
Family Trio Phasing and Missing Data Recovery*

Dumitru Brinza, Jingwu He[†], Weidong Mao and
Alexander Zelikovsky[‡]

Department of Computer Science, Georgia State University, 34 Peachtree
Str., suite 1450, Atlanta, GA 30303

Phone: (404) 463-2888

Fax: (404) 463-9912

{dima,jingwu,weidong,alexz}@cs.gsu.edu

*Authors are listed in alphabetical order.

[†]The work of J. He has been supported by Georgia State University Molecular Basis of Disease Fellowship.

[‡]Corresponding author. The work of A. Zelikovsky has been partially supported by NIH Award 1 P20 GM065762-01A1.

Abstract: Although there exist many phasing methods for unrelated adults or pedigrees, phasing and missing data recovery for data representing family trios is lagging behind. This paper is an attempt to fill this gap by considering the following problem. Given a set of genotypes partitioned into family trios, find for each trio a quartet of parent/offspring haplotypes explaining each trio without recombinations and recovering the SNP values missed in given genotype data. Our contributions include (i) formulating the pure-parsimony trio phasing without recombinations and the trio missing data recovery problems, (ii) proposing new greedy and integer linear programming based solution methods, and (iii) extensive experimental validation of proposed methods showing advantage over the previously known methods.

Keywords: haplotypes, genotypes, SNP, family trio data, phasing.

Reference to this paper should be made as follows: Brinza, D., He, J., Mao, W. and Zelikovsky, A. (xxxx) 'Phasing and Missing data recovery in Family Trios', Int. J. of Bioinformatics Research and Applications, Vol. x, No. x, pp.xxx-xxx.

Biographical notes:

Dumitru Brinza received B.S in Mathematics at Moldova State University. He is a Ph.D. candidate at the Department of Computer Science, Georgia State University. His research interests include computational biology, bioinformatics, ad-hoc and sensors networks.

Jingwu He received M.S in Chemistry at Georgia Institute of Technology and M.S Computer Science at Georgia State University. He is a PhD candidate in Computer Science at Georgia State University. His research interests include computational biology, bioinformatics, algorithms and networks. He is author of research studies published at international journals and conference proceedings.

Weidong Mao received M.S in Computer Science at the Wuhan University of Technology in China. He is a Ph.D candidate at the Department of Computer Science, Georgia State University. His research interests include bioinformatics, haplotype phasing and disease association. He has published several national and

international conference proceedings.

Alexander Zelikovsky received the Ph.D. degree in Computer Science from the Institute of Mathematics of the Belorussian Academy of Sciences in Minsk (Belarus) in 1989 and worked at the Institute of Mathematics in Kishinev (Moldova) (1989-1995). He was a Research Scientist at University of Virginia (1995-1997) and a Postdoctoral Scholar at UCLA (1997-1998). Dr. Zelikovsky is an Associate Professor at Computer Science Department of Georgia State University which he joined in 1999. He is the author of more than 100 refereed publications. Dr. Zelikovsky's research interests include computational biology, discrete algorithms, ad-hoc wireless networks and VLSI physical layout design.

1 Introduction

In diploid organisms (such as human) each chromosome has two "copies" which are not completely identical. Each of two single copies is called a haplotype, while a description of the data consisting of mixture of the two haplotypes is called a genotype. The underlying data that forms a haplotype is either the full DNA sequence in the region, or more commonly the values of single nucleotide polymorphisms (SNPs) in that region. A SNP is a single nucleotide site where two or more (out of four) different nucleotides occur in a large percentage of the population.

In general, it is costly and time consuming to examine the two copies of a chromosome separately, and genotype data rather than haplotype data are only available, even though it is the haplotype data that will be of greatest use for investigating complex diseases. Data from m sites (SNPs) in n individual genotype are collected, where each site has one of two alleles denoted by 0 and 1 (the case of three- and four-allelic SNPs can be reduced to the biallelic case). The input to the phasing problem consists of n genotype vectors, each of length m , where each value in the vector is either 0, 1, or 2. The position in the genotype vector has a value of 0 or 1 if the associated chromosome site has that allele on both copies (it is a homozygous site), and has a value of 2 otherwise (the chromosome site is heterozygous).

Usually, genotype data miss certain SNP values due to failures during genotyping. Although the missing data rate decreases, still unavailable SNP values constitute a substantial part of the entire data (as large as 16% of the genotype data in Daly et al. [4] data and 10% in Gabriel et al. [5]). The problem of missing data recovery attracts a great deal of attention [9, 14].

