

Phasing of 2-SNP Genotypes based on Non-Random Mating Model

Dumitru Brinza* and Alexander Zelikovsky**

Department of Computer Science, Georgia State University, Atlanta, GA 30303
{dima,alex}@cs.gsu.edu

Abstract. Emerging microarray technologies allow genotyping of long genome sequences resulting in huge amount of data. A key challenge is to provide an accurate phasing of very long single nucleotide polymorphism (SNP) sequences. In this paper we explore phasing of genotypes with 2 SNPs adjusted to the non-random mating model and then apply it to the haplotype inference of complete genotypes using maximum spanning trees. The runtime of the algorithm is $O(nm(n+m))$, where n and m are the number of genotypes and SNPs, respectively. The proposed phasing algorithm (2SNP) can be used for comparatively accurate phasing of large number of very long genome sequences. On datasets across 79 regions from HapMap[7] 2SNP is several orders of magnitude faster than GERBIL and PHASE while matching them in quality measured by the number of correctly phased genotypes, single-site and switching errors. For example, 2SNP requires 41 s on Pentium 4 2Ghz processor to phase 30 genotypes with 1381 SNPs (ENm010.7p15:2 data from HapMap) versus GERBIL and PHASE requiring more than a week of runtime and admitting no less errors than 2SNP¹.

Keywords: haplotypes, genotypes, SNP, algorithm, phasing

1 Introduction

The difference between individual DNA sequences mostly occurs at a single-base site, in which more than one nucleic acid or gap is observed across the population. Such variations are called single nucleotide polymorphisms (SNPs). The number of sufficiently frequent SNPs in the human population is estimated to be around 10 million [11]. For complex diseases caused by more than a single gene it is important to identify a set of alleles inherited together. Identification of haplotypes, the sequences of alleles in contiguous SNP sites along a chromosomal region, is a central challenge of the International HapMap project [7]. The number of simultaneously typed SNPs for association and linkage studies is reaching 250,000 for SNP Mapping Arrays (Affymetrix).

Diploid organisms, like human, have two near-identical copies of each chromosome. Most experimental techniques for determining SNPs do not provide

* Supported by GSU Molecular Basis of Disease Fellowship

** Partially supported by NIH Award 1 P20 GM065762-01A1

¹ 2SNP software is publicly available at <http://alla.cs.gsu.edu/~software/2SNP>

the haplotype information separately for each of the two chromosomes. Instead, they generate for each site an unordered pair of allele readings, one from each copy of the chromosome, which is called a genotype.

The input to the phasing problem consists of n genotype vectors each with m coordinates corresponding to SNPs. The phasing problem asks for explaining each genotype with two haplotypes corresponding to chromosomes. In general as well as in common biological setting, there are 2^{k-1} of possible haplotype pairs for the same input genotype with k heterozygous sites.

Computational inferring of haplotypes from the genotypes (or *phasing*) has been initiated by Clark[2] who proposed a parsimony-based approach. It has been shown later that the likelihood based expectation-maximization (EM) is more accurate [13]. Markov chain Bayesian haplotype reconstruction methods have been used in PHASE [14] and HAPLOTYPER [12]. A combinatorial model based on the perfect phylogeny tree assumption was suggested in [5]. HAP [6] exploits perfect phylogeny model and block structure showing good performance on real genotypes with low error rates. Recently, GERBIL [10] has combined block identification and phasing steps for reliable phasing of long genotypes.

In this paper we first explore phasing of genotypes with 2 SNPs which have ambiguity when the both sites are heterozygous. Then there are two possible phasing and the phasing problem is reduced to inferring their frequencies. Given phasing solution for 2-SNP genotypes, complete haplotypes for a given genotype can be inferred based on the maximum spanning tree of a complete graph with vertices corresponding to heterozygous sites and edge weights given by inferred 2-SNP frequencies. In tests on real datasets across 79 different genomic regions from HapMap [7] 2SNP was several orders of magnitude faster than GERBIL and PHASE while matching them in quality. We also found that HAPLOTYPER is almost as fast as 2SNP and close in quality, but currently does not handle populations with more than 500 SNPs. A brief description of 2SNP software can be found in the application note [1].

