

Linearly Independent Tagging of Genotypes

Jingwu He^{1,2}, Kelly Westbrooks¹ and Alexander Zelikovsky^{1,3}

Keywords: Single nucleotide polymorphism, tag SNP, linear independence

1 Introduction

Constructing a complete human haplotype map is helpful when associating complex diseases with their related SNPs. Unfortunately, the number of SNPs is very large and it is costly to sequence many individuals. A common approach is to identify a small number of informative SNPs called *tag SNPs* in a population of known genotypes (or haplotypes), then to type the tag SNPs and reconstruct a genotype (or a pair of haplotypes) for each unknown individual. We can consider the following:

Predictive Tag SNP Selection Problem. Given a sample S of a population P of genotypes on m SNPs, find positions of k ($k < m$) tag SNPs such that one can reconstruct an entire genotype (or haplotype) $g \in P$ from its restriction g' on k tag SNPs.

We assume that genotypes in the sample S are phased and do not contain any missing data (otherwise, one can phase S using, e.g., PHASE [5]). On the other hand, some tag SNPs in the restricted genotype (or haplotype) g' may contain missing data.

We propose a new linear algebra-based method for selecting and using tag SNPs. Our method is purely combinatorial and can be combined with linkage disequilibrium (LD) and block-based methods. We compare with the existing prediction methods of [2] and [6] using a leave-one-out framework. We also explore the dependency of the accuracy of genotype reconstruction on the number of tag SNPs and sample size.

2 Linearly Independent Tagging

One cannot straightforwardly apply linear dependency to haplotypes since *complimentary* columns are linearly independent. To overcome this obstacle, we replace 0's with -1 's as in [3]. The proposed method removes SNP sites which are linearly dependent on other SNP sites:

- Apply $O(s^2m)$ Gauss-Jordan elimination on the $s \times m$ matrix S to obtain the reconstruction matrix $F = rref(S)$. Extract $r = rank(S)$ linearly independent columns $\{t_1, \dots, t_r\}$ as tag SNPs of S .
- If $k = r$, for each genotype/haplotype g , one can reconstruct g from its restriction g' via matrix multiplication, i.e. $g = g' \times F$.
- If $k < r$, we greedily choose k tag SNP columns $\{t_1, \dots, t_k\}$ from S such that the other columns can be represented as closely as possible as a linear combination of tag SNP sites $\{t_1, \dots, t_k\}$.
- If $k > r$, for each non-tag site, we select the majority values from each reconstruction of all r independent sites.

¹Department of Computer Science, Georgia State University, Atlanta, GA 30303. E-mail: {jingwu, kelly, alexz}@cs.gsu.edu

²Supported by GSU Molecular Basis of Disease Fellowship.

³Partially supported by NIH Award 1 P20 GM065762-01A1.

3 Experimental Results.

Our experiments in [4] show that for sufficiently long haplotypes (> 25000 SNPs), knowing only tag SNPs constituting 0.4% of all SNPs the proposed linear reduction method reconstructs an unknown haplotype with the error rate below 2% while the tag are selected based on 10% of the population.

We apply leave-one-out cross-validation to evaluate the quality of the solution given by the tagging SNP selection algorithm. The haplotype left out is reconstructed based on the tag SNPs and the reconstruction matrix. The average number of errors in reconstruction over all haplotypes is used as a measure of the overall accuracy of the tagging method. The comparison results with [2] and [6] are shown as follows:

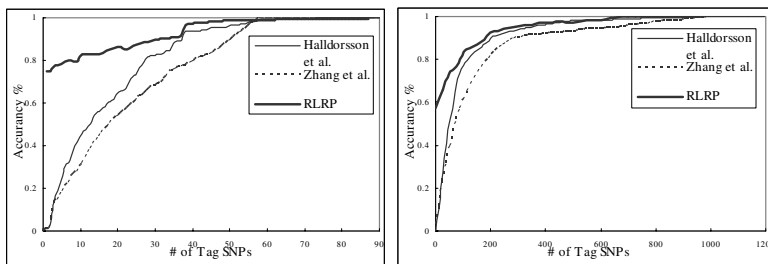


Figure 1: Leave-one-out tests on real data. (Left) Results from the LPL data set. (Right) Results from the first 1000 SNPs of Chromosome 21 data set.

We can apply our method to genotype tagging and reconstruction. The average number of errors in reconstruction over genotypes is used as a measure of the overall accuracy of the genotype tagging method. The accuracy of genotype reconstruction depends on the number of selected tag SNPs and sample percentage in the population (see Figure 2).

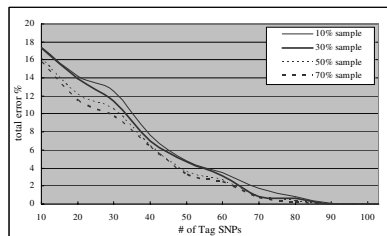


Figure 2: The accuracy of genotype reconstruction depending on the number of selected tag SNPs and sample sizes equal to 10%, 30%, 50% and 70% of population of [1].

References

- [1] Daly, M., Rioux, J., Schaffner, S., Hudson, T. and Lander, E. 2001. *Nature Genetics*, 29:229–232.
- [2] Halldorsson, B.V, et al. 2004. *Genome Research*, 14:1633–1640.
- [3] He, J. and Zelikovsky, A. 2004. *Proceedings WABI'04*, 3240:242–253.
- [4] He, J. and Zelikovsky, A. 2004. *Proceedings of EMBC'04*, 2840–2843.
- [5] Stephens, M., et al. 2001. *Am. J. Human Genetics*, 68:978–989.
- [6] Zhang, K., et al. 2004. *Genome Research*, 14:908–916.