs (minor allele frequencies) are smaller than a threshold (say lower than 30%). In this case, the blue dot will *skip* several yellow dots when user clicks the “Left SNP” or “Right SNP” button (*i.e.* the next target SNP might be a few SNPs away from the current target SNP).

5.3 Computing hotspot-SNP associations

To calculate the hotspot-SNP associations, user first need specify the boundaries of a hotspot. To do that, user can drag one of the sliders above the chart of recombination profiles (Figure 9) and move it to the left or to the right. Two red vertical lines show on the line chart will move with the sliders dragged by user. The numbers in two boxes above the sliders show the physical positions of the hotspot boundaries. After fixing the boundaries of a hotspot, user can calculate the association of the hotspot with each candidate SNP by clicking the button labeled “Calculate p-values” below the line chart (Figure 9).

A window of genomic region may contain multiple hotspots, and one hotspot can be studied at a time. While it is time consuming to calculate recombination profiles (Section 3), these data can be reused for different hotspots. With recombination profile data available, it will take only seconds to calculate the associations of a specified hotspot with every candidate SNP.

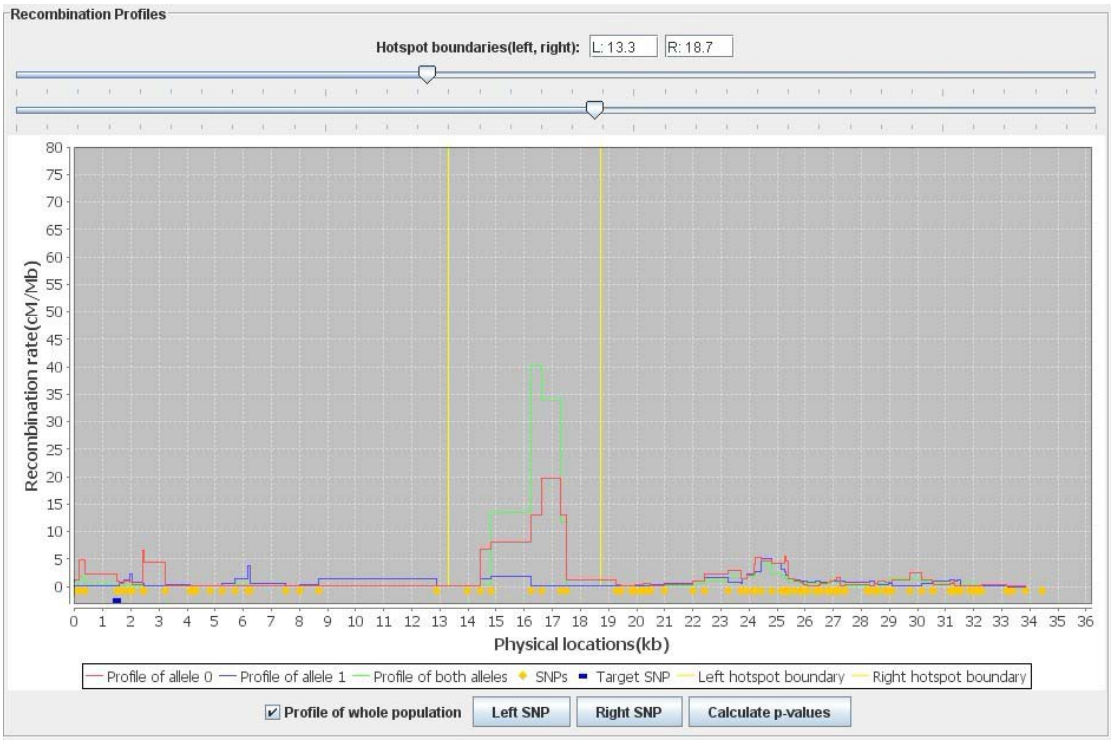

Figure 9. Using sliders to set the left and right boundaries of a hotspot.

The *p*-value of association between a hotspot and a SNP is estimated as follows. For each candidate SNP, LDsplit first divides the population of chromosomes into two subpopulations by SNP alleles, and then calls LDhat to estimate the recombination rates for each sub-population (Section 3). The difference of hotspot strengths between the SNP alleles (denoted by $\Delta\rho$) is defined as $(\rho_0 - \rho_1)/(\rho_0 + \rho_1)$, where ρ_0, ρ_1 denote the hotspot strengths in two different sub-populations. Then, 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.* randomly split populations into pseudo-populations to calculate random $\Delta\rho$ values).