The rest of the paper is organized as follows. The next section describes proposed 2-SNP genotypes phasing. Section 3 describes the phasing of long genotypes using 2-SNP genotypes phasing. Section 4 describes tested datasets and defines phasing accuracy measures and compares the quality and runtime of PHASE, GERBIL, HAPLOTYPER and 2SNP algorithms.

2 Phasing of 2-SNP genotypes

In this section we first formally introduce the phasing problem and suggest a LD-based formula for the expected frequencies of cis- or trans-phasing of 2-SNP genotypes. We conclude with adjusting of expected haplotype frequencies to deviation from the random mating model.

The input to the phasing problem consists of n genotype vectors each with m coordinates corresponding to SNPs. SNP values belong to $\{0, 1, 2, ?\}$, where 0's and 1's denote homozygous sites with major allele and minor allele, respectively; 2's stand for heterozygous sites, and ?'s denote missed SNP values. Phasing

replaces each genotype vector by two haplotype vectors with SNP values in $\{0, 1\}$ such that any genotype 0-SNP (resp. 1-SNP or 2-SNP) is replaced with two haplotype 0-SNPs (resp. two 1-SNPs or 0-SNP and 1-SNP). A 2-SNP genotype 22 can be cis-phased, i.e., represented as 00 and 11 haplotypes, or trans-phased, i.e. represented as 01 and 10 haplotypes.

Certainty of cis- or trans- phasing. It is natural for the certainty of cis- or trans- phasing of homozygous SNPs i and j to take in account odds ratio of phasing being cis- or trans-, which is $\lambda = \frac{F_{00} \times F_{11}}{F_{01} \times F_{10}}$ where $F_{00}, F_{01}, F_{10}, F_{11}$ are unknown true frequencies of haplotypes with the first and the second binary index denoting alleles of the i -th and j -th SNP, respectively. In our experiments, we have noticed that the modified odds ratio $\lambda' = \frac{F_{00} + F_{11}}{F_{01} + F_{10}}$ better describes real cis- or trans- phasing. We measure LD (linkage disequilibrium) between endpoints by the ratio of the modified odds ratio λ' over the expected value of λ' computed in assumption of no linkage between endpoints, $LD_{ij} = \frac{\lambda'}{\exp(\lambda')}$. Finally, it has been observed[12, 13, 10] the higher LD between pairs of closer SNPs. In order to discard falsely encountered LD between non-linked SNPs which are far apart, we divide LD by the square of the distance between the SNPs obtaining $c_{ij} = \frac{\log LD_{ij}}{(i-j)^2}$. The complete formula for the certainty of cis- or trans-phasing of two homozygous SNPs i and j is logarithm of linkage disequilibrium for the cis-/trans- odds ratio divided by squared distance between corresponding SNPs.

$$c_{ij} = \log \left(\frac{n + (F_{00}F_{11} - F_{01}F_{10}) / (F_{01} + F_{01})}{n - (F_{00}F_{11} - F_{01}F_{10}) / (F_{00} + F_{11})} \right) / (i - j)^2 \quad (1)$$

where n is number of input genotypes, and $F_{00}, F_{01}, F_{10}, F_{11}$ are frequencies of haplotypes with the first and the second binary index denoting alleles of the i -th and j -th SNP, respectively. Haplotype frequencies are computed based on all genotype frequencies except 22. For 22 genotypes, the haplotype frequencies are chosen to fit best Hardy-Weinberg equilibrium adjusted to observed deviation in single-site genotype distribution.

Adjusting observed frequencies to deviations from the random mating model. The certainty formula (1) cannot be directly used in phasing since the participating haplotype frequencies are *true* frequencies and, therefore, are unknown. We have access only to *observed* haplotype frequencies which can be extracted from all types of 2-SNP genotypes except 22, genotypes heterozygous in the both SNPs.

