Highly Scalable Genotype Phasing by Entropy Minimization

Bogdan Paşaniuc and Ion Măndoiu

***Abstract*— A *Single Nucleotide Polymorphism (SNP)* is a position in the genome at which two or more of the possible four nucleotides occur in a large percentage of the population. SNPs account for most of the genetic variability between individuals, and mapping SNPs in the human population has become the next high-priority in genomics after the completion of the Human Genome project. In diploid organisms such as humans, there are two non-identical copies of each autosomal chromosome. A description of the SNPs in a chromosome is called a haplotype. At present, it is prohibitively expensive to directly determine the haplotypes of an individual, but it is possible to obtain rather easily the conflated SNP information in the so called genotype. Computational methods for genotype phasing, i.e., inferring haplotypes from genotype data, have received much attention in recent years as haplotype information leads to increased statistical power of disease association tests. However, existing algorithms have impractical running time for phasing large genotype datasets such as those generated by the international HapMap project. In this paper we propose a highly scalable algorithm based on entropy minimization. Our algorithm is capable of phasing genotype data coming from either unrelated individuals or families consisting of a child and one or both parents. Experimental results show that our algorithm achieves a phasing accuracy close to that of best existing methods while being several orders of magnitude faster.**

I. Introduction

After the completion of the Human Genome Project has provided us with a blueprint of the DNA present in each human cell, genomics research is now focusing on the study of DNA variations that occur between individuals and understanding how these variations confer susceptibility to common diseases such as diabetes or cancer. The most common form of genomic variation are the so called *single nucleotide polymorphisms* (SNPs), i.e., the presence of different DNA nucleotides, or *alleles*, at certain chromosomal locations. Over 9 million common SNPs have already been catalogued in the dbSNP database maintained by NCBI.

In diploid organisms such as humans, there are two non-identical copies of each autosomal chromosome, one inherited from the mother and one inherited from the father. The combinations of SNP alleles in the maternal and paternal chromosomes are referred to as the individual's *haplotypes*. Although it is possible to directly determine the haplotypes of an individual by experimental techniques, such methods are prohibitively expensive and time consuming. In contrast, there are many cost-effective high-throughput techniques for determining the conflated SNP information called *genotype*, which specifies the identities of the two alleles at each SNP position, but does not assign the alleles to specific chromosomes for *heterozygous* SNP positions i.e., SNP positions at which the individual has two different alleles.

Since haplotypes determine the exact sequence (and hence function) of proteins encoded by the genes, finding the haplotypes in human populations is an important step in determining the genetic basis of complex diseases. For this reason, computational inference of haplotypes from genotype date, known as the *genotype phasing problem*, has received an increasing amount of attention in the literature in the past few years, see, e.g., [1], [2], [3], [4] for recent surveys. While many of the existing methods achieve high haplotype reconstruction accuracy, their runtime does not scale well with the number of SNPs and the number of genotypes in the sample. In particular, existing methods are vastly inadequate for handling datasets of the size envisioned to be produced by next generation of genome-wide association studies. These studies are expected to result in thousands of individual genotypes with 500,000 or more SNPs [5] by leveraging recent advances in genotyping technologies such as the Affymetrix Mapping 500K Array Set [6].

In this paper we propose a highly scalable algorithm based on the entropy minimization principle that has previously been proposed in the context of genotype phasing and haplotype missing data recovery by Halperin and Karp [7]. Unlike the simple greedy algorithm employed in [7], we use a local optimization algorithm, which in practice results in genotype phasings with lower entropy. After formalizing the problem in Section II, in Section III we describe a simple yet very efficient implementation of this algorithm, and a novel *overlapping window* approach for handling genotypes with large numbers of SNPs. We also describe the extension of our algorithm to the case when the input genotypes come from a mixture of unrelated individuals and families consisting of a child and one or both parents. Phasing related genotypes is likely to gain in importance in future genotyping studies since relationships between genotypes can be exploited to reliably infer haplotype phase for a substantial fraction of the SNPs based on the no-recombination assumption [5]. Finally, in Section IV we present experimental results on large real datasets extracted from the HapMap repository [8] showing that our algorithm achieves a phasing accuracy close to that of best existing methods while being several orders of magnitude faster.

This work was supported in part by NSF CAREER award IIS-0546457 and NSF award DBI-0543365.

