Introduction

The program implements the method for estimating the distribution of fine-scale recombination rates and recombination hotspots along sequences using genetic data sampled from one or more populations.

The program accounts for the effect of population subdivision and other effects such as uncertainty in genotype phasing, mutation rate and other parameters. Simulation studies and analyses on real data showed that the method performed well in terms of identifying recombination hotspots and estimating recombination rates.

The program is distributed as a free software, so that people can make use of the software to analyze their data, especially now that the program can be applicable to genomic data from cancer patients by making use of the single-cell sequencing technique. If you have any question using the software, please drop me an email. My email address is yingw09@gmail.com. I am happy to work with you helping you with interpreting your cancer genomic data.

InferRho

The program reflect newest changes to older versions. The current version is 2.0. If you have any questions regarding the use of the program, please email me.

For compiling, type

```
Make InferRho
```

Note that the software requires libraries Open MP (omp) and gsl (It is a temporary version. The gsl library is not optimal. We will make changes to that later.). You will need specify the path to gsl library in the Makefile.

There are four example data files are given. Three of the files span the beta-globin region from two humans, CEPH and Beni, and chimpanzee. One of them is from a region of chr21 from human population YRI.

For running the example files, please type:

```
./inferRho CEPH_Bglobin_pedfile_web_IR -adjIter 50000 -iter 2000000 -nChains 3
./inferRho Beni_Bglobin_pedfile_web_IR -adjIter 50000 -iter 2000000 -nChains 3
./inferRho Chimp_Bglobin_pedfile_web_IR -adjIter 50000 -iter 2000000 -nChains 3
./inferRho ir_input_yri_0 -adjIter 50000 -iter 2000000 -nChains 3
```

The program will generate several output files.

1) *.log: log file

2) *.rec: sampled recombination along chromosomes

3) *.rho: sampled rho between markers

4) *.rhoBackg: sampled background rho between markers

5) *.monitor: log-likelihood, log-prior, and the some of the two, from each sampling iterations

6) *.hotsp: sampled hotspots along chromosomes

7) *.tmrcas: sampled TMRCA for each markers

8) *.pars: sampled parameters theta and rhoBackgLambda

9) *.mcmc: internal data at the end of the given iteration

10) *.hap: sampled haplotypes if the data are genotypes

11) *.miss: sampled missing data if the data contain missing information

12) *.roothap: sampled haplotypes for the most recent common ancestor of the sampled chromosomes

Input file format

For the example input file “Beni_Bglobin_pedfile_web_IR”,

```
1     #the first line indicates the number of intervals in the input file

g     #g (genotype) or h (haplotype)

47    # the number of genotypes

30     #the number of markers

C C G C C T G G T T A G C G G T G A A C T A C C G C G C A A

C C G C C T G G T T A G C G G T G A A T T A C G G T G C A A

                                        …

C A A C C C G C G T A G A C G N A A A C T A C C G T G C A A

T C G G G T G G T T A T C G G N G C A T T A C C G T G C A A

# genotypes in the data

0 1248 4975 5488 14703 16553 17046 17088 17249 17270 17402 17458 17559 17641 17696 18187
18672 18679 19957 20050 20057 20127 20316 20336 20416 20875 21478 23197 26858 29802
#positions of the markers
```

When the number of markers in the input file is larger, we usually divided the interval into sub-intervals, and specify the number of intervals on the first line. For example, in the example input file “ir_input_yri_0”, the first line is the number of interval which is 20 for the file. Similar to the above example, the second line specifies it is g (genotype) or h (haplotype). The third and fourth lines specify the number of genotypes and the number of markers. Then the genotypes are given followed by the positions of the markers. Then you continue to specify these attributes for the second interval until the last interval. Please see the example input file “ir_input_yri_0” for formatting the input file. If you have question regarding the input file format, please contact me.

Output files

1) The most important output files is *.hotsp which gives the sampled hotspots along the chromosome over iterations.

   It first gives the number of iterations followed by the number of recombination hotspots at the iteration, followed by the recombination hotspots including starting position, ending position, and the intensity of the hotspot.

   You will need to write a program to process the output. You will want to know how often a position is within a recombination hotspots over iterations as an estimate of the posterior probability of hotspot for a given position.

2) Another useful output file is “*.rho” which gives the sampled recombination rate (in rho) between markers.

   The format is quite straightforward. The first line is the header line including iteration and the mid-point of marker positions, followed by the iteration number and the sample recombination rate between makers.

Other output file could be useful, and the meaning of each file is given above. If you have questions regarding the interpretation of the output file or the processing of the output file, please contact me. I can be reached at yingw09@gmail.com.