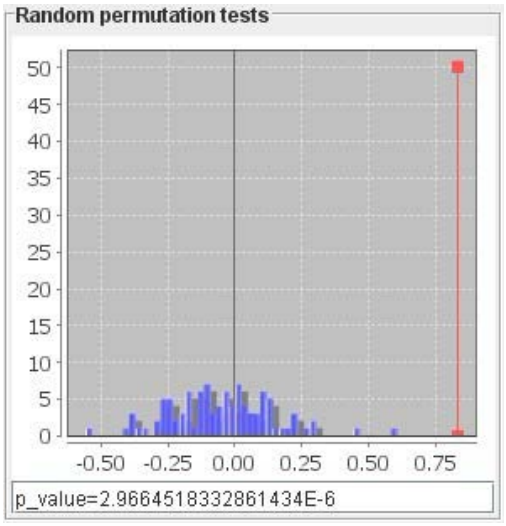

Figure 10. Histogram of random $\Delta\rho$ values from permutation tests (blue) and the observed $\Delta\rho$ value (red line)

In Figure 10, the blue bins represent the histogram of the random $\Delta\rho$ from permutation tests, and the red vertical line marks the observed $\Delta\rho$ value. The *p*-value is shown below the histogram. Usually, the random $\Delta\rho$ values are in Normal distribution, but sometimes they are not. In the latter case, user should take the *p*-value with caution. This histogram is shown at the top-right corner of the LDsplit main frame.

At the bottom-right of the main frame, the *p*-values of all candidate SNPs will be shown in a table, in which each row corresponds to a SNP (Figure 11). There are three columns: (1) "Index" indicates the ID of a SNP counting from the left most SNP in the window; (2) "SNP Location" is the physical position (in kb) of a SNP in the window; (3) "P value" is the *p*-value measuring the statistical significance of association of a hotspot-SNP pair. The table can be exported to a text file (in CSV format) by clicking the "Save" button. The table of *p*-values can be used to predict *cis*-regulatory loci of a recombination hotspot chosen by user, such as to search for genomic elements (*e.g.* repeats, transcription factor binding sites) nearby the SNPs with significant associations (*i.e.* small *p*-values).

If user chooses a different hotspot (using sliders) or chooses a different target SNP, the histogram of random $\Delta\rho$ and *p*-value table will be updated accordingly (Figure 12).

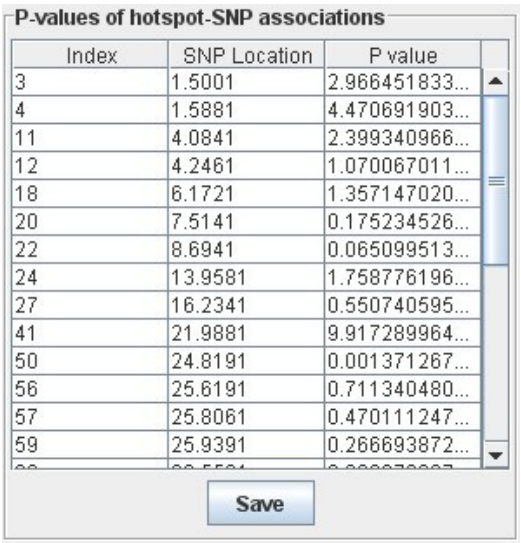

Figure 11. Table of candidate SNPs and their *p*-values for associations with a hotspot.

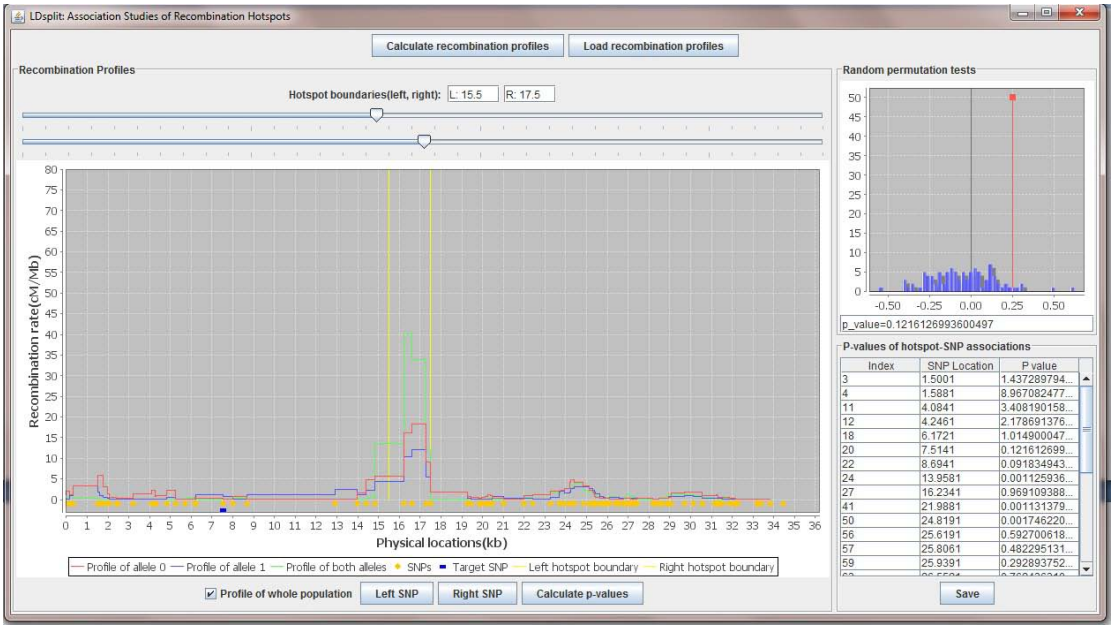

Figure 12. A screenshot of the full interface of LDsplit.

