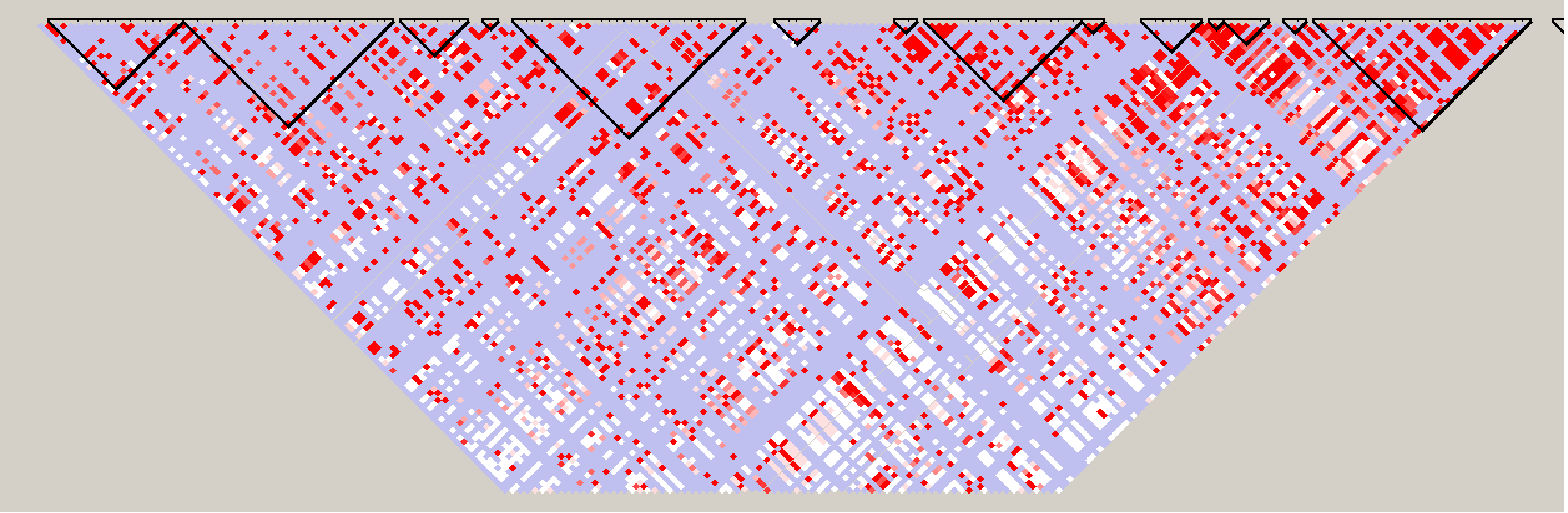
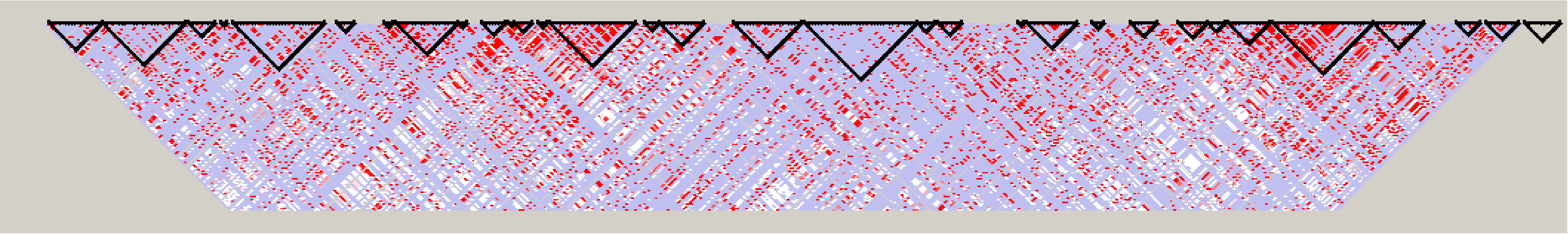
GENOME: Whole Genome Coalescent Simulator (2Mb region) **Haploview Plot**

N=10000 diploid individuals, n=200, nPOP=1, fragment=20000, length=100, numChr=1, #SNP=Poisson (result=4943), rec=1e-6, mut=1e-8

