

Linkage Disequilibrium Based Single Individual Genotyping from Low-Coverage Short Sequencing Reads

Sanjiv Dinakar¹, Jorge Duitama¹, Yözen Hernández², Justin Kennedy¹, Ion I. Măndoiu¹, Yufeng Wu¹

1 Introduction

Recent advances in *high-throughput sequencing* (HTS) technologies such as the Roche 454 FLX, Illumina Genome Analyzer, ABI SOLiD, and Helicos HeliScope have enabled cost-effective shotgun sequencing of individual genomes. Indeed, several individual genomes have already been published [2, 5, 8, 10], and there are ongoing efforts (e.g., [1]) to sequence thousands more. Sequencing can be used to discover new SNPs and other forms of sequence variation such as small insertions and deletions, copy number variants, genome rearrangements, etc., thus providing a complete picture of individual genome variation. The ideal outcome of such a sequencing project is the diplotype genome of the individual – i.e., full haplotype sequences for the individual’s maternal and paternal chromosomes – since haplotype sequences provide the detailed context required for accurate *functional characterization* of genomic variants [6] and studying genome evolution. This motivates:

Diplotype Genome Reconstruction Problem: *Reconstruct the two haplotype sequences from HTS shotgun sequencing reads.*

The sensitivity of detecting heterozygous loci from shotgun sequencing reads is severely limited by coverage depth. Wendl and Wilson [9] estimate that a coverage as high as $21\times$ may be required to achieve a genotype calling accuracy comparable to the “finished” sequence standards established for the human genome project. This estimate is based on simple genotype calling methods that treat different loci as unlinked and postulate that the number of reads covering each allele at a heterozygous locus follow a binomial distribution. Unfortunately, at coverages such as those used in [5, 10] ($7.5\times$), the binomial tests were found to detect only 75% of the heterozygous SNPs, and sensitivity drops rapidly at even lower coverage depths.

Genotype calling methods such as those in [5] disregard *linkage disequilibrium* (LD), i.e. correlations between variable genome loci located in close proximity of each other. In the following, we describe a linkage-disequilibrium based method for genotype and haplotype reconstruction with common SNPs. This method allows cost-effective reconstruction of accurate multilocus genotypes and haplotypes over an increasingly large set of variable loci.

2 Methodologies

Our analysis pipeline (Figure 1) combines low coverage shotgun sequencing data with LD information extracted from a *reference* population panel. At present, such information is available for 11 human populations for which sample individuals have been typed at up to 4 million common SNPs as part of the international Hapmap project [7]. An expected outcome of the 1000 genomes project [1] will be the construction of LD maps covering even more variable loci (SNPs and small indels with a minor allele frequency of at least 1%). Thus, our method will allow cost-effective reconstruction of accurate multilocus genotypes and haplotypes over an increasingly large set of variable loci.

Our multilocus statistical model (graphical model in Figure 1), can be thought of as a *hierarchical factorial HMM* (HF-HMM). Indeed, we use a distributed state characteristic of factorial HMMs [4] to exploit the independence between maternal and paternal chromosomes implied by the assumption of random mating, while also employing a multilevel state representation as in hierarchical HMMs [3] to capture the structured nature of the data. The hierarchical factorial structure of the model leads to a reduced number of parameters and modular estimation procedures, and enables highly scalable inference algorithms, with runtime scaling linearly with both the number of shotgun reads and that of SNP loci. Genotype calling is performed using posterior probabilities efficiently computed based on the HF-HMM using a forward-backward algorithm.

¹Computer Science and Engineering Department, University of Connecticut, 371 Fairfield Way, Unit 2155, Storrs, CT 06269-2155. E-mail: {sdinakar, jduitama, j1k02019, ion, ywu}@engr.uconn.edu.

²Department of Computer Science, Hunter College, 695 Park Avenue, New York, NY 10021. E-mail: yzhernand@gmail.com

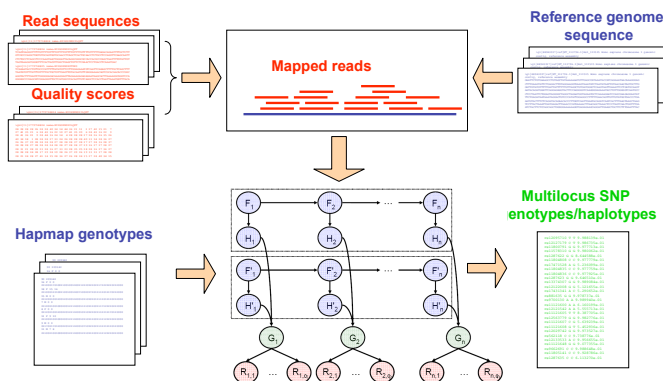


Figure 1: LD-based genotype and haplotype reconstruction pipeline for common SNPs.

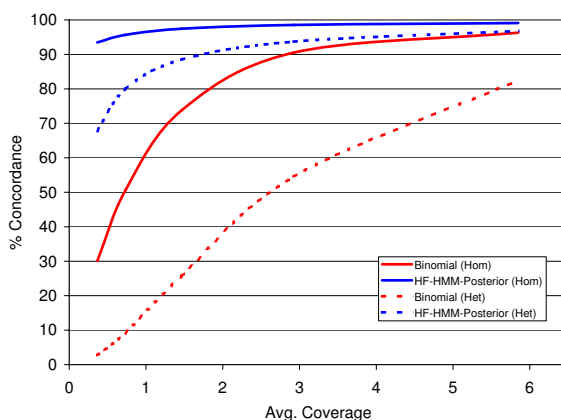


Figure 2: Concordance between SNP genotypes inferred using the HF-HMM from sequencing data with varying average coverage and genotypes determined using replicated Affymetrix microarray experiments.

3 Results

Experiments on publicly available whole-genome datasets generated using 454 [10], Illumina [2], and SOLiD technologies show that the HF-HMM based posterior decoding algorithm significantly outperforms binomial test methods, especially for heterozygotes. For example, experiments on the Watson 454 sequencing data (Figure 2) show that the posterior algorithm matches the accuracy achieved at $5.85\times$ average coverage by the binomial test of [10] using only one *quarter* of the reads, even when the latter is relaxed to allow a minimum coverage of 1 for each allele.

References

- [1] 1000 Genomes Project Consortium. <http://www.1000genomes.org/>.
- [2] D.R. Bentley et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59, 2008.
- [3] S. Fine, Y. Singer, and N. Tishby. The hierarchical hidden Markov model: Analysis and applications. *Mach. Learn.*, 32(1):41–62, 1998.
- [4] Z. Ghahramani and M.I. Jordan. Factorial hidden Markov models. *Mach. Learn.*, 29(2-3):245–273, 1997.
- [5] S. Levy et al. The diploid genome sequence of an individual human. *PLoS Biology*, 5(10):e254+, 2007.
- [6] P.C. Ng et al. Genetic variation in an individual human exome. *PLoS Genet*, 4(8):e1000160+, 2008.
- [7] The International HapMap Consortium. A second generation human haplotype map of over 3.1 million snps. *Nature*, 449:851–861, 2007.
- [8] J. Wang et al. The diploid genome sequence of an asian individual. *Nature*, 456(7218):60–65, November 2008.
- [9] M.C. Wendl and R.K. Wilson. Aspects of coverage in medical DNA sequencing. *BMC Bioinformatics*, 9:239+, 2008.
- [10] D.A. Wheeler et al. The complete genome of an individual by massively parallel DNA sequencing. *Nature*, 452:872–876, 2008.