The distribution of cis- and trans-phasing of 22-genotypes can be adjusted to the unknown mating model as follows. Let C_{22} and P_{22} denote unknown numbers of trans- and cis-phasings, then $C_{22} + P_{22} = G_{22}$, where G_{22} is the observed number of 22-genotypes. Then the adjusted odds ratio λ' can be expressed as $\lambda' = \frac{F_{00}^* + F_{11}^* + P_{22}}{F_{01}^* + F_{10}^* + C_{22}}$ where F_{ij}^* denote haplotype frequencies observed from all genotypes except 22. The best-fit values of C_{22} and P_{22} should minimize the sum of differences between expected and observed genotype frequencies. It is easy to compute the expected genotype frequencies for the random mating model, which unfortunately can significantly deviate from the real mating. Unknown for pairs of SNPs, such deviation can be accurately measured for individual SNPs. In

our algorithm, we assume that the deviation for pairs of SNPs is similar to the deviation observed for the corresponding single SNP. Therefore, the expected 2-SNP frequencies are adjusted proportionally to the observed single-site deviation.

3 Phasing of Complete Genotypes

Below we describe phasing of long genotype using certainty of cis- or trans-phasing between any pairs of 2's computed in the previous section. We then explain how we resolve missing data recovery and conclude with runtime analysis of the 2SNP algorithm.

Genotype Graph. For each genotype g , 2SNP constructs a *genotype graph*, which is a complete graph with vertices corresponding to 2's (i.e., heterozygous sites) of g . The weight of the edge between heterozygous sites i and j represents the certainty (formula 1) in that i and j are cis- or trans-phased. The maximum spanning tree of the genotype graph uniquely determines the phasing of the corresponding genotype since it gives cis-/trans- phasing for any two 2's. Obviously, if for any pair of 2's we know if they are cis- or trans-phased, then the entire phasing is known. Note that [10] have applied the same construction for preliminary estimation of haplotype frequencies rather than phasing *per se*. Therefore, for the edge weight, they have chosen LD-based formula over probabilities of *full* (i - j)-haplotypes given by maximum-likelihood solution. Instead, edge weights in 2SNP do not account for SNPs between i and j .

Missing data recovery. Missing data (?'s) are recovered after phasing of 2's. For each haplotype h we find the closest (w.r.t. Hamming distance) haplotype(s) h' and recover ?'s in h with the corresponding values from h' .

Runtime 2SNP Algorithm. The runtime of 2SNP algorithm has two bottlenecks. The first is computing of the observed haplotype frequencies for each pair of SNPs, which takes $O(nm^2)$ since we have n genotypes each with m SNPs. The second is recovering of the missing data, which needs $O(n^2m)$ runtime since it results in computing all pairwise Hamming distances between $2n$ haplotypes each with m SNPs. As a result, the total runtime of the algorithm is $O(nm(n + m))$, where n and m are the number of genotypes and SNPs, respectively.

4 Results

In this section we first describe the datasets and quality measures. Then, we compare our 2SNP method with PHASE-2.1.1[14], HAPLOTYPYPER[12] and GERBIL[10].

Data Sets. The comparison of phasing methods were performed on 46 real datasets from 79 different genomic regions and on 4 simulated datasets. All real datasets represent family trios – the computationally inferred offspring haplotypes for offspring have been compared with haplotypes inferred from parental genotypes.

- *Chromosome 5q31*: 129 genotypes with 103 SNPs derived from the 616 KB region of human Chromosome 5q31 [3].

- *Yoruba population (D)*: 30 genotypes with SNPs from 51 various genomic regions, with number of SNPs per region ranging from 13 to 114 [4].
- *HapMap datasets*: 30 genotypes of Utah residents and Yoruba residents available on HapMap by December 2005. The number of SNPs varies from 52 to 1381 across 40 regions including ENm010, ENm013, ENr112, ENr113 and ENr123 spanning 500 KB regions of chromosome bands 7p15:2, 7q21:13, 2p16:3, 4q26 and 12q12 respectively, and two regions spanning the gene STEAP and TRPM8 plus 10 KB upstream and downstream. Two more datasets with 60 genotypes each were obtained by mixing two populations YRI+CEU where SNPs with non-zero frequency in the both populations are kept.
- *Random matching 5q31*: 128 genotypes each with 89 SNPs from 5q31 cytokine gene cluster generated by random matching from 64 haplotypes of 32 West African reported by [9].
- *MS-simulated data*: 258 populations have been generated by MS[8] haplotype generator (using recombination rates 0,4 and 16). From each population of 100 haplotypes with 103 SNPs we have randomly chosen one haplotype and generated 129 genotypes by random matching.