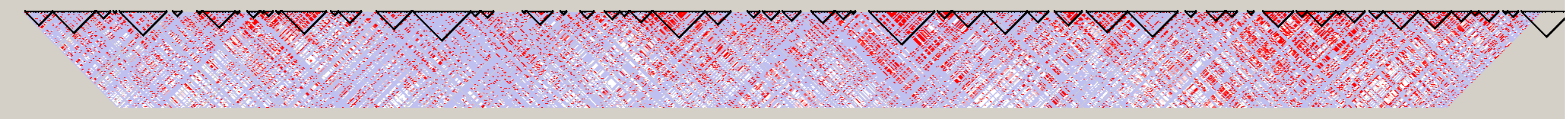
Marker1-200



Marker 1-500



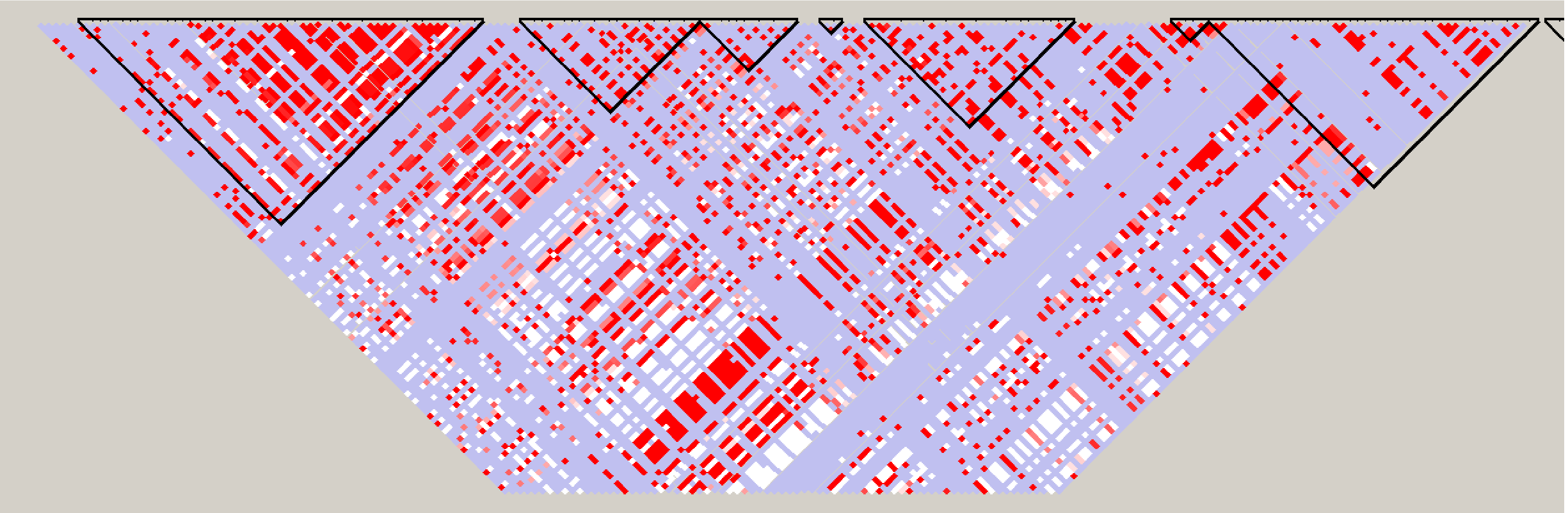
Marker 1-1000



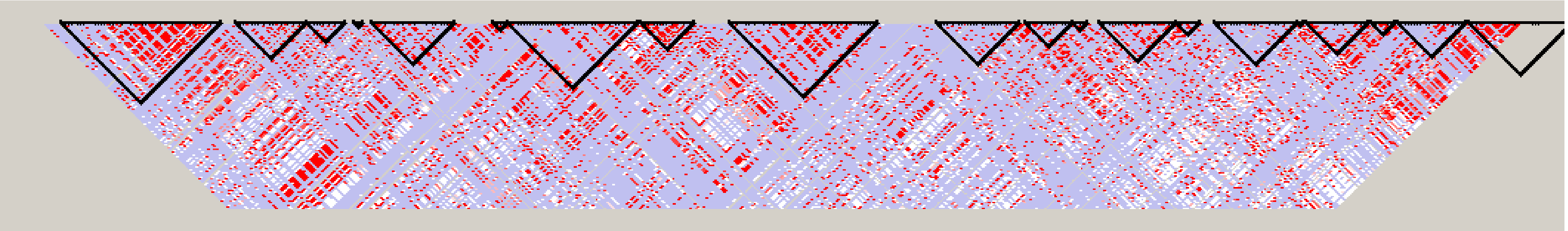
Hudson's ms (2Mb region, equivalent settings) **Haploview Plot**