Authors address: University of Connecticut, Computer Science & Engineering Department, 371 Fairfield Rd., Storrs, CT 06269-2155, {bogdan,ion}@engr.uconn.edu

II. Problem Formulation

Following the standard practice, in this paper we restrict our attention to bi-allelic SNPs, which form the vast majority of known SNPs. In this case a haplotype can be represented as a 0/1 vector – typically by representing the most frequent SNP allele as a 0 and the alternate allele as a 1. A genotype can be viewed as a 0/1/2 vector, where 0 (1) means that both chromosomes contain the 0 (1) allele while 2 means that the two chromosomes contain different alleles.

We say that haplotype h is *compatible* with genotype g if $g(i) = h(i)$ whenever $g(i) \in \{0, 1\}$. A pair of haplotypes (h_1, h_2) *explains* genotype g if $h_1(i) = h_2(i) = g(i)$ whenever $g(i) \in \{0, 1\}$, and $h_1(i) \neq h_2(i)$ whenever $g(i) = 2$. For a given pair (h_1, h_2) that explains g we say that h_1 and h_2 are complements with respect to g.

A *phasing* of a set of genotypes G, each of length k, is a function $\phi : G \rightarrow \{0, 1\}^k \times \{0, 1\}^k$, such that, for every $g \in G$, $\phi(g)$ is a pair of haplotypes that explain g. For a haplotype h and a phasing ϕ, the *coverage of h under ϕ*, denoted by $cov(h, \phi)$, is the number of genotypes $g \in G$ such that $\phi(g) = (h, h')$ or $\phi(g) = (h', h)$ plus twice the number of of genotypes $g \in G$ such that $\phi(g) = (h, h)$. As in [7], we define the *entropy* of a phasing ϕ as

$$\mathcal{H}(\phi) = \sum_{h:cov(h,\phi)\neq 0} -\frac{cov(h, \phi)}{2|G|} \log \frac{cov(h, \phi)}{2|G|} \qquad (1)$$

The **Minimum Entropy Genotype Phasing Problem** can then be defined as follows: Given a set of genotypes, find a phasing with minimum entropy.

III. Algorithm

Halperin and Karp [7] proposed a greedy algorithm for the related minimum-entropy set cover problem, and showed that a variant of this algorithm can be applied to genotype phasing. However, this algorithm cannot be applied directly to phasing long genotypes, i.e., genotypes with large numbers of SNPs. Indeed, in this case each haplotype is likely to be compatible with a single genotype, and thus all phasings are likely to have the same entropy of $-\log \frac{1}{2|G|}$. Furthermore, even for short genotypes, the greedy algorithm in [7] is producing phasings whose entropy can be further decreased. In this paper we use the entropy minimization objective of [7] within a local improvement framework. In Section III-A we describe the local improvement algorithm for phasing short genotypes of unrelated individuals. Then, in Sections III-B and III-D we describe extensions of the local improvement algorithm to the problem of phasing long genotypes of unrelated, respectively related individuals.

A. Short genotype phasing

We have implemented a simple local improvement algorithm for entropy minimization. Our algorithm which we refer to as ENT, starts from a random phasing, then, at each step, finds the genotype whose re-explanation yields the largest decrease in phasing entropy (see Figure 1). The use of random initial phasings is justified by observing

Input: Set G of genotypes
Output: Phasing ϕ of the genotypes in G

1. Generate a random phasing ϕ for genotypes in G
2. **Repeat forever**
   - 2.1 Find the pair $(g, (h'_1, h'_2))$ such that $\mathcal{H}(\phi')$ is minimized, where ϕ' is obtained from ϕ by re-explaining g with (h'_1, h'_2)
   - 2.2 **If** $\mathcal{H}(\phi') < \mathcal{H}(\phi)$, **then** $\phi \leftarrow \phi'$ **Else** exit the **repeat** loop
3. Output ϕ

Fig. 1. ENT phasing of short genotypes.

that a random phasing of a genotype with i heterozygous positions matches the real phasing with probability 2^{-i}. E.g., for the Daly children dataset (see Section IV), random phasing results in an average of 46% correct haplotypes over windows of 5 consecutive SNPs. We have also experimented with a version of the algorithm in which the initial phasing is obtained by running the greedy algorithm of [7]. However, the use of random initial phasings was found to yield convergence to final phasings with lower entropy.