Commonly, genotype data represent family trios consisting of the two parents and their offspring since that allows to recover haplotypes with higher confidence. A simple logical analysis allows to substantially decrease uncertainty of phasing. For example, for two SNPs in a trio with parent genotypes $f = 22$ and $m = 21$, and the offspring genotype $k = 02$, there is a feasible phasing of the parents: $f_1 = 00$, $f_2 = 11$, $m_1 = 01$, $m_2 = 11$ such that the haplotypes $k_1 = 00$ and $k_2 = 01$ are transmitted to the offspring by parents f and m , respectively. In fact, there is another feasible phasing of the parent f , i.e., $f_1 = 01$, $f_2 = 10$, but then it assumes a crossover between these two sites in the transmitted haplotype k_1 . It is not difficult to check that logical ambiguity in offspring phasing exists only if all three genotypes have 2's in the same SNP site.

This paper deals with problems of phasing and missing data recovery on family trios data assuming no recombinations in parents. Although there exist many phasing methods

for unrelated adults or pedigrees, phasing and missing data recovery for trios is lagging behind. Unfortunately, all well-known computational methods for phasing Daly et al. [4] family trio data introduce recombinations in the transmitted haplotypes and do not allow to set the 0 recombination rate.

Formally, given a set of genotypes partitioned into family trios, the Trio Phasing Problem without Recombinations (TPPWR) requires to find for each trio a quartet of parent/offspring haplotypes which agree with all three genotypes. Trio Missing Data Recovery Problem (TMDRP) asks for the SNP values missed in given genotype data.

Our contributions include

- Formulating the Pure-parsimony Trio Phasing Problem (PTPPWR) and the Trio Missing Data Recovery Problem (TMDRP).
- Proposing two new greedy and integer linear programming (ILP) based methods solving PTPPWR and TMDRP.
- Extensive experimental validation of proposed methods and comparison with the previously known methods.

The rest of the paper is organized as follows. The next section summarizes the previous works on phasing and missing data recovery in family trio data. In Section 3 we describe pure-parsimony trio phasing problem without recombinations and propose greedy and ILP based method to solve it. In Section 4 we present our experimental study of the suggested methods on simulated and real data.

2 Previous Work

In this section we overview previous research and on phasing based on statistical methods (Phamily and PHASE, HAPLOTYPYER, HAP and greedy algorithms). The ILP based approaches will be discussed in the next section.

Stephens et al. [16] introduced a Bayesian statistical method PHASE for phasing genotype data. It exploits ideas from population genetics and coalescent theory that make phased haplotypes to be expected in natural populations. It also estimates the uncertainty associated with each phasing. The software can deal with SNP in any combination, any size of population and missing data are allowed. The drawback of this method is that it takes long time for large population

Acherman et al. [1] described the tool Phamily for phasing the trio families based on well-known phasing tool PHASE [16]. It first uses the logical method described above to infer the SNPs in the parental haplotypes. Then offspring genotypes are discarded while the parental genotypes and known transmitted haplotypes are passed to PHASE. Unfortunately, Phamily does not have an option forbidding recombinations in transmitted haplotypes.

Niu et al [15] proposed a new Monte Carlo approach HAPLOTYPYER for phasing genotype data. It first partition the whole haplotype into smaller segments then use the Gibbs sampler both to construct the partial haplotypes of each segment and to assemble all the segments together. This method can accurately and rapidly infer haplotypes for a large number of linked SNPs. The drawback of HAPLOTYPYER is that it can not handle lengthy genotype with large population. It is limited to 100 SNPs and 500 population.

Halperin et al. [9] used the greedy method for phasing and missing data recovery. For each trio the author introduce four partially resolved haplotypes with the coordinates 0, 1

and ?. The values of 0 and 1 correspond to fully resolved SNPs which can be found via logical resolution from the section 1, while the ?'s corresponds to ambiguous and missing positions. The greedy algorithm iteratively finds the complete haplotype which covers the maximum possible number of partial haplotypes, removes this set of resolved partial haplotypes and continues in that manner. The authors replace each genotype in Daly et al [4] data with a pair of logically partial resolved haplotypes referring to each ambiguous SNP value as a ?. The ?'s constitute 16% of all data. Then extra 10% of data are erased (i.e., replaced with ?'s) and the resulted 26% of ambiguous SNP values are inferred by the greedy algorithm minimizing haplotype variability within blocks. When measured on the additionally erased 10% of data, the error rate for the greedy algorithm is 2.8% [9] which has been independently confirmed in our computational experiments. The resulted phasing also introduces considerable amount of crossovers in the parental haplotypes transmitted to offspring.