ms 200 1 -t 800 -r 800 20000, 4Nr=800, 4Nu=800, fragment=20000, result=4658 SNPs

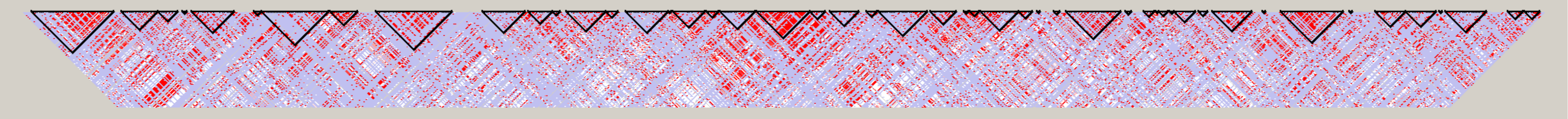
Marker1-200



Marker 1-500



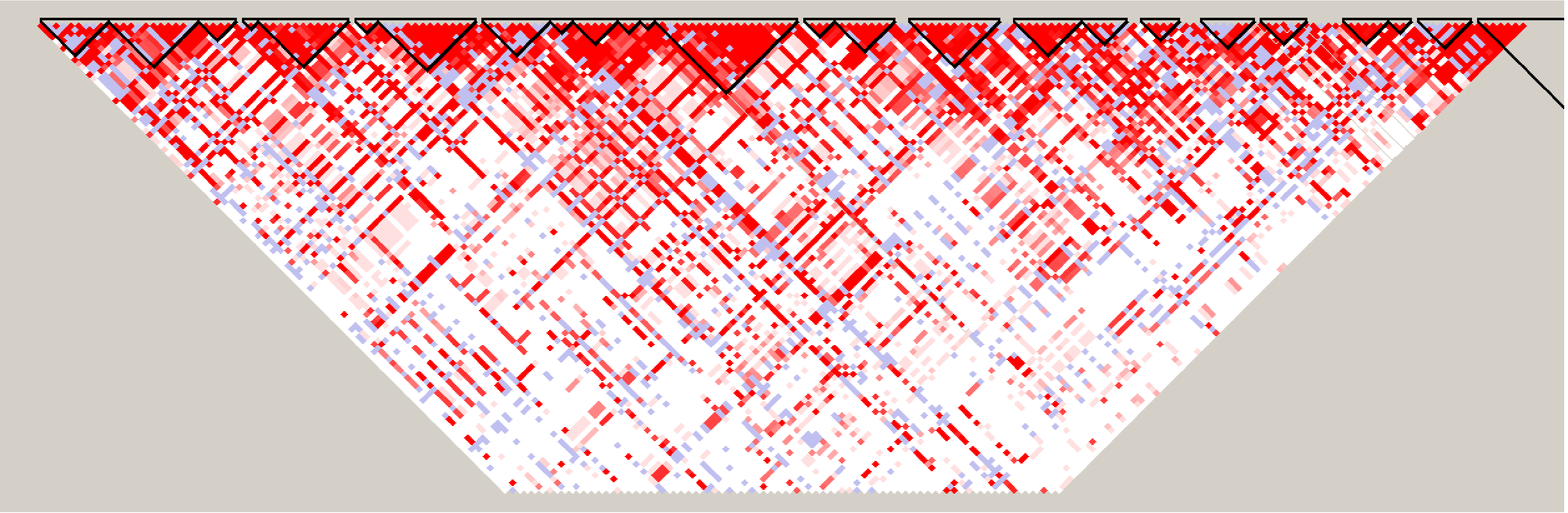
Marker 1-1000



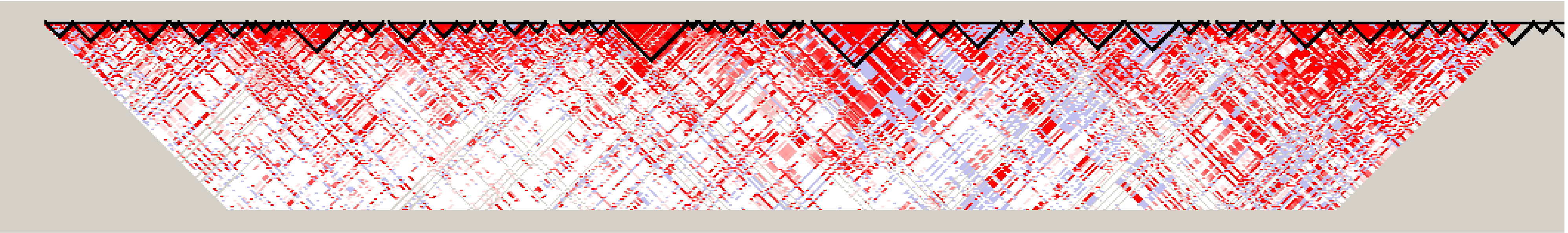
GENOME: Whole Genome Coalescent Simulator (SNPs with MAF > 0.05, 2597 common SNPs) **Haploview Plot**

N=10000 diploid individuals, n=200, nPOP=1, fragment=20000, length=100, numChr=1, #SNP=Poisson (result=4943), rec=1e-6, mut=1e-8