Error measures. A *single-site error*[15] is the percent of erroneous SNPs among all SNPs in phased haplotypes. An *individual error*[12] is the percent of genotypes phased with at least one error among all genotypes. A *switching error*[10] is the percent of switches (among all possible switches) between inferred haplotypes necessary to obtain a true haplotype. For each dataset we bootstrapped phasing result 100 times, and for each bootstrap sample we computed an error. The 95% confidence interval for the error mean was computed based on the 100 error values.

Comparison of phasing methods. Table 1 shows performance of four phasing methods on real datasets, and Table 2 shows the performance for simulated datasets. All runs were performed on computer with Intel Pentium 4, 2.0Ghz processor and 2 Gigabytes of Random Access Memory. HAPLOTYPED was run with 20 rounds and the 2.1.1 version of PHASE was run with the default parameters. The *-marked switching error is 2.9% for the earlier version PHASE-2.0.2. The dashes in HAPLOTYPED columns correspond to the cases when it does not output valid phasing.

The tables show that 2SNP is several orders of magnitude faster than two other phasing methods handling large datasets in a matter of seconds. The reported mean errors with the respective 95% confidence intervals show that GERBIL, PHASE, and 2SNP have the same accuracy for real data (Chromosome 5q31, Yoruba(D), HapMap datasets). On the other hand, 2SNP and GERBIL are considerably outperformed by PHASE and HAPLOTYPED on some simulated data (Random matching 5q31). Poor performance of 2SNP can be caused by the absence of deviation from the Hardy-Weinberg equilibrium observed on real data. For the mixed populations one can see deterioration of all phasing methods.

In conclusion, we have presented a new extremely fast and simultaneously highly accurate phasing algorithm 2SNP based on 2-SNP haplotypes. We hope

Table 1. Mean single-site, individual, and switching errors with 95% confidence intervals and runtimes of PHASE, GERBIL, HAPLOTYPYPER (HTYPER), and 2SNP on real datasets.