In [13] for haplotyping pedigree data, the objective is to minimize recombinations. That objective is not suitable for TPPWR since there always exists phasing with no recombinations. Note that it is easy to find a feasible solution to TPPWR but the number of feasible solutions is exponential and it is necessary to choose a criteria for comparing such solutions.

3 Pure-Parsimony Trio Phasing Without Recombinations

Everywhere further in this section, we assume that there is no recombinations in the haplotypes transmitted to the offspring. As noted above the number of feasible solutions is exponential and we need to choose an objective. Following [3, 6] we pursue parsimonious objective, i.e., minimization of the total number of haplotypes while forbidding recombinations.

The drawback of pure parsimony is that when the number of SNPs becomes large (as well as the number of recombinations), then the quality of pure parsimony phasing is diminishing [6]. Therefore, following the approach in [7], we suggest to partition the genotypes into blocks, i.e., substrings of bounded length, and find solution for the pure parsimony problem for each block separately. Note that in case of family trios we have great advantage over the method of [7] since we do not need to solve the problem of joining blocks. Indeed, for each family trio we can make four haplotype templates (partially resolved by logic means of haplotypes) that imply unique way of gluing together blocks to arrange complete haplotypes for the entire sequence of SNPs.

3.1 Problem Formulation

Formally, let genotype be a vector with m coordinates each corresponding to an SNP and having one of the following values: 0 (homozygote with major allele), 1 (homozygote with minor allele), 2 (heterozygote), or ? (missing SNP value). Let haplotype be a vector with m coordinates where each coordinate is either 0 or 1. We say that two haplotypes explain a genotype if

- for any 0 (resp. 1) in the genotype vector, the corresponding coordinates in the both haplotype vectors are 0's (resp. 1's),
- for any 2 in the genotype vector, the corresponding coordinates in the two haplotype vectors are 0 and 1,

- for any ? in the genotype vector, the corresponding coordinates in the haplotypes are unconstrained (can be arbitrary).

We say that four haplotypes h_1, h_2, h_3, h_4 explain a family trio of genotypes (f, m, k) , if h_1 and h_2 explain the genotype f , h_3 and h_4 explain the genotype m , and h_1 and h_3 explain the genotype k .

Pure-Parsimony Trio Phasing Without Recombinations (PPTPWR). Given $3n$ genotypes corresponding to n family trios find minimum number of distinct haplotypes explaining all trios without recombinations.

3.2 Greedy Method for Trio Phasing

We apply the greedy algorithm from Halperin [9] for trio phasing. For each trio we introduce four partial haplotypes with the coordinates 0, 1 and ?. The values of 0 and 1 correspond to fully resolved SNPs which can be found via logical resolution from the section 1, while the ?'s corresponds to ambiguous and missing positions. The greedy algorithm iteratively finds the complete haplotype which covers the maximum possible number of partial haplotypes, removes this set of resolved partial haplotypes and continues in that manner. The drawback of this greedy method is that its result in general cannot be explained without recombinations. In the future, we will try to modify the greedy algorithm to overcome this shortcoming.

3.3 Integer Linear Program for Trio Phasing Without Recombinations

We have implemented the following integer linear program (ILP) formulation (1)-(4) for pure-parsimony trio phasing. It uses 0-1 variable x_i for each possible haplotype with the minimization objective:

$$\text{Minimize } \sum x_i \tag{1}$$

For each trio we introduce four template haplotypes, i.e., haplotypes with the coordinates 0,1,2 and ?: 0's and 1's correspond to fully resolved SNPs, 2's come in pairs corresponding to the genotype 2's and ?'s correspond to unconstrained SNPs. For each 2 in each template we introduce a 0-1-variable y and a constraint binding it with the variable y' corresponding to the complimentary 2:

$$y + y' = 1 \tag{2}$$

Instead of completely resolving templates, we can resolve only 2's. Then several haplotypes can fit partially resolved templates and at least one of the corresponding x -variables should be set to 1. This results in the following constraint: For any y -assignment of 2's in each template T ,

$$\sum_{x \text{ fits all } y\text{'s in template } T} x \geq 1 + \sum_{i \in I_1} y_i - |I_1| + \sum_{i \in I_2} (1 - y_i) - |I_2| \tag{3}$$

We should guarantee that each template is resolved. This is done by the following constraint: For each template T ,

$$\sum_{x \text{ fits template } T} x \geq 1 \tag{4}$$

4 Experimental Results

In this section we compare our greedy and ILP based methods suggested in Section 3 with previously known phasing methods such as Phamily [1], PHASE[16] and HAPLOTYPER[15] applied to phasing and missing recovery on family trio data. We first describe the test data sets then give experimental results of five methods for phasing and then for missing data recovery of family trio data.

