Genotype Error Detection and Imputation using Hidden Markov Models of Haplotype Diversity*

Justin Kennedy, Ion Măndoiu, and Bogdan Paşaniuc

Department of Computer Science & Engineering, University of Connecticut, 371 Fairfield Rd., Unit 2155, Storrs, CT 06269-2155, USA
E-mail: {jlk02019,ion,bogdan}@engr.uconn.edu

1 Introduction

Recent advances in Single Nucleotide Polymorphism (SNP) genotyping technologies have made possible large scale genome-wide association studies that promise to uncover the genetic basis of complex human diseases. The validity of associations uncovered by these studies critically depends on the accuracy of the genotype data. Despite recent progress in genotype calling algorithms, SNP genotyping errors remain present at levels that can invalidate statistical test for disease association, particularly for methods based on haplotype analysis. Furthermore, since causal SNPs are unlikely to be typed directly due to the limited coverage of current genotyping platforms, imputation of genotypes at untyped SNP loci has recently emerged as a powerful technique for increasing the power of association studies [8, 10, 12, 7].

In this poster we introduce GEDI, a software package for *Genotype Error Detection and Imputation* of genotypes at untyped SNP loci based on reference haplotypes such as those available in HapMap. Detection of genotyping errors and imputation of missing genotypes is based on multi-locus genotype likelihoods efficiently computed using a Hidden Markov Model (HMM) that captures the Linkage Disequilibrium (LD) observed in the population under study. With a runtime that scales linearly both in the number of markers and the number of typed individuals, GEDI is able to handle very large datasets while achieving high accuracy rates for both error detection and imputation.

2 Methods

At the core of GEDI is a left-to-right HMM used to represent haplotype frequencies in the underlying population [5]. Our HMM has a structure similar to that of models recently used for other haplotype analysis problems including genotype phasing, testing for disease association, and imputation [6, 8–11]. The HMM has K states for every SNP locus, where K is a user-specified parameter (typically a small constant, we used $K = 7$ in our experiments). Although each state can emit each allele with non-zero probability, during model training emission probabilities of most states are strongly biased towards one allele or another. Unlike

* Work supported in part by NSF awards IIS-0546457 and DBI-0543365.

the models in [8, 10], which estimate a single recombination rate for every pair of consecutive SNP loci, our model has independent transition probabilities for all pairs of states corresponding to consecutive SNP loci. All model parameters (emission and transition probabilities) are estimated from haplotype data using the classical Baum-Welch algorithm [1]. Intuitively, the HMM represents a number of K founder haplotypes along high-probability "horizontal" paths of states, while capturing observed recombinations between pairs of founder haplotypes via probabilities of "non-horizontal" transitions.

For imputation of genotypes at untyped SNP loci, HMM training must be done using haplotypes from a reference panel (such as HapMap) that includes both the typed SNPs and the SNPs to be imputed. For genotype error detection and imputation of missing genotypes at typed SNP loci the training can be done based either on reference haplotypes or on haplotypes inferred from the available genotype data for the population under study. The latter option is to be preferred particularly when genotype data is available for families of related individuals, which enables high accuracy haplotype inference. To facilitate HMM training when only genotype data is available, GEDI automatically performs haplotype inference using the highly scalable ENT phasing algorithm of [4].

We denote by 0 and 1 the major and minor alleles at every SNP locus, by 0, 1, and 2 the three possible SNP genotypes (homozygous major/minor, respectively heterozygous), and by '?' a missing SNP genotype. The probability of a multilocus haplotype $H \in \{0,1\}^n$ under a trained HMM model M, denoted $P(H|M)$, is the sum over all possible HMM paths π of length n of the joint probability that M follows path π and emits H. $P(H|M)$ can be computed in $O(nK^2)$ time using the standard forward algorithm. Similarly, the probability under M of a multilocus genotype $G \in \{0,1,2,?\}^n$ is given by $P(G|M) = \sum P(H|M)P(H'|M)$, where the sum is over all pairs of haplotypes (H, H') compatible with G, and can be computed efficiently in $O(nK^3)$ time using an optimized two-path version of the forward algorithm [9].

Let $G_{g_i \leftarrow x}$ denote the multilocus genotype obtained from G by replacing the i-th SNP genotype with x, where $x \in \{0,1,2\}$. After HMM training, GEDI computes *all* probabilities $P(G_{g_i \leftarrow x}|M)$ for $i = 1, \ldots, n$ and $x \in \{0,1,2\}$. This is done in $O(nK^3)$ time per multilocus genotype using the speed-up idea of [9] in combination with a forward-backward algorithm. For a missing genotype g_i imputation is done by replacing g_i with $x^* = \text{argmax}_{x \in \{0,1,2\}} P(G_{g_i \leftarrow x}|M)$. For each non-missing genotype g_i, GEDI computes the *log-likelihood ratio*

$$\log\left(\frac{\max_{x \in \{0,1,2\}} P(G_{g_i \leftarrow x}|M)}{P(G|M)}\right)$$

Genotype g_i is replaced by $x^* = \text{argmax}_{x \in \{0,1,2\}} P(G_{g_i \leftarrow x}|M)$ and flagged as a potential error whenever the log-likelihood ratio exceeds either a global or locus-specific detection threshold specified by the user.

