Linkage Disequilibrium Based Genotype Calling from Low-Coverage Shotgun Sequencing Reads

Jorge Duitama[1], Justin Kennedy[1], Sanjiv Dinakar[2], Yözen Hernández[3]and Yufeng Wu[1], Ion I. Măndoiu*[1]

[1]Department of Computer Science & Engineering, University of Connecticut,371 Fairfield Rd., Unit 2155, Storrs, CT 06269-2155, USA
[2]Department of Computer Science, University of Maryland,College Park, Maryland 20742, USA.
[3]Department of Computer Science, Hunter College,695 Park Avenue, New York, NY 10021, USA.

Email: Jorge Duitama - jduitama@engr.uconn.edu; Justin Kennedy - jlk02019@engr.uconn.edu; Sanjiv Dinakar - sdinakar@cs.umd.edu; Yözen Hernández - yzhernand@gmail.com; Yufeng Wu - ywu@engr.uconn.edu; Ion I. Măndoiu*- ion@engr.uconn.edu;

*Corresponding author

Abstract

Background: Recent technology advances have enabled sequencing of individual genomes, promising to revolutionize biomedical research. However, deep sequencing remains more expensive than microarrays for performing whole-genome SNP genotyping.

Results: In this paper we introduce a new multi-locus statistical model and computationally efficient genotype calling algorithms that integrate shotgun sequencing data with linkage disequilibrium (LD) information extracted from reference population panels such as Hapmap or the 1000 genomes project. Experiments on publicly available 454, Illumina, and ABI SOLiD sequencing datasets suggest that integration of LD information results in genotype calling accuracy comparable to that of microarray platforms from sequencing data of low-coverage. A software package implementing our algorithm, released under the GNU General Public License, is available at http://dna.engr.uconn.edu/software/GeneSeq/.

Conclusions: Integration of LD information leads to significant improvements in genotype calling accuracy compared to prior LD-oblivious methods, rendering low-coverage sequencing as a viable alternative to microarrays for conducting large-scale genome-wide association studies.

Background

Recent advances in massively parallel sequencing have dramatically increased throughput compared to the classic Sanger technology, with several commercially available platforms including 454, Illumina, ABI SOLiD, and Helicos delivering billions of bases per day. This has enabled sequencing of several individual genomes [1–8], ushering the era of personal genomics. Thousands of other individual genomes are currently being sequenced as part of large scale projects such as the international 1000 genomes project [9], and whole genome sequencing is likely to become routine as sequencing costs continue to decrease. However, analysis of whole genome sequencing data remains challenging [10] and experimental design optimization has only recently started to receive attention [11].

In this paper we focus on one of the most fundamental genomic analyses, namely determining the genotypes at known loci of genome variation such as single nucleotide polymorphisms (SNPs). Diploid organisms including humans inherit two (possibly identical) variants or *alleles* at autosomal loci, and most medical applications of personal genomics require accurate identification of both variants, the combination of which is referred to as *genotype*. Of particular interest are loci that are heterozygous, i.e., loci for which the two chromosomes carry different alleles. However, identifying heterozygous loci from low-coverage whole-genome sequencing data poses a significant challenge. Sequencing data is obtained using the so called "shotgun" approach, whereby millions of short DNA fragments called reads are generated from randomly selected locations on the two chromosomes. If, for example, there are only two reads generated from a heterozygous locus, there is a 50% chance that one allele would be missed. To compensate for sequencing errors, existing methods for detecting heterozygous loci have even higher minimum allele coverage requirements, e.g., in [3, 8], calling an allele requires the presence of at least two reads supporting it. Consequently, due to the relatively low sequencing depth used in these two studies (about 7.5×), the reported sensitivity of detecting heterozygous SNPs was of only 75%.

A simple way to improve genotype calling accuracy is to increase sequencing depth, as the probability of "missing" an allele decreases with the number of reads. After taking into account the effect of sequencing errors it has been estimated that, in the absence of additional information, achieving 99% sensitivity at detecting heterozygous SNPs would require an average sequencing depth of over 21× [12]. Our main contribution is to demonstrate that high accuracy SNP genotypes can be inferred from shotgun sequencing data of much lower depth by exploiting the correlation between alleles at nearby SNP sites, commonly referred to as *linkage disequilibrium* (LD).