6. Discussions

Recently, we have presented our software on the 8th International Symposium of Integrative Bioinformatics [5] and then extended this work to a journal paper on *BMC Bioinformatics* [6]. Since then, several researchers in genetics have contacted us to inquire the usage of LDsplit for their projects. Thompson P. *et al.* conducted *in silico* analysis with LDsplit that indicated that recombination rates at *DNA3* were influenced by SNPs identified in childhood ALL association studies [7]. In addition, Guo *et al.* (from our group) designed an efficient algorithm to automatically predict parameter values and monitor the mixing process of LDhat [8], which may be used to speed up LDsplit in future. Meanwhile, we have used LDsplit to further analyze the DNA sequence motif CCTCCCT bound by PRDM9 with results consistent with its regulatory role in meiotic recombination hotspots. In parallel, our group also conducted research on the *trans*-regulators of recombination hotspots , which highlighted epigenetic mechanisms of recombination [9]. Thus we also make efforts to build integrative model of genetic and epigenetic factors [10]. In future, we will integrate LDsplit software with large-scale genetic and epigenetic data into a data mining system, which will speed up the discovery of regulatory mechanisms of meiotic recombination hotspots and genome instability.

7. Bugs and Idiosyncrasies

The authors have carried out extensive testing and debugging of the program, which should be generally stable and functional if the instructions in this manual are followed. It is our hope that, with the GUI, users without much computational sophistication can also apply our algorithm to analyze their data smoothly. However, since the authors are not professional programmers, LDsplit does not behave or look like a commercial software. Please send your suggestions,

comments, reports of bugs and errors to Jie Zheng at zhengjie@ntu.edu.sg. Thank you!

8. Acknowledgements

This project is currently supported by Tier 1 AcRF Grant MOE RG 32/11 (M4010977.020) from Ministry of Education, Singapore. Previously it was partially supported by: Startup Grant M4080108.020, College of Engineering, Nanyang Technological University, Singapore; Intramural Research Program of National Library of Medicine, National Institutes of Health, U.S.A.

References

1. Zheng J, Khil PP, Camerini-Otero RD, Przytycka TM: **Detecting sequence polymorphisms associated with meiotic recombination hotspots in the human genome**. *Genome Biol* 2010, **11**(10):R103.
2. Auton A, McVean G: **Recombination rate estimation in the presence of hotspots**. *Genome Res* 2007, **17**(8):1219-1227.
3. Jeffreys AJ, Neumann R: **Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot**. *Nat Genet* 2002, **31**(3):267-271.
4. Myers S, Bowden R, Tumian A, Bontrop RE, Freeman C, MacFie TS, McVean G, Donnelly P: **Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination**. *Science* 2010, **327**(5967):876-879.
5. Yang P, Wu M, Kowh CK, Khil PP, Camerini-Otero RD, Przytycka TM, Zheng J: **Predicting DNA sequence motifs of recombination hotspots by integrative visualization and analysis**. In: *International Symposium on Integrative Bioinformatics; Hangzhou, China*. 2012: 52 - 58.
6. Yang P, Wu M, Kowh CK, Khil PP, Camerini-Otero RD, Przytycka TM, Zheng J: **LDsplit: screening for cis-regulatory motifs stimulating meiotic recombination hotspots by analysis of DNA sequence polymorphisms**. *BMC Bioinformatics* 2014, **15**:48.
7. Thompson P, Urayama K, Zheng J, Yang P, Ford M, Buffler P, Chokkalingam A, Lightfoot T, Taylor M: **Differences in Meiotic Recombination Rates in Childhood Acute Lymphoblastic Leukemia at an MHC Class II Hotspot Close to Disease Associated Haplotypes**. *PloS one* 2014, **9**(6):e100480.
8. Guo J, Jain R, Yang P, Fan R, Kowh C-K, Zheng J: **Reliable and Fast Estimation of Recombination Rates by Convergence Diagnosis and Parallel Markov Chain Monte Carlo**. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2014, **11**(1):63-72.
9. Wu M, Kwoh CK, Przytycka TM, Li J, Zheng J: **Epigenetic functions enriched in transcription factors binding to mouse recombination hotspots**. *Proteome Sci* 2012, **10 Suppl 1**:S11.
10. Wu M, Kwoh CK, Przytycka TM, Li J, Zheng J: **Integration of Genomic and Epigenomic Features to Predict Meiotic Recombination Hotspots in Human and Mouse**. In: *The 2012 ACM International Conference on Bioinformatics, Computational Biology and Biomedicine (ACM-BCB 2012); Orlando, Florida, USA*. 2012: 297-304.