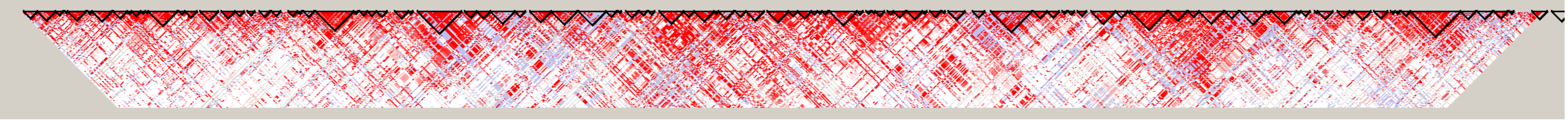
Marker1-200



Marker 1-500



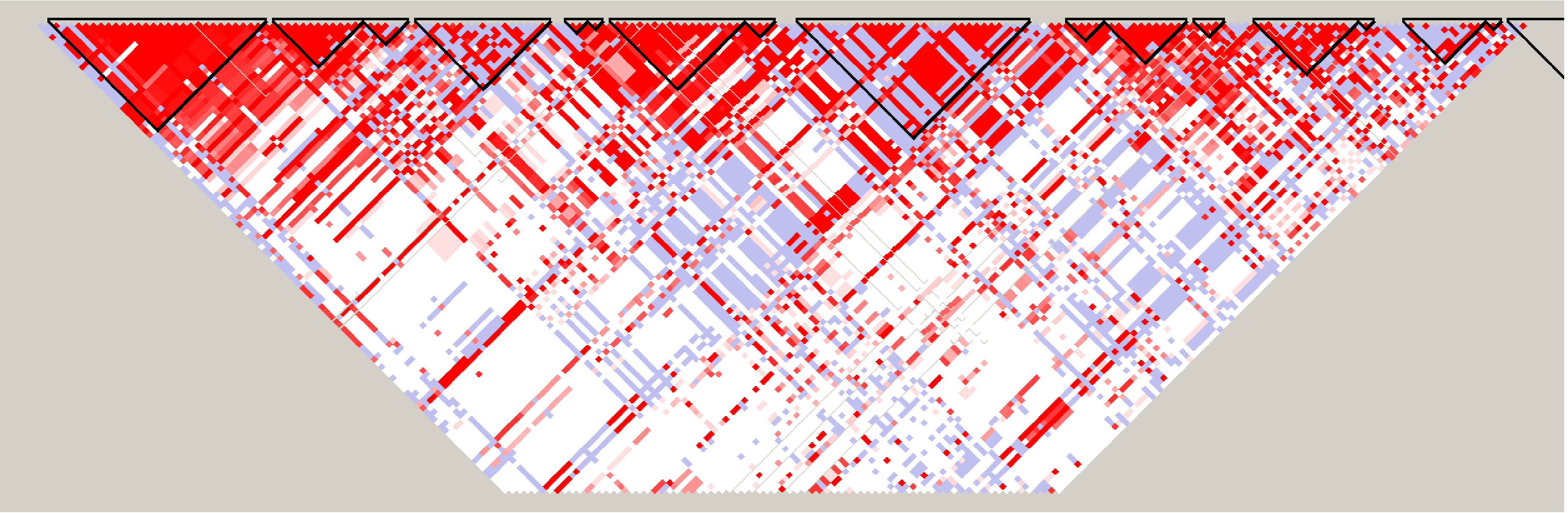
Marker 1-1000



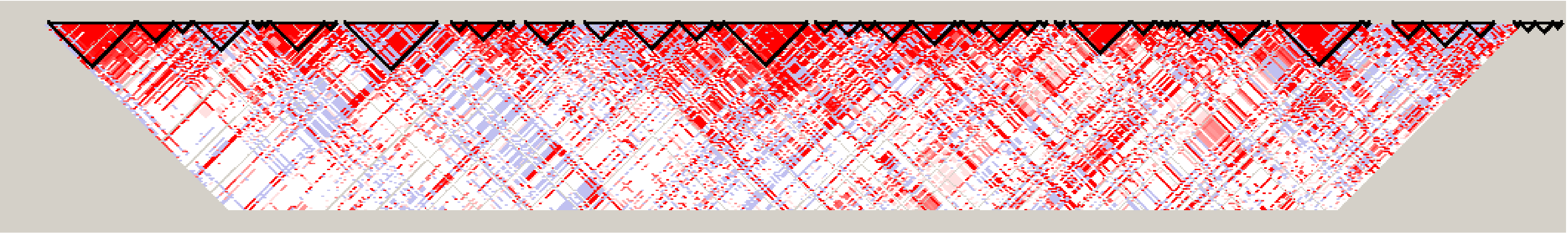
Hudson's ms (SNPs with MAF > 0.05, 2426 common SNPs) **Haploview Plot**