4.1 Test Data Sets

Our algorithms are evaluated on real and simulated data. The data set collected by Daly et al. [4] is derived from the 616 kilobase region of human Chromosome 5q31. Another real data set is collect by Gabriel et al. [5]. This data consists of genotypes of SNPs from 62 region. The both data sets contain about 10% missing data.

The simulated data is generated using ms [10], a well-known haplotype generator based on the coalescent model of SNP sequence evolution. The ms generator emits a haplotype population for the given number of haplotypes, number of SNPs, and the recombination rate. We have simulated Daly et al. [4] data by generating 258 populations, each population with 100 individuals and each haplotype with 103 SNPs, then randomly choosing one haplotype from each population. We only simulate parents's haplotypes, then we obtain offspring haplotypes by random matching the parental haplotypes thus obtaining family trios without recombinations in transmitted haplotypes.

4.2 Family Trio Phasing Validation

It is clear how to validate a phasing method on simulated data since the underlying haplotypes are known. The validation on real data is usually performed on the trio data. E.g., a phasing method is applied to parents (respectively, to offspring) genotypes and the resulted haplotypes are validated on offspring's (respectively, on parents') genotypes. Unfortunately, in our case, one can not apply such validation since a trio phasing method may rely on both offspring and parents' genotypes. Therefore, we suggest to validate trio phasing by erasing randomly chosen SNP values and recording the errors in the erased SNP sites. In Tables 1, 2, 3, each row corresponds to an instance of real data (Daly et al. or Gabriel et al.) or simulated data (ms) and the column (E) shows the percent of erased data (0% - no data erased, 1%-10% - percent of SNP values erased) .

The value of phasing errors is measured by the Hamming distance from the method's solution to the closest phasing without recombinations. In Tables 1 and 2, for parents (P) we report the percent of SNP values that should be inverted out of the total number of SNP values that should be inferred (i.e., number of 2 plus number of unknown values). For offspring (C), we report the percent of SNP which should be inverted with respect to the total number of SNPs. The total number of errors (T) is the percent of SNP's that should be inverted in order to obtain a feasible phasing solution.

In Table 2, we also report true error for phasing simulated genotype data which is the Hamming distance between inferred and actual simulated underlying haplotypes for offspring (C), for parents (P) and the total error (T).

Table 1 The results for five phasing methods on the real data sets of Daly et al.[4] and Gabriel et al. [5] and on simulated data. The second column corresponds to the ratio of erased data. The C corresponds to the error of offspring. The P corresponds to the error of parents. The T corresponds to the total error.