LD patterns over millions of common SNPs have been mapped for several populations as part of the

Hapmap project [13]. The strong LD observed in human populations has already been exploited by methods for imputation of genotypes at untyped SNP loci based on nearby SNP genotypes [14–19], see [20] for a recent review, and more recently, for improving genotype calling accuracy from microarray hybridization signals [21]. Another striking demonstration of the power of LD has been the inference of Watson's APOE status [22] despite the removal of sequencing reads covering this region from the published dataset [8]. In this work we introduce a novel hierarchical factorial Hidden Markov Model (HMM) that allows integrated analysis of LD information extracted from reference population panels such as Hapmap and short-read sequencing data generated by current technologies. Although the ensuing multilocus genotype inference is computationally hard, we develop a scalable heuristic similar to the posterior decoding algorithm for HMMs. A software package implementing this algorithm has been released under the GNU General Public License and is available at http://dna.engr.uconn.edu/software/GeneSeq/.
We also present experimental results on publicly available 454, Illumina, and ABI SOLiD whole-genome sequencing datasets showing that integration of LD information leads to significant improvements in genotype calling accuracy compared to prior LD-oblivious methods. For example, at 6× average mapped read coverage, our algorithm calls heterozygous SNP genotypes with about 96% accuracy, and accuracy can be further increased to 98-99% by leaving uncalled a small percentage of SNP genotypes with low posterior probabilities. This accuracy is comparable to that achieved by microarray-based genotyping platforms. Coupled with continued decreases in sequencing costs, the reduced sequencing depth required when using LD information renders low-coverage sequencing as a potentially more cost-effective alternative to microarrays for the next generation of genome wide association studies (GWAS). For example, the ABI SOLiD 4hq is expected to deliver 300Gb of sequencing data per run, or the equivalent of 16 individual genomes at 6× coverage, with a cost of only $600 per genome [23]. Undoubtedly, cost will be an important factor in future GWAS studies, which are expected to use much higher sample sizes compared to past studies in order to enable the study of gene-gene and gene-environment interactions [24].

Methods

In this section we begin by describing a simplified statistical model that assumes independence between loci, then extend it to include dependences between alleles at different SNPs due to LD. We next formalize the multilocus genotype calling problem in the context of the extended model and show that computing the most likely multilocus genotype is computationally hard. Finally, we present a posterior decoding heuristic which independently selects the most likely genotype at each locus conditional on the entire set of reads.

Notations

We use uppercase italic letters (e.g., X) to denote random variables and lowercase italic letters (e.g., x) to denote generic values taken by them. Vectors of random variables and generic variables are denoted by boldface uppercase (e.g., $\mathbf{X}$), respectively boldface lowercase letters (e.g., $\mathbf{x}$). When there is no ambiguity on the underlying probabilistic event we use $P(x)$ to denote $P(X = x)$, with similar shorthands used for joint and conditional probabilities of multiple events. For simplicity we consider only bialelic SNPs on autosomes. For every SNP locus, we denote the two possible alleles by 0 and 1, and the three genotypes by 0, 1, and 2, with 0 and 2 denoting the homozygous 0 and homozygous 1 genotypes, and 1 denoting the heterozygous genotype.

Single SNP Genotype Calling

In this section we describe a genotype inference model that assumes the SNPs to be unlinked as in [8], but further incorporates allele uncertainty quantified by sequencing quality scores, read mapping uncertainty, and population genotype frequencies estimated from a reference panel.

Let r be a read mapped onto the genome. If r covers SNP locus i, we denote by $r(i)$ the allele observed in the read at this locus. Since our focus is on genotyping SNPs represented in a reference panel, we further assume that panel SNPs at which the individual under study has novel allele variants (observed in [8] at only 0.02% of the markers) have been identified in a preliminary analysis, e.g., by using binomial probability test of [8]. Based on this assumption, all reads with alleles not represented in the panel population are ignored, and for remaining reads r we have that $r(i) \in \{0, 1\}$. The probability that allele $r(i)$ is affected by a sequencing error is denoted by $\varepsilon_{r(i)}$. In our experiments we set $\varepsilon_{r(i)} = 10^{-q_{r(i)}/10}$, where where $q_{r(i)}$ denotes the Phred quality score of $r(i)$ [25].