ms 200 1 -t 800 -r 800 20000, 4Nr=800, 4Nu=800, fragment=20000, result=4658 SNPs

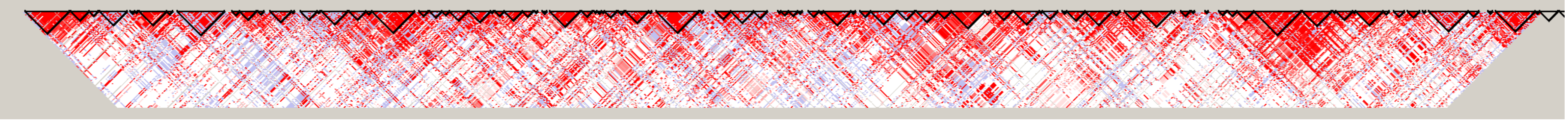
Marker1-200

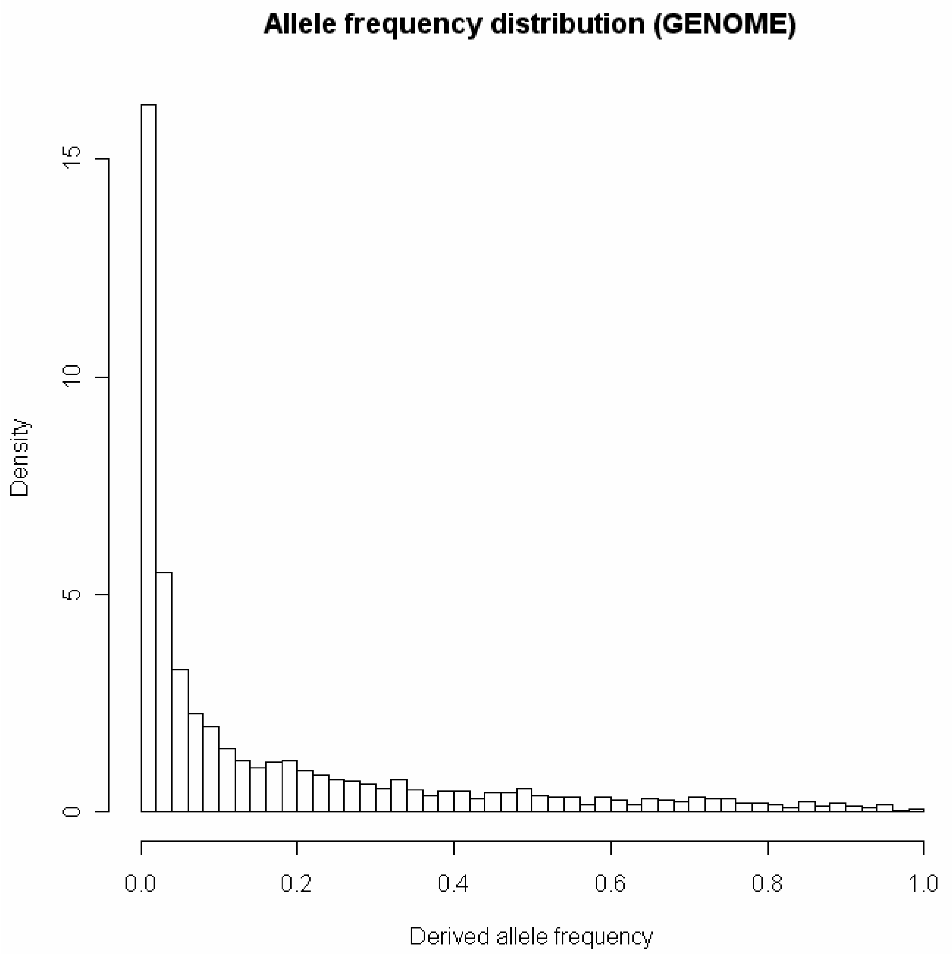


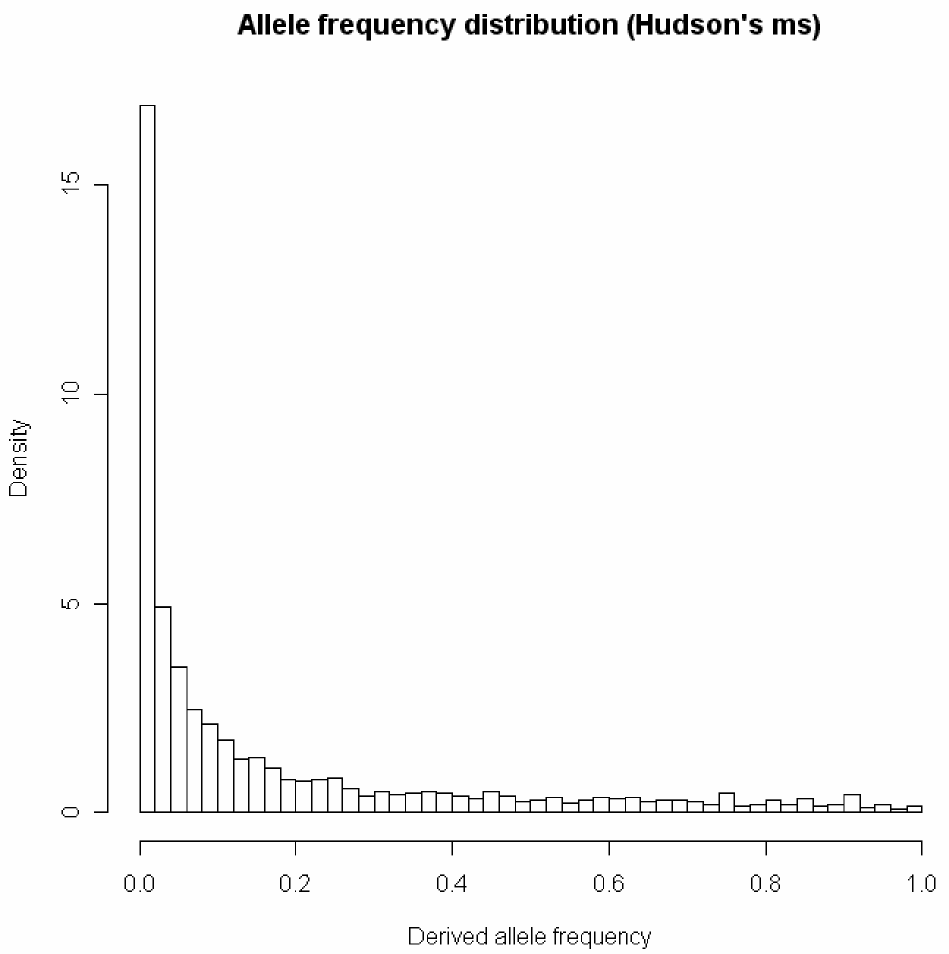