Data	Measure	PHASE	GERBIL	HTYPER	2SNP
5q31, Daly et al.[3] # genotypes = 129 # SNPs = 103	single-site	1.9 ± 0.3	1.8 ± 0.3	2.2 ± 0.4	1.5 ± 0.3
	individual	25.9 ± 4.4	30.4 ± 5.2	29.2 ± 4.9	25.8 ± 4.5
	switching	$3.5^* \pm 0.1$	3.3 ± 0.1	4.0 ± 0.1	3.0 ± 0.1
	runtime	1.4×10^4	1.0×10^2	1.2×10^2	2.0×10^0
Gabriel et al.[4] average on 51bk # genotypes = 29 # SNPs = 50	single-site	2.7 ± 0.6	3.3 ± 0.7	4.4 ± 1.0	3.6 ± 0.8
	individual	26.6 ± 5.5	32.2 ± 6.7	34.7 ± 6.9	32.9 ± 7.0
	switching	13.8 ± 2.9	15.5 ± 3.2	23.0 ± 3.3	16.1 ± 3.4
	runtime	5.6×10^2	8.0×10^0	8.0×10^0	8.0×10^{-2}
ENm010.7p15:2[7] CEU # genotypes = 30 # SNPs = 1381	single-site	5.8 ± 0.4	6.1 ± 0.3	–	5.6 ± 0.2
	individual	50.0 ± 2.2	50.0 ± 2.5	–	50.0 ± 2.7
	switching	13.6 ± 0.2	11.9 ± 0.1	–	10.9 ± 0.2
	runtime	1.1×10^6	3.7×10^5	–	4.1×10^1
ENm010.7p15:2[7] CEU-nonrs # genotypes = 30 # SNPs = 459	single-site	0.4 ± 0.1	0.5 ± 0.1	0.6 ± 0.1	0.4 ± 0.1
	individual	40.0 ± 2.1	36.6 ± 1.1	43.3 ± 0.7	36.6 ± 1.8
	switching	19.2 ± 0.5	21.0 ± 0.6	28.7 ± 0.7	16.8 ± 0.6
	runtime	1.3×10^4	5.0×10^1	1.0×10^1	4.0×10^0
ENm010.7p15:2[7] YRI # genotypes = 30 # SNPs = 371	single-site	5.7 ± 0.7	5.7 ± 0.5	6.9 ± 1.1	5.5 ± 0.4
	individual	50.0 ± 1.1	50.0 ± 1.3	50.0 ± 0.4	50.0 ± 1.4
	switching	24.3 ± 0.2	23.1 ± 0.3	32.6 ± 0.2	22.7 ± 0.3
	runtime	4.5×10^4	1.3×10^3	4.9×10^1	3.0×10^0
ENr123.12q:12[7] CEU-nonrs # genotypes = 30 # SNPs = 1026	single-site	3.3 ± 0.1	2.4 ± 0.2	–	2.8 ± 0.1
	individual	39.7 ± 1.4	46.6 ± 3.0	–	36.6 ± 1.4
	switching	2.2 ± 0.1	1.0 ± 0.1	–	2.0 ± 0.1
	runtime	2.2×10^5	3.3×10^5	–	3.7×10^1
ENm013.7q21:13[7] YRI # genotypes = 30 # SNPs = 758	single-site	2.5 ± 0.2	2.3 ± 0.1	–	2.4 ± 0.2
	individual	41.6 ± 2.2	38.3 ± 1.8	–	36.6 ± 1.5
	switching	5.1 ± 0.2	3.7 ± 0.1	–	4.6 ± 0.2
	runtime	9.9×10^4	2.2×10^4	–	1.6×10^1
ENr113.4q26[7] CEU # genotypes = 30 # SNPs = 1017	single-site	3.8 ± 0.1	4.8 ± 0.3	–	4.6 ± 0.4
	individual	41.6 ± 2.8	43.3 ± 3.2	–	43.3 ± 3.2
	switching	2.3 ± 0.1	1.5 ± 0.1	–	1.6 ± 0.1
	runtime	5.0×10^5	6.1×10^5	–	3.9×10^1
ENr113.4q26[7] YRI # genotypes = 30 # SNPs = 885	single-site	5.5 ± 0.2	6.8 ± 0.3	–	6.5 ± 0.1
	individual	50.0 ± 2.7	50.0 ± 2.8	–	50.0 ± 2.6
	switching	6.0 ± 0.1	5.5 ± 0.1	–	6.0 ± 0.1
	runtime	2.1×10^5	1.3×10^5	–	2.5×10^1
ENr112.2p16:3[7] YRI # genotypes = 30 # SNPs = 1090	single-site	6.3 ± 0.2	5.1 ± 0.1	–	5.3 ± 0.1
	individual	50.0 ± 2.2	48.0 ± 2.8	–	50.0 ± 2.9
	switching	8.0 ± 0.2	3.9 ± 0.1	–	4.6 ± 0.1
	runtime	5.6×10^5	3.3×10^5	–	4.2×10^1
ENCODE project[7] average over 40 datasets	single-site	3.4 ± 0.1	3.5 ± 0.1	–	3.4 ± 0.1
	individuals	43.4 ± 1.9	44.9 ± 2.0	–	43.6 ± 1.9
	switching	10.2 ± 1.8	10.1 ± 1.9	–	9.7 ± 1.7
	runtime	3.0×10^5	1.7×10^5	–	2.7×10^1