If there exists more than one pair $(g, (h_1, h_2))$ with minimum $\mathcal{H}(\phi')$ in step 2.1 of the algorithm, then we pick the pair $(g, (h_1, h_2))$ maximizing $Prob(h_1) \times Prob(h_2)$, where $Prob(h)$ is defined as $\Pi_1^n p(h[i])$, and $p(h[i])$ is the probability of seeing allele $h[i]$ at position i.

B. Long genotype phasing

A common approach to phasing long genotypes is to phase small *non-overlapping windows* of the input genotypes and then stitch together the resulting haplotypes using various statistical approaches. Recently, Eskin, Sharan, and Halperin [9] proposed a dynamic programming algorithm for selecting a set of tiling windows maximizing a natural maximum likelihood function. Our algorithm also uses a window-based approach to phasing long genotypes, however, unlike previous approaches, it employs a set of *overlapping windows*. Each window consists of a set of l "locked" SNPs, which have been previously phased, and a set of f "free" SNPs, which are currently being phased. For each window, the phasing algorithm proceeds as described in the previous section, except that only re-explanations consistent with the already determined haplotypes of the locked SNPs are considered in the local improvement step (see Figure 2).

The basic implementation of the ENT algorithm takes l and f as input parameters. We have also implemented variants of the algorithm that dynamically compute the number of locked, respectively free SNPs based on the input data. These variants pick l and f as large as possible subject to the constraint that the numbers of ambiguous (heterozygous or missing) SNP genotypes in the locked, respectively free region of the current window do not exceed twice the number of genotypes. The number of free SNPs f is further constrained to disallow having more than 7 ambiguous SNPs in the free region of any genotype.

# Free SNPs	# Locked SNPs									
	1	2	3	4	5	6	7	8	9	Variable
1	19.5	37.9	38.6	38.6	39.2	42.1	44.0	43.6	42.0	39.4
2	15.1	10.3	13.4	25.6	24.7	30.2	26.9	30.3	28.2	18.2
3	12.0	8.0	6.0	10.7	17.2	23.0	19.2	22.5	24.3	11.4
4	11.0	8.0	5.1	5.0	6.7	8.2	13.3	19.8	15.6	6.9
5	8.5	6.5	5.6	4.7	4.4	5.1	7.0	7.2	7.9	5.2
6	7.6	6.4	4.4	5.2	4.9	5.4	5.5	5.5	5.5	5.0
7	8.2	5.6	6.2	5.3	5.0	5.0	5.4	5.4	5.4	5.1
8	6.8	6.1	5.5	6.0	5.1	5.2	5.4	4.8	5.9	5.4
9	6.6	5.6	4.9	4.7	4.5	5.2	5.5	5.3	5.2	4.8
Variable	7.0	5.6	4.5	4.9	4.7	4.7	6.5	6.5	7.3	4.2

TABLE I

ENT SWITCHING ERROR RATES (%) FOR VARIOUS WINDOW SETTINGS ON THE DALY DATASET.

Input: Set G of genotypes
Output: Phasing ϕ of the genotypes in G

1. Divide the genotypes in groups of f consecutive SNPs from left to right
2. For each group, add the preceding l SNPs to create a window of size $l + f$ SNPs (leftmost window has no locked SNPs and is of size f)
3. Run the phasing algorithm in Figure 1 for each window, in left to right order, where the haplotypes over the locked l SNPs are not allowed to change
4. Output the resulting phasing ϕ

Fig. 2. ENT phasing of long genotypes.

C. Time complexity

When phasing n unrelated genotypes over k SNPs, the algorithm in Figure 1 is run on $\lceil k/f \rceil$ windows. For each window, the algorithm evaluates at most $n \times 2^f$ candidate pairs of haplotypes for finding the best pair in Step 2.1. Computing the entropy gain for each candidate pair takes constant time. Indeed, $\mathcal{H}(\phi')$ differs from $\mathcal{H}(\phi)$ in at most four terms corresponding to the haplotypes that can change their coverages, namely the haplotypes explaining g in ϕ and ϕ'. Empirically, the number of iterations required in Step 2 of the algorithm in Figure 1 is linear in the number n of genotypes, resulting in an overall runtime of $O(n^2 2^f k/f)$. The number of iterations can be reduced to nearly constant by re-expaining multiple genotypes per iteration. This speed-up technique – which results in a runtime that depends nearly linearly on the number of genotypes – will be included in the next implementation of our algorithm, but was not used for obtaining the experimental results in Section IV.