Marker 1-500



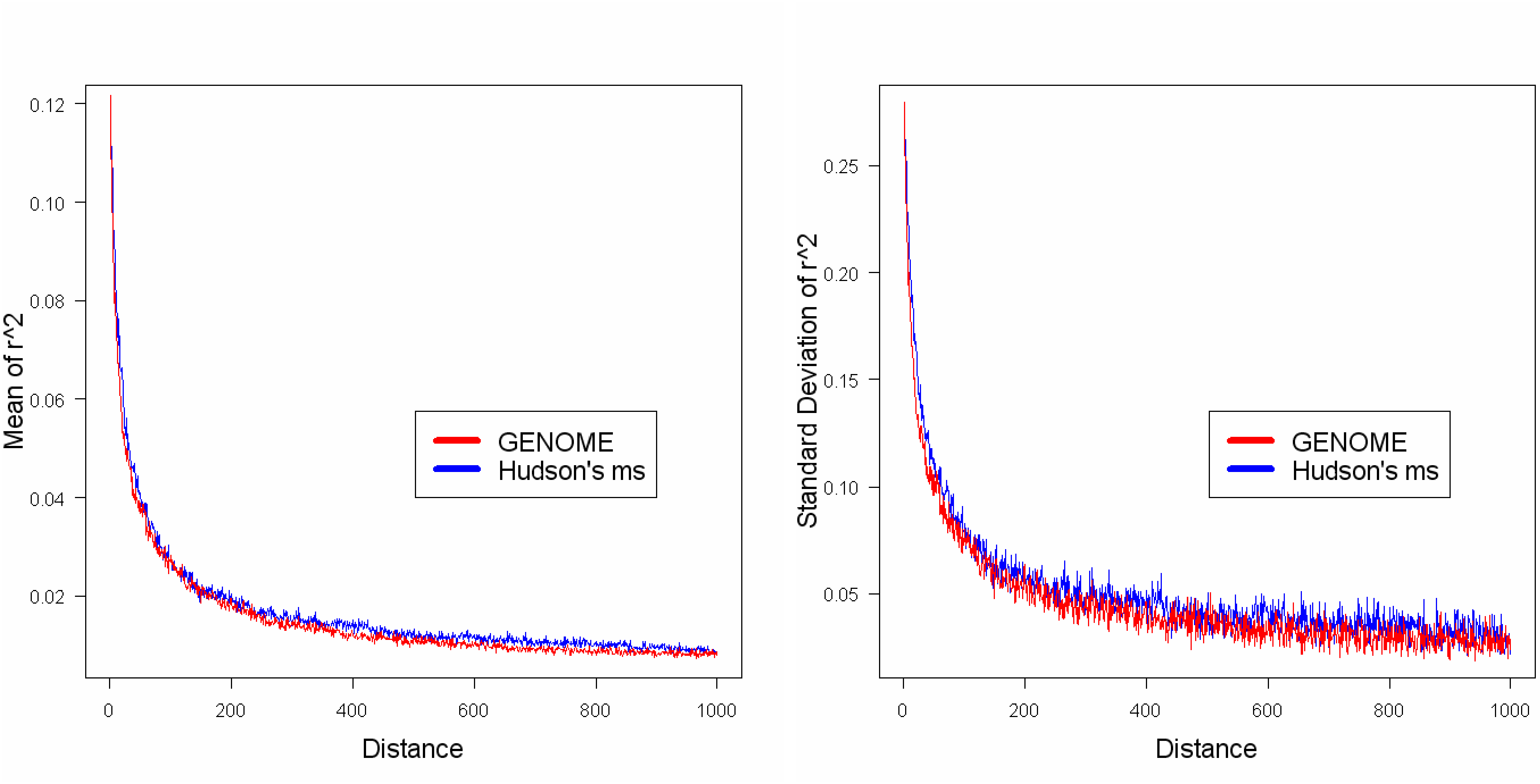
Marker 1-1000




Allele frequency distribution (GENOME)
Density
0
5
10
15
0.0
0.2
0.4
0.6
0.8
1.0
Derived allele frequency