Data	Measure	PHASE	GERBIL	HTYPER	2SNP
TRPM8[7]	single-site	2.4 ± 0.1	3.3 ± 0.2	2.1 ± 0.1	2.6 ± 0.2
CEU	individual	35.0 ± 2.3	35.0 ± 1.7	31.7 ± 1.8	28.0 ± 5.6
# genotypes = 30	switching	2.8 ± 0.1	6.5 ± 0.2	2.3 ± 0.1	1.7 ± 0.1
# SNPs = 315	runtime	8.8×10^3	1.4×10^3	2.3×10^1	3.0×10^0
TRPM8[7]	single-site	2.8 ± 0.2	4.4 ± 0.4	4.7 ± 0.5	3.6 ± 0.4
YRI	individual	36.6 ± 1.8	48.3 ± 2.2	43.3 ± 2.3	43.3 ± 2.9
# genotypes = 30	switching	5.0 ± 0.2	6.8 ± 0.2	11.1 ± 0.2	6.4 ± 0.2
# SNPs = 290	runtime	9.7×10^3	9.8×10^2	2.7×10^1	3.0×10^0
STEAP[7]	single-site	0.3 ± 0.1	0.6 ± 0.2	0.5 ± 0.1	0.6 ± 0.2
YRI	individual	6.6 ± 0.6	13.3 ± 0.7	11.6 ± 0.7	13.3 ± 0.6
# genotypes = 30	switching	5.2 ± 0.6	7.9 ± 0.4	5.9 ± 0.5	7.8 ± 0.6
# SNPs = 52	runtime	1.2×10^2	2.0×10^0	4.0×10^0	1.0×10^{-3}
STEAP[7]	single-site	0.6 ± 0.1	1.0 ± 0.2	1.1 ± 0.2	1.0 ± 0.1
CEU	individual	6.6 ± 0.7	10.0 ± 0.9	11.7 ± 1.0	8.3 ± 1.2
# genotypes = 30	switching	7.3 ± 0.9	11.1 ± 1.0	11.6 ± 1.0	11.5 ± 1.1
# SNPs = 60	runtime	1.3×10^2	3.0×10^0	5.0×10^0	1.0×10^{-3}
STEAP[7]	single-site	1.3 ± 0.1	2.1 ± 0.2	2.0 ± 0.2	1.9 ± 0.2
YRI+CEU	individuals	9.2 ± 0.9	15.8 ± 1.3	15.0 ± 1.2	15.8 ± 1.2
# genotypes = 60	switching	15.1 ± 1.0	24.6 ± 1.0	22.2 ± 0.9	22.9 ± 1.0
# SNPs = 49	runtime	3.2×10^2	5.0×10^0	1.2×10^1	1.0×10^{-3}
TRPM8[7]	single-site	4.0 ± 0.9	5.6 ± 1.1	6.5 ± 0.9	5.2 ± 1.0
YRI+CEU	individuals	43.3 ± 1.2	47.5 ± 1.7	47.5 ± 1.6	46.6 ± 1.7
# genotypes = 60	switching	18.7 ± 0.3	28.6 ± 0.3	36.1 ± 0.3	27.7 ± 2.5
# SNPs = 231	runtime	3.7×10^4	4.5×10^2	4.7×10^1	2.0×10^0

that it will be very useful for high-throughput genotype data processing, e.g., SNP Mapping Arrays (Affymetrix). We are going to extend our method by applying 3-SNP haplotype analysis.

References

1. Brinza, D. and Zelikovsky, A. (2006) 2SNP: Scalable Phasing Based on 2-SNP Haplotypes. *Bioinformatics*, **22(3)**, 371–374.
2. Clark, A. (1990) Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol Biol Evol.*, **7**, 111–122.
3. Daly, M., Rioux, J., Schaffner, S., Hudson, T. and Lander, E. (2001) High resolution haplotype structure in the human genome. *Nat Genet.*, **29**, 229–232.
4. Gabriel, G., Schaffner, S., Nguyen, H., Moore, J., Roy, J., Blumenstiel, B., Higgins, J., et al. (2002) The structure of haplotype blocks in the human genome. *Science*, **296**, 2225–2229.
5. Gusfield, D. (2003) Haplotype inference by pure parsimony. *Proc. Symp. on Comb. Pattern Matching*, LNCS **2676**, 144–155.
6. Halperin, E. and Eskin, E. (2004) Haplotype Reconstruction from Genotype Data using Imperfect Phylogeny. *Bioinformatics*, **20**, 1842–1849.