D. Phasing related genotypes

A trio is a nuclear family composed of the two parents plus a child. In the no-recombination assumption each parent passes one of its chromosomes to the child. That is, the child shares one haplotype with the mother and the other one with the father. The no-recombination assumption provides very useful information about phasing all members of a trio. The only situation when there is phasing ambiguity for a given SNP is when all three genotypes are heterozygous at that SNP. For example, in the CEU and YRI trio populations of HapMap [8], the phase of only around 15% of the SNPs is ambiguous, while the phase of the remaining 85% of the SNPs can be inferred based on the no-recombination assumption.

The ENT algorithm described above can be easily adapted to phase families of related genotypes under the no-recombination assumption. In order to enforce the no-recombination assumption, at each local improvement step, we re-explain a whole family, rather than an individual genotype. The entropy can still be recomputed in constant time after each update by a straightforward extension of the method described in Section III-A. Our current implementation handles trio genotype data as well as mixtures of independent genotypes, full trios, and partial trios consisting of one parent and one child.

IV. EXPERIMENTS

In a first set of experiments we assessed the effect of the windowing strategy (number of free and locked SNPs) on phasing accuracy of the ENT algorithm. We conducted these experiments on a well-known dataset from Daly et al. [10]. This dataset contains 129 trios from a European population. Each individual was genotyped at 103 SNP positions in the 5q31 region of chromosome 5. The trio genotypes were used to infer as much as possible out of the "true" haplotypes of the children under the no-recombination assumption. We use the following three measures [5] to assess phasing accuracy on the unrelated genotypes of the children in the Daly dataset:

Switching error. Given inferred haplotypes (h, h') of a genotype g with true haplotypes (t, t'), the number of switches is defined as the number of times one has to switch between h and h' to obtain t. The number of ambiguous SNPs in a genotype g is the number of 2's (heterozygous positions) plus the number of missing SNP genotypes. The switching error rate (given in percents) for a set G of n genotypes is defined as the ratio between the total number of switches and the total number of ambiguous SNPs minus n, since the maximum number of switches in a genotype is one less than the number of ambiguous SNPs.

Haplotype accuracy. The percentage of haplotypes correctly recovered.

SNP accuracy. The number of correctly phased SNPs as percentage of the total number of SNPs.

Table I reports the switching error obtained by our ENT algorithm with various settings for the number of free and locked SNPs on the Daly dataset. We varied the number of locked and free SNPs from 1 to 9, and also included in comparison the ENT variants which dynamically choose either one or both l and f as described in Section III-B. The version that chooses both l and f dynamically yields the smallest switching error, with next best results being obtained by using fixed window sizes with 5 locked and 5 free SNPs.

Table II gives all three accuracy measures for the two best performing variants of ENT, the widely used PHASE [11], and the more recent GERBIL [12] and 2SNP [13] phasing algorithms. The two ENT variants have slightly worse, yet very close accuracy compared to the other methods.

Unfortunately, the methods in [11], [12], [13] do not directly handle trio data. An extension of PHASE to trio data has been described in [5], however, its runtime does not scale well to very large trio datasets such as those generated in the HapMap project [8]. To test the scalability of the ENT algorithm, in a second set of experiments we used two datasets from HapMap Phase I release 16a, each one consisting of 30 trios. The first dataset was collected from a population of Utah residents with ancestry from northern and western Europe (CEU), and the other one from a population of Yoruba people of Ibadan, Nigeria (YRI).

As reported on the HapMap website, phasing these datasets using the trio version of the PHASE algorithm [5] requires extensive computational resources (months of CPU time on two clusters with a combined total of 238 nodes) and for this reason the haplotypes can be recomputed only for major releases of the datasets. In contrast, only a few hours on a 2.8GHz Pentium Xeon computer were required by the ENT variant which dynamically picks the number of locked and free SNPs. In Table III we report the accuracy of ENT phasing with respect to the results obtained by PHASE. The accuracy is computed as the number of *trio ambiguous SNPs* – i.e., positions in which all three members of the trio have ambiguous SNPs – that are differently phased by ENT and PHASE, as percentage of the total number of trio ambiguous SNPs. Of the approximately 15% of the SNPs that are trio ambiguous, only 7-12% are inferred differently by ENT and PHASE, depending on the population and the chromosome. Thus, our method results in a SNP genotyping difference of 1-2% with respect to PHASE, while being many orders of magnitude faster.