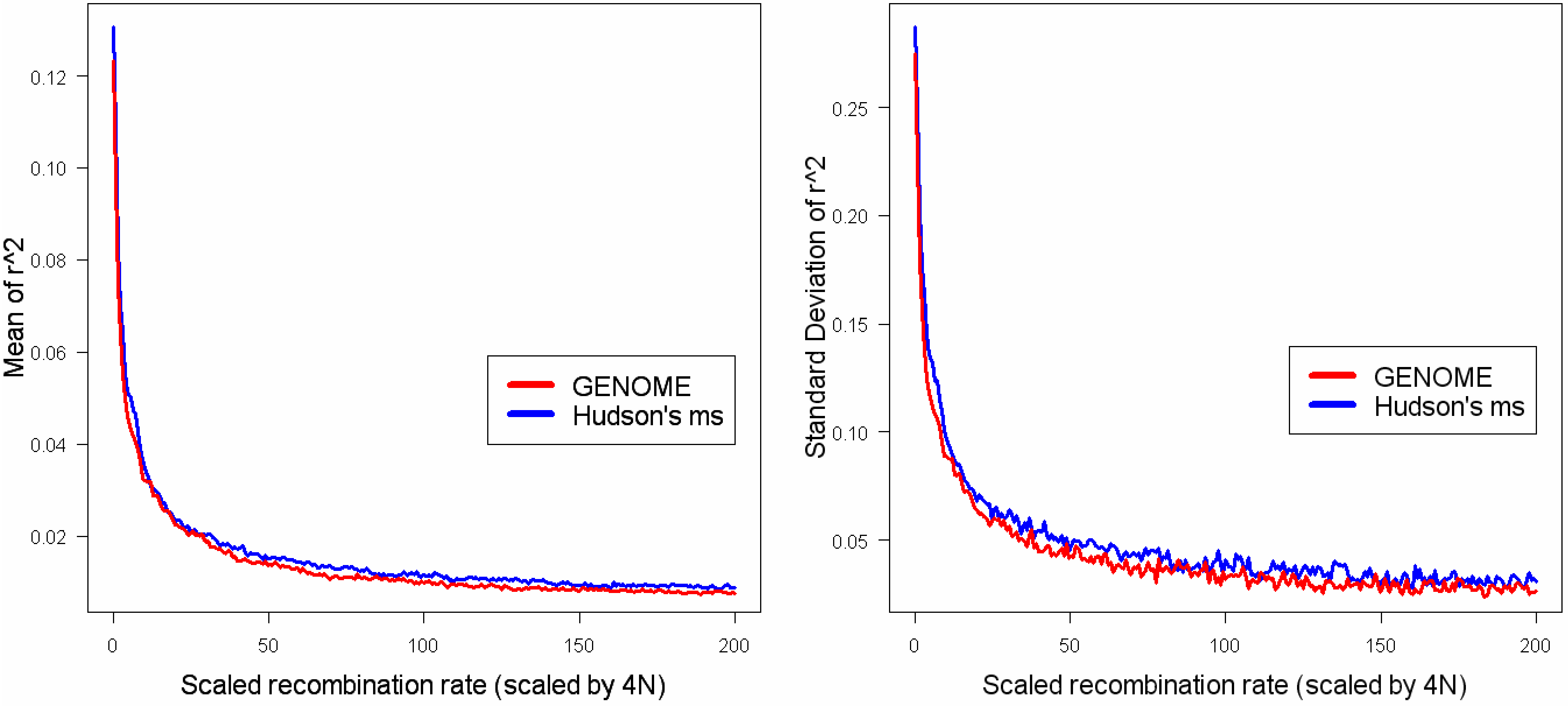

Allele frequency distribution (Hudson's ms)
Density
0
5
10
15
0.0
0.2
0.4
0.6
0.8
1.0
Derived allele frequency