Data	E	ILP			Greedy			Phamily			PHASE			HAPLOTYPYER		
		C	P	T	C	P	T	C	P	T	C	P	T	C	P	T
Daly et al. [4]	0	0.0	0.0	0.0	4.9	16.2	3.8	1.3	0.0	0.7	1.1	0.0	0.6	2.2	0.0	1.2
	1	0.2	0.5	0.2	4.8	16.8	3.8	1.2	1.4	0.7	1.3	0.2	0.7	2.1	1.0	1.6
	2	0.3	0.7	0.4	5.0	16.9	4.0	1.3	1.8	0.9	1.3	0.5	0.8	2.2	2.3	1.7
	5	0.8	2.6	1.2	5.3	17.1	4.0	1.3	1.0	1.0	1.6	0.9	1.0	2.3	7.0	2.9
	10	1.8	6.7	3.0	5.9	17.2	4.7	1.5	2.2	1.3	1.5	1.9	1.2	2.6	9.8	4.1
Gabriel et al. [5]	0	0.0	0.0	0.0	2.9	11.5	2.2	3.0	0.0	2.0	2.2	0.0	1.3	4.4	0.0	2.7
	1	0.2	0.6	0.2	2.9	12.1	2.3	3.1	0.2	2.0	2.8	0.2	1.7	4.6	1.7	1.5
	2	0.3	1.2	0.5	3.2	12.2	2.4	3.3	0.4	2.1	2.9	0.6	1.8	4.9	3.1	1.6
	5	0.8	3.4	1.1	3.4	12.2	2.9	3.4	1.3	2.5	3.0	1.4	1.6	5.4	6.3	2.1
	10	1.5	6.2	1.5	4.3	12.4	3.7	3.9	2.4	2.5	3.3	3.1	2.1	6.1	15.7	6.3
ms [10]	0	0.0	0.0	0.0	2.6	13.2	1.9	9.4	0.0	4.7	5.6	0.0	6.5	8.1	0.0	5.4
	1	0.3	1.0	0.4	2.9	13.5	1.9	10.1	0.8	4.3	5.8	1.2	5.4	8.4	2.2	5.6
	2	0.5	1.9	0.7	3.1	13.7	2.1	10.4	1.8	7.8	5.9	2.3	5.5	8.9	4.3	6.0
	5	1.3	3.8	1.9	4.3	13.9	3.1	10.6	3.8	7.6	6.1	4.7	5.9	9.2	10.2	7.0
	10	2.5	7.7	3.6	5.3	14.0	4.4	11.9	9.5	9.2	6.9	10.5	6.0	11.5	17.1	8.0

Table 2 The results for five phasing methods on simulated data sets. The column E represents the percent of erased data. The C corresponds to the true error of offspring. The P corresponds to the true error of parents. The T corresponds to the true total error.

Data	E	ILP			Greedy			Phamily			PHASE			HAPLOTYPYER		
		C	P	T	C	P	T	C	P	T	C	P	T	C	P	T
ms [10]	0	1.2	1.3	1.3	1.4	1.4	1.4	2.1	2.2	2.2	3.3	3.2	3.2	2.9	2.7	2.8
	1	1.3	1.3	1.3	1.3	1.4	1.4	4.5	4.0	4.3	3.2	3.3	3.2	3.0	3.2	3.1
	2	1.5	1.6	1.6	1.6	1.6	1.6	4.4	4.3	4.4	3.4	3.3	3.4	3.2	3.3	3.3
	5	2.2	2.5	2.4	2.1	2.3	2.2	4.3	4.2	4.3	3.6	3.5	3.5	3.4	3.7	3.6
	10	3.0	3.7	3.5	3.3	3.3	3.3	5.2	5.2	5.2	3.1	3.0	3.0	3.9	4.2	4.1

4.3 Missing Data Recovery in Family Trios

Table 3 compares five methods (ILP, Greedy, Phamily, PHASE and HAPLOTYPYER) on trio missing data recovery on the real data sets (Daly [4] and Gabriel [5]) and simulated data. We erase random data in trio genotypes with certain amount(1%, 2%, 5% and 10%) of the entire data. We report the error as the number of incorrectly recovered erased positions of the genotypes on offspring (C*), parents (P*) and trios (T*) divided the total number of erased positions in parent genotypes in percentage. We count only half error if the compared paired SNP is 2 and 0 (or 1).

5 Conclusions & Future Work

In this paper we have formulated the pure-parsimony trio phasing without recombinations and the trio missing data recovery problems, proposed two new greedy and integer linear programming based solution methods, and experimentally validated of proposed methods showing advantage over the previously known methods. It has been shown that the simple greedy algorithm is very stable in missing data recovery while the ILP method is superior in trio data phasing. Our future work will include design of new methods which will combine the feasibility and stability of ILP and greedy methods.

Table 3 The results for missing data recovery on the real and simulated data sets with five methods. The second column corresponds to the ratio of erased data. The C* corresponds to the error of offspring. The P* corresponds to the error of parents. The T* corresponds to the total error.