In the above experiments, the ENT algorithm was run on the genotypes inferred from the PHASE haplotypes since the corresponding genotypes (which most likely have missing data) are not available at [8]. In order to test the capacity of our method to recover missing alleles we randomly deleted 1%, 2%, 5%, and 10% of the genotype SNPs and used the genotypes with missing data as input to ENT. In Table IV we report the number of SNPs where our algorithm recovers correctly the missing alleles as percentage of the total number of deleted SNPs. The recovery accuracy varies with the percentage of deleted data, but is on the average over 97.5% for the CEU population and over 95.8% percent for the YRI population.

	Switching error (%)	Haplotype accuracy (%)	SNP accuracy (%)
PHASE 2.1	3.09	55.81	98.29
GERBIL	3.18	44.96	97.89
2SNP	3.18	51.16	98.58
ENT Var/Var	4.21	42.64	97.96
ENT 5/5	4.36	45.74	97.85

TABLE II

PHASING ACCURACY OF DIFFERENT METHODS ON THE DALY DATASET.

V. CONCLUSIONS

In this paper we have presented a highly scalable algorithm for genotype phasing based on entropy minimization. Experimental results on large datasets extracted from the HapMap repository show that our algorithm is several orders of magnitude faster than existing phasing methods while achieving a phasing accuracy close to that of best existing methods. The source code of our implementation is available from the authors upon request.

REFERENCES

[1] D. Gusfield, "An overview of combinatorial methods for haplotype inference," in *Proc. DIMACS/RECOMB Satellite Workshop on Computational Methods for SNPs and Haplotype Inference*, 2004, pp. 9–25.

[2] B. Halldorsson, V. Bafna, N. Edwards, R. Lippert, S. Yooseph, and S. Istrail, "A survey of computational methods for determining haplotypes," in *Proc. DIMACS/RECOMB Satellite Workshop on Computational Methods for SNPs and Haplotype Inference*, 2004, pp. 26–47.

[3] T. Niu, "Algorithms for inferring haplotypes," *Genet. Epid.*, vol. 27, pp. 334–347, 2004.

[4] R. Salem, J. Wessel, and N. Schork, "A comprehensive literature review of haplotyping software and methods for use with unrelated individuals," *Human Genomics*, vol. 2, pp. 39–66, 2005.

[5] J. Marchini, D. Cutler, N. Patterson, M. Stephens, E. Eskin, E. Halperin, S. Lin, Z. Qin, H. Munro, G. Abecasis, P. Donnelly, and International HapMap Consortium, "A comparison of phasing algorithms for trios and unrelated individuals," *American Journal of Human Genetics*, vol. 78, pp. 437–450, 2006.

[6] `http://www.affymetrix.com/products/arrays/specific/500k.affx`.

[7] E. Halperin and R. Karp, "The minimum-entropy set cover problem," in *Proc. Annual International Colloquium on Automata, Languages and Programming (ICALP)*, 2004.

[8] `http://www.hapmap.org/`.

[9] E. Eskin, E. Halperin, and R. Sharan, "Optimally phasing long genomic regions using local haplotype predictions," (to appear).

[10] M. Daly, J. Rioux, S. Schaffner, T. Hudson, and E. Lander, "High-resolution haplotype structure in the human genome." *Nature Genetics*, vol. 29, no. 2, pp. 229–232, 2001.

[11] M. Stephens and N. J. Smith and Peter Donnelly, "A new statistical method for haplotype reconstruction from population data," *American Journal of Human Genetics*, vol. 68, pp. 978–989, 2001.

[12] G. Kimmel and R. Shamir, "Gerbil: Genotype resolution and block identification using likelihood," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, pp. 158–162, 2005.

[13] D. Branza and A. Zelikovsky, "2snp: scalable phasing based on 2-snp haplotypes," *Bioinformatics*, vol. 22, no. 3, pp. 371–373, 2006.