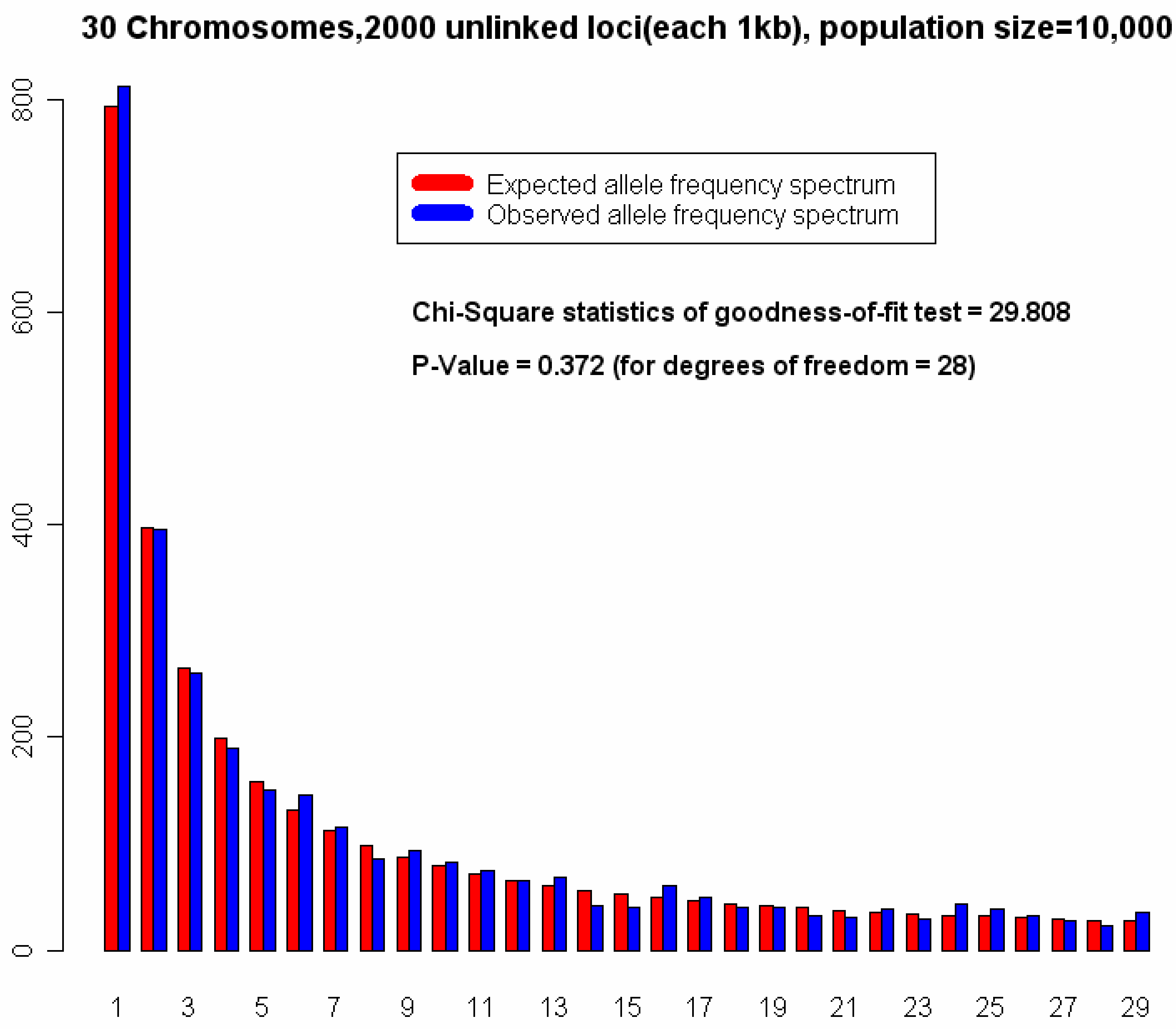
Distribution of LD by distance



Distance is defined as the number of intervening SNPs.

Distribution of LD by genetic distance





30 Chromosomes,2000 unlinked loci(each 1kb), population size=10,000

Chi-Square statistics of goodness-of-fit test = 29.808

P-Value = 0.372 (for degrees of freedom = 28)

Run Time Comparison

Settings:

Effective population size, N=10000 diploid individuals (for 2 pops, N=the size of each subpopulation)
1200 chromosomes (600/600 for 2 populations)
15000 fragments for 150Mb region.
Migration rate = $2.5*10^{-4}$ per generation, (4Nm=10)
Mutation rate = 10^{-8} per base pair (4Nu=60000 for 150 Mb)
Recombination rate =10^{-8} per base pair (4Nr=60000 for 150Mb)

Length of Region	**Number of Populations**	**GENOME (seconds)**	**ms (seconds)**
150Mb	1	1556 (25.9 mins)	13416 (3.7 hrs)
150Mb	2	3964 (66.1 mins)	45138 (12.5 hrs)

We simulate 300 sequences of 20Mb, 30Mb or 40Mb using COSI v1.1 and GENOME. COSI crashes when simulating 50Mb and 300 sequences or for longer region and more sequences. Observed that COSI took about 2.8G memory when simulating 30Mb region.

	COSI v1.1 (seconds)	**GENOME (seconds)**
40M	3089	294
30M	1658	240
20M	677	150