Let G_i be a random variable denoting the unknown SNP genotype at locus i, and let $\mathbf{r}_i = \{r_{i,1}, \ldots, r_{i,c_i}\}$ be the arbitrarily ordered set of shotgun reads covering locus i, where c_i is the coverage at this locus. Since for a homozygous genotype the allele of origin for a read is the same regardless of which chromosome is sampled, we get:

$$P(\mathbf{r}_i|G_i = 0) = \prod_{\substack{r\in\mathbf{r}_i \\ r(i)=0}} (1 - \varepsilon_{r(i)}) \prod_{\substack{r\in\mathbf{r}_i \\ r(i)=1}} (\varepsilon_{r(i)}) \tag{1}$$

and

$$P(\mathbf{r}_i|G_i = 2) = \prod_{\substack{r\in\mathbf{r}_i \\ r(i)=1}} (1 - \varepsilon_{r(i)}) \prod_{\substack{r\in\mathbf{r}_i \\ r(i)=0}} (\varepsilon_{r(i)}) \tag{2}$$

For a read r covering a heterozygous SNP locus i allele $r(i)$ can be observed either as the result of sampling r from the chromosome bearing allele $r(i)$ and correctly sequencing it, or as the result of sampling the other chromosome followed by a sequencing error. Hence:

$$P(\mathbf{r}_i|G_i = 1) = \prod_{r \in \mathbf{R}_i} \left(\frac{1}{2}(1 - \varepsilon_{r(i)}) + \frac{1}{2}\varepsilon_{r(i)} \right) = \left(\frac{1}{2} \right)^{c_i} \tag{3}$$

A natural approach to single-locus SNP genotyping is to call a genotype of $\widehat{g}_i = \text{argmax}_{g_i \in \{0,1,2\}} P(g_i|\mathbf{r}_i)$ for every SNP locus i, where the posterior probabilities $P(g_i|\mathbf{r}_i)$ are obtained from (1)-(3) by applying Bayes' formula:

$$P(G_i = g_i|\mathbf{r}_i) = \frac{P(g_i)P(\mathbf{r}_i|G_i = g_i)}{\sum_g P(G_i = g)P(\mathbf{r}_i|G_i = g)} \tag{4}$$

and $P(G_i = g)$ denotes the population frequency of genotype g, estimated from the reference panel.

If read mapping uncertainty is available in the form of probabilities $m(r)$ that read r is mapped at the correct position, such information can be accounted for in genotype calling by replacing the above conditional probabilities with genotype weights obtained from (1)–(3) by rising the terms corresponding to read r to power $m(r)$. Although the resulting weights can no longer be interpreted as conditional genotype probabilities, they naturally allow interpreting the presence of a read r with mapping confidence $m(r) < 1$ as the equivalent of observing an $m(r)$ fraction of an identical read mapped with confidence 1.

A Statistical Model for Multilocus Genotype Inference

In this section we introduce a statistical model that allows us to integrate shotgun sequencing data and LD information in the inference of SNP genotypes. Our model, represented graphically in Fig. 1, can be thought of as a *hierarchical factorial HMM* (HF-HMM). Indeed, we use a distributed state (characteristic of factorial HMMs [26]) 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 [27] to capture the structured nature of the data. This structure leads to a reduced number of model parameters and enables highly scalable inference algorithms.

At the core of the model are two left-to-right HMMs M and M' (dotted boxes in Fig. 1), each emitting haplotypes with frequencies corresponding to those in the populations of origin for the sequenced individual's parents. Under M and M', each haplotype is viewed as a mosaic formed as a result of historical recombination among a set of K founder haplotypes, where K is a population specific model parameter. Formally, for every SNP locus $i \in \{1, \ldots, n\}$, we let H_i (H'_i) be a random variable representing the allele observed at this locus on the maternal (paternal) chromosome of the individual under study, and

F_i (F'_i) be a random variable denoting the founder haplotype from which H_i (respectively H'_i) originates. As in previous works [15, 17, 28–30], we assume that F_i form the states of a first order HMM with emissions H_i, and estimate probabilities $P(f_1)$, $P(f_{i+1}|f_i)$, and $P(h_i|f_i)$) using the classical Baum-Welch algorithm [31] based on haplotypes inferred from a panel representing the population of origin of the individual's mother. Probabilities $P(f'_1)$, $P(f'_{i+1}|f'_i)$, and $P(h'_i|f'_i)$ are estimated in the same way based on haplotypes inferred from a panel representing the population of origin of the individual's father.
We define $P(g_i|h_i, h'_i)$ to be 1 if $g_i = h_i + h'_i$ and 0 otherwise. Finally, assuming that each read covers no more than a SNP locus, we set

$$P(r_{i,j}|G_i = g_i) \quad = \quad \frac{g_i}{2}(\varepsilon_{r(i)})^{1-r(i)}(1-\varepsilon_{r(i)})^{r(i)} + \frac{2-g_i}{2}(\varepsilon_{r(i)})^{r(i)}(1-\varepsilon_{r(i)})^{1-r(i)} \tag{5}$$