CEU population (30 trios)				
Chr #	# SNPs	Trio Ambig. SNPs (%)	Diff. from PHASE (%)	CPU sec.
1	61814	16.4	10.3	401
2	69753	16.3	8.5	939
3	56737	15.8	8.9	682
4	48952	16.0	9.0	450
5	48831	15.5	8.8	445
6	53458	16.0	8.7	452
7	41046	15.4	10.2	292
8	60234	16.0	8.5	855
9	47682	15.2	8.5	628
10	38940	16.1	9.1	353
11	36287	15.5	9.7	391
12	39189	15.9	9.9	340
13	28816	15.1	10.6	157
14	24128	16.2	9.2	215
15	21138	16.6	9.5	147
16	19922	16.2	11.2	126
17	19767	15.5	10.1	215
18	32177	16.0	9.3	465
19	14175	16.8	12.0	86
20	17096	16.6	10.5	144
21	16199	15.2	11.5	187
22	15548	15.2	9.9	176
Total	**811889**	**-**	**-**	**8155**
YRI population (30 trios)				
Chr #	# SNPs	Trio Ambig. SNPs (%)	Diff. from PHASE (%)	CPU sec.
1	68579	16.6	8.8	273
2	74275	16.1	6.4	399
3	56617	16.1	6.7	254
4	49807	16.0	7.4	211
5	47400	16.2	7.0	209
6	55376	16.8	6.9	188
7	39126	16.3	7.7	142
8	64461	16.0	6.5	410
9	50258	15.5	6.3	414
10	42002	16.1	7.6	155
11	36268	16.4	7.1	171
12	40666	16.1	8.3	149
13	31627	16.3	7.2	118
14	23968	17.2	7.1	95
15	21504	16.6	7.6	63
16	20237	16.4	9.3	69
17	19744	16.5	8.6	77
18	35094	15.8	7.1	196
19	14007	16.8	9.5	52
20	16580	16.4	12.2	56
21	17897	16.0	7.2	82
22	16386	15.9	7.2	72
Total	**841879**	**-**	**-**	**3866**

TABLE III

PERCENTAGE OF TRIO AMBIGUOUS SNPS WITH DIFFERENT ENT AND PHASE PHASINGS.

CEU population (30 trios)					
Chr#	#SNPs	Deleted SNPs			
		1%	2%	5%	10%
1	61814	97.79	97.74	97.64	97.40
2	69753	98.30	98.23	98.10	97.83
3	56737	98.23	98.13	98.00	97.74
4	48952	98.15	98.07	97.93	97.72
5	48831	98.18	98.08	97.96	97.72
6	53458	98.24	98.19	98.00	97.79
7	41046	97.90	97.89	97.72	97.46
8	60234	98.54	98.48	98.39	98.16
9	47682	98.39	98.25	98.12	97.84
10	38940	97.82	97.82	97.66	97.42
11	36287	98.09	98.07	97.94	97.65
12	39189	97.85	97.84	97.74	97.49
13	28816	98.09	97.98	97.82	97.63
14	24128	98.01	97.87	97.86	97.58
15	21138	97.66	97.70	97.59	97.32
16	19922	97.34	97.24	97.10	96.84
17	19767	97.41	97.36	97.23	96.92
18	32177	98.41	98.31	98.16	97.96
19	14175	96.95	97.06	96.84	96.54
20	17096	97.29	97.35	97.31	96.93
21	16199	98.11	98.12	98.17	97.84
22	15548	97.97	97.95	97.71	97.46
Averages	**-**	**97.94**	**97.90**	**97.77**	**97.51**
YRI population (30 trios)					
Chr#	#SNPs	Deleted SNPs			
		1%	2%	5%	10%
1	68579	96.17	96.03	95.85	95.54
2	74275	97.15	97.03	96.79	96.32
3	56617	96.92	96.93	96.64	96.25
4	49807	96.82	96.77	96.46	96.09
5	47400	96.80	96.71	96.51	96.02
6	55376	96.93	96.81	96.60	96.23
7	39126	96.39	96.25	96.04	95.64
8	64461	97.66	97.62	97.31	96.99
9	50258	97.34	97.13	96.93	96.53
10	42002	96.47	96.41	96.18	95.79
11	36268	96.97	96.70	96.51	96.06
12	40666	96.36	96.22	96.04	95.70
13	31627	96.69	96.58	96.40	96.04
14	23968	96.62	96.56	96.31	95.82
15	21504	95.98	96.07	95.70	95.22
16	20237	95.90	95.76	95.36	94.84
17	19744	96.00	95.88	95.52	95.11
18	35094	97.14	97.02	96.89	96.54
19	14007	95.22	95.27	94.94	94.51
20	16580	94.66	94.67	94.32	93.86
21	17897	97.31	97.15	96.81	96.43
22	16386	96.70	96.67	96.33	95.98
Averages	**-**	**96.55**	**96.47**	**96.20**	**95.80**

TABLE IV

PERCENTAGE OF DELETED SNPS CORRECTLY RECOVERED ON HAPMAP CEU AND YRI TRIO POPULATIONS [8].