Data	E	ILP			Greedy			Phamily			PHASE			HAPLOTYPYER		
		C*	P*	T*	C*	P*	T*	C*	P*	T*	C*	P*	T*	C*	P*	T*
Daly et al. [4]	1	2.3	7.8	5.7	3.9	6.0	5.2	0.3	2.3	1.5	0.3	3.1	2.0	1.9	26.1	16.7
	2	3.1	8.6	6.5	4.0	6.0	5.2	0.2	4.7	3.0	0.2	3.7	2.4	1.7	24.5	15.9
	5	3.9	9.9	7.8	4.5	4.8	4.7	0.2	3.6	2.5	0.1	3.4	2.3	1.3	20.5	13.9
	10	5.7	13.5	10.8	4.6	5.8	5.4	0.6	4.4	3.1	0.5	4.0	2.8	1.5	21.8	14.8
Gabriel et al. [5]	1	7.7	8.0	7.9	5.6	6.4	6.1	0	2.5	1.6	0.4	3.1	2.1	1.6	21.8	14.5
	2	7.1	8.6	8.1	4.9	5.7	5.5	0	2.8	1.9	0.5	3.1	2.2	1.0	20.7	14.1
	5	7.9	8.7	8.4	5.6	5.8	5.7	0	2.3	1.5	0.1	3.3	2.2	2.5	20.7	14.6
	10	7.4	9.5	8.8	6.1	6.6	6.5	0.1	2.1	1.5	0.3	3.1	2.1	2.3	25.1	17.5
ms [10]	1	10.9	13.3	12.4	11.5	9.2	10.1	1.0	16.0	10.2	0.7	15.2	9.6	4.3	26.4	17.9
	2	11.4	12.3	11.9	11.2	8.6	9.6	1.7	15.3	10.3	0.3	15.6	10.0	4.6	20.6	14.7
	5	13.1	12.1	12.4	12.3	7.8	9.3	0.9	14.8	10.0	0.7	14.9	10.0	3.6	23.1	16.4
	10	12.0	12.4	12.3	11.6	8.9	9.8	2.3	14.4	10.3	0.7	13.9	9.3	3.4	21.9	15.5

References and Notes

- Ackerman, H. et al (2003) 'Haplotypic analysis of the TNF locus by association efficiency and entropy', *Genome Biology*, 4:R24.
- Brown, D.G. and Harrower, I.M. (2004) 'A new integer programming formulation for the pure parsimony problem in the haplotype association', *Workshop on Algorithms in Bioinformatics*, v.3240, 3-540-23018-1
- Clark, A. (1990) 'Inference of haplotypes from PCR-amplified samples of diploid populations', *Mol. Biol., Evol.*, 7:111–122.
- Daly, M., Rioux, J., Schaffner, S., Hudson, T. and Lander, E. (2001) 'High resolution haplotype structure in the human genome', *Nature Genetics*, 29:229–232.
- Gabriel, G., Schaffner, S., Nguyen, H., Moore, J., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E., Daly, M. and Altshuler, D. (2002) 'The structure of haplotype blocks in the human genome', *Science*, 296:2225—2229.
- Gusfield, D. (2003) 'Haplotype inference by pure parsimony', In R. Baeza-Yates, E. Chavez, and M. Chrochemore, ed. 14'th Annual Symposium on Combinatorial Pattern Matching, v. 2676 of Springer LNCS, 144–155.
- Halperin, E. and Eskin, E. (2004) 'Haplotype reconstruction from genotype data using imperfect phylogeny', *Bioinformatics*, 20(12):1842-9.
- Halperin, E. and Karp, R.M. (2003) 'Large Scale Reconstruction of Haplotypes from Genotype Data', *International Conference on Research in Computational Molecular Biology*, 104–113.
- Halperin, E. and Karp, R.M. (2004) 'Perfect phylogeny and haplotype assignment', *International Conference on Research in Computational Molecular Biology*, 1-58113-755-9.
- Hudson, R. (1990) 'Gene genealogies and the coalescent process', *Oxford Survey of Evolutionary Biology*, 7:1–44.
- He, J. and Zelikovsky, A. (2004) 'Linear Reduction for Haplotype Inference', *Proc. Workshop on Algorithms in Bioinformatics*, September 2004, *Lecture Notes in Bioinformatics*, 3240:242-253.
- He, J. and Zelikovsky, A. (2004) 'Linear Reduction Methods for Tag SNP Selection', *Proc. International Conf. of the IEEE Engineering in Medicine and Biology*, 2840-2843.
- Li, J. and Jiang, J. (2003) 'Efficient Rule-Based Haplotyping Algorithm for Pedigree Data. In Proc.', *International Conference on Research in Computational Molecular Biology*, 197-206

- 14 Lin, S., Chakravarti, A. and Cutler, D.J. (2004) 'Haplotype and Missing Data Inference in Nuclear Families', *Genome Res*,14(8):1624-32.
- 15 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.
- 16 Stephens, M., Smith, N. and Donnelly, P. (2001) 'A new statistical method for haplotype reconstruction from population data', *Am. J. Human Genetics*, 68:978-989.