GEDI also implements extensions of the above error detection and imputation methods to the case when the input is genotype data of related individuals. In this case testing/imputation is still done one SNP genotype at a time, but

imputation probabilities and log-likelihood ratios are computed over genotype data of entire mother-father-child trios (see [5] for more details). When input genotype data includes family trios, GEDI has the additional option of automatically estimating locus-specific detection thresholds based on allele frequencies and observed Mendelian inconsistency rates using the method of [3].

3 Experimental Results

In this section we provide preliminary results on GEDI's imputation accuracy; error detection accuracy with an earlier version of GEDI can be found in [5]. We started our experiment from the genotype data of the 1958 birth cohort of the WTCCC study [2]. 1,444 individuals from this cohort were typed using both the Affymetrix 500k platform and a custom Illumina chip containing 15k SNPs. In our experiment we used the Affymetrix data in conjunction with the CEU HapMap haplotypes to impute genotypes at the SNP loci present of the Illumina chip and not on the Affymetrix chip. The actual Illumina genotypes were then used to estimate imputation accuracy.

In Figure 1 we compare different estimates of the frequency of 0 alleles for Illumina SNPs on chromosome 1. As shown in Figure 1(a), allele frequencies estimated from the HapMap CEU haplotypes are well corelated with allele frequencies derived from the Illumina genotypes, suggesting that the HapMap CEU haplotypes are a good reference panel for the 1958 birth cohort. Figure 1(b) shows that allele frequencies learned by the HMM used in GEDI match very well allele frequencies in the HapMap haplotypes that are used to train it, suggesting that the Baum-Welch training method is appropriate. Finally, Figure 1(c) shows a high corelation between allele frequencies derived from genotypes imputed by GEDI and those derived from Illumina genotypes. Indeed, over all chromosomes GEDI achieves a 1.5% discordance between Illumina genotype calls and genotypes imputed with a confidence of 95% confidence or higher. A comprehensive evaluation of GEDI's imputation accuracy and comparison with existing imputation methods including [8, 10, 12, 7] is ongoing.

References

1. L.E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist.*, 41:164–171, 1970.
2. The Wellcome Trust Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447:661–678, 2007.
3. D. Gordon, S.C. Heath, and J. Ott. True pedigree errors more frequent than apparent errors for single nucleotide poloymorphisms. *Hum. Hered.*, 49:65–70, 1999.
4. A. Gusev, B. Paşaniuc, and I.I. Măndoiu. Highly scalable genotype phasing by entropy minimization. to appear.

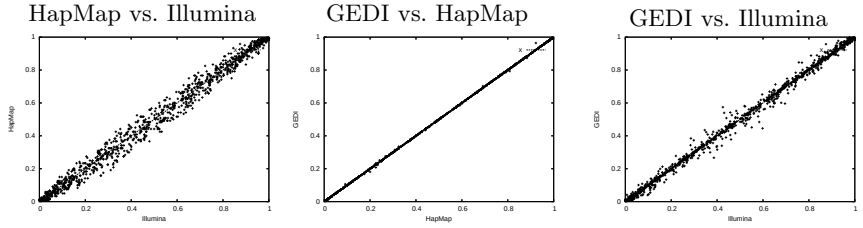

Fig. 1. Estimates of the frequency of 0 alleles for Illumina SNPs on chromosome 1. (a) HapMap based estimates vs. estimates based on Illumina genotypes; (a) GEDI HMM estimates vs. HapMap based estimates; (c) Imputation based estimates vs. estimates based on Illumina genotypes.

5. J. Kennedy, I.I. Mandoiu, and B. Pasaniuc. Genotype error detection using hidden Markov models of haplotype diversity. In *Proc. 7th Workshop on Algorithms in Bioinformatics*, Lecture Notes in Computer Science, pages 73–84, 2007. KMP07.ppt.
6. G. Kimmel and R. Shamir. A block-free hidden Markov model for genotypes and its application to disease association. *Journal of Computational Biology*, 12:1243–1260, 2005.
7. Y. Li and G. R. Abecasis. Mach 1.0: Rapid haplotype reconstruction and missing genotype inference. *American Journal of Human Genetics*, page 2290, 2006.
8. J. Marchini, B. Howie, S. Myers, G. McVean, and P. Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.*, 39:906–913, 2007.
9. P. Rastas, M. Koivisto, H. Mannila, and E. Ukkonen. Phasing genotypes using a hidden Markov model. In I.I. Măndoiu and A. Zelikovsky, editors, *Bioinformatics Algorithms: Techniques and Applications*, pages 355–372. Wiley, 2008.
10. P. Scheet and M. Stephens. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *American Journal of Human Genetics*, 78:629–644, 2006.
11. R. Schwartz. Algorithms for association study design using a generalized model of haplotype conservation. In *Proc. CSB*, pages 90–97, 2004.
12. X. Wen and D. L. Nicolae. Association studies for untyped markers with tuna. *Bioinformatics*, 24:435–437, 2008.