Table 2. Mean single-site, individual, and switching errors with 95% confidence intervals and runtimes of PHASE, GERBIL, HAPLOTYPYPER (HTYPER), and 2SNP on simulated datasets.

Data	Measure	PHASE	GERBIL	HTYPER	2SNP
random mating	single-site	3.9 ± 0.1	8.7 ± 0.7	2.8 ± 0.5	9.1 ± 0.5
Hull et al.[9]	individual	25.0 ± 4.1	47.2 ± 6.9	20.9 ± 3.1	46.4 ± 6.8
# genotypes = 128	switching	4.1 ± 0.1	12.1 ± 0.1	4.9 ± 0.1	11.8 ± 0.1
# SNPs = 89	runtime	3.9×10^4	4.7×10^1	2.4×10^1	2.0×10^0
MS[8]	single-site	0.1 ± 0.1	0.1 ± 0.1	0.5 ± 0.1	0.2 ± 0.1
recomb rate=0	individual	1.6 ± 0.1	2.7 ± 0.2	5.7 ± 0.3	4.6 ± 0.3
# genotypes = 100	switching	0.7 ± 0.1	1.2 ± 0.1	3.3 ± 0.3	2.1 ± 0.2
# SNPs = 103	runtime	1.5×10^2	6.0×10^0	3.0×10^0	1.0×10^0
MS[8]	single-site	0.2 ± 0.1	0.3 ± 0.1	0.3 ± 0.1	0.5 ± 0.2
recomb rate=4	individual	2.0 ± 0.1	3.66 ± 0.2	6.5 ± 0.2	6.6 ± 0.3
# genotypes = 100	switching	0.6 ± 0.1	1.4 ± 0.2	2.8 ± 0.2	3.4 ± 0.4
#SNPs = 103	runtime	1.2×10^2	6.0×10^0	3.0×10^0	1.0×10^0
MS[8]	single-site	0.4 ± 0.1	0.5 ± 0.1	0.4 ± 0.1	1.0 ± 0.1
recomb rate=16	individual	6.7 ± 0.5	6.3 ± 0.3	5.3 ± 0.3	9.8 ± 0.4
# genotypes = 100	switching	1.7 ± 0.2	2.1 ± 0.2	2.6 ± 0.2	3.2 ± 0.2
# SNPs = 103	runtime	1.3×10^2	6.0×10^0	3.0×10^0	1.0×10^0

7. International HapMap Consortium. (2003) The International HapMap Project. *Nature*, **426**, 789–796, <http://www.hapmap.org>.
8. Hudson, R. (1990) ‘Gene genealogies and the coalescent process’, *Oxford Survey of Evolutionary Biology*, **7**, 1–44.
9. Hull, J., Rowlands, K., Lockhart, E., Sharland, M., Moore, C., Hanchard, N., Kwiatkowski, D.P. (2004) Haplotype mapping of the bronchiolitis susceptibility locus near IL8. *Am J Hum Genet.*, **114**, 272–279
10. Kimmel, G. and Shamir, R. (2005) GERBIL: Genotype resolution and block identification using likelihood. *Proc Natl Acad Sci.*, **102**, 158–162.
11. Kruglyak, L. and Nickerson, D. A. (2001) Variation is the spice of life. *Nat Genet.*, **27**, 234–236.
12. Niu, T., Qin, Z., Xu, X. and Liu, J.S. (2002) Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *Am J Hum Genet.*, **70**, 157–169.
13. Niu T. (2004) Algorithms for inferring haplotypes. *Genet Epidemiol.*, **27**(4), 334–47.
14. Stephens, M., Smith, N. and Donnelly, P. (2001) A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet.*, **68**, 978–989.
15. Stephens, M., and Donnelly, P. (2003) ‘A Comparison of Bayesian Methods for Haplotype Reconstruction from Population Genotype Data’, *Am. J. Human Genetics*, **73**:1162–1169.