This implies that $P(\mathbf{r}_i|g_i)$ are given by Equations (1)-(3), and in the following we will assume that probabilities $P(\mathbf{r}_i|g_i)$ are precomputed in $O(m)$ time, where $m = \sum_{i=1}^{n} c_i$ is bounded above by the total number of reads. We can now formulate the following:

Multilocus Genotyping Problem (MGP)

Given: *Trained HMM models M, M' and set of shotgun reads $\mathbf{r} = (\mathbf{r}_1, \ldots, \mathbf{r}_n)$*

Find: *Multilocus genotype $\mathbf{g}^* \in \{0, 1, 2\}^n$ with maximum posterior probability, i.e.,*

$$\mathbf{g}^* = \text{argmax}_{\mathbf{g}} P(\mathbf{g}|\mathbf{r}, M, M') \tag{6}$$

Computational complexity

In this section we show that MGP is NP-hard. Let *Maximum Multilocus Genotype Probability Problem (MMGPP)* denote the optimization version of MGP that requires finding $\max_{\mathbf{g}} P(\mathbf{g}|\mathbf{r}, M, M')$.

Theorem 1. *For any $\epsilon > 0$, MMGPP cannot be approximated within $O(n^{\frac{1}{2}-\epsilon})$ unless P=NP, and it cannot be approximated within $O(n^{1-\epsilon})$ unless ZPP=NP. Furthermore, this holds even if $M' = M$.*

Proof. Lyngsø et al. [32] give an approximation preserving reduction from the clique problem to the problem of computing the maximum probability of a string emitted by an HMM. It is not difficult to modify their construction to show that this reduction holds for left-to-right HMMs that emit 0/1 strings of fixed length. Next, we show that computing the maximum probability of a string emitted by such an HMM M_0 can be reduced in approximation preserving manner to MMGPP with $M' = M$. The haplotype models M and M' are obtained from M_0 as follows (see the schematic state diagram in Fig. 2):

- The number of SNPs n is set to one plus the length of the strings emitted by M_0.
- At the first SNP, for two founder states f_1^1 and f_1^2 we have $P(f_1^i) = 1/2$; all other founder states have zero initial probability.
- For every SNP locus $i > 1$ we add a new founder f_1^i as well as a set of founders corresponding to the states at "column" $i-1$ of M_0.
- All founder f_i^1, $i = 1, \ldots, n$, emit 0 with probability 1. Furthermore, $P(f_i^1 | f_{i-1}^1) = 1$ for every $i = 2, \ldots, n$.
- Founder f_1^2 emits 1 with probability 1, and has transitions to founders f_2^j, $j > 1$, according to the initial probabilities of M_0.
- All other emission and transition probabilities are identical to those for the corresponding states of M_0.

Finally, we set $\mathbf{r} = \{r_0, r_1\}$ where r_0 is a read that supports allele 0 at first SNP and r_1 is a read that supports the allele 1 at first SNP. Error probabilities for both alleles are set to zero.
Note that $P(\mathbf{g}|\mathbf{r}, M, M') \neq 0$ only for multilocus genotypes with $g_1 = 1$ and $g_i \in \{0, 1\}$ for $i = 2, \ldots, n$. Furthermore, for such a genotype $\mathbf{g}$,

$$\begin{aligned} P(\mathbf{g}|\mathbf{r}, M, M') &= \frac{P(\mathbf{r}|\mathbf{g})P(\mathbf{g}|M, M')}{P(\mathbf{r})} \\ &= \frac{1}{4P(\mathbf{r})} P(\mathbf{g}|M, M') \\ &= \frac{1}{4P(\mathbf{r})} \frac{P(g_2, \ldots, g_n|M_0)}{2} \end{aligned} \tag{7}$$

The last equality comes from the fact that $\mathbf{g}$ can only be observed when the maternal haplotype is 0^n and the paternal haplotype is $\mathbf{g}$ or vice-versa, and each of these configurations have a probability of $P(g_2, \ldots, g_n|M_0)/4$. The innaproximability result follows from [32] since, by (7), $P(\mathbf{g}|\mathbf{r}, M, M')$ is constant fraction of $P(g_2, \ldots, g_n|M_0)$. □

Since an algorithm similar to the forward algorithm for HMMs can be used to compute in polynomial time the marginal probability of a given genotype, Theorem 1 implies the following:

Corollary 2. *MGP is NP-Hard.*

Posterior decoding algorithm

We next present an MGP heuristic similar to the posterior decoding algorithm for HMMs. Specifically, the algorithm selects for each SNP locus i the genotype $\widehat{g}_i$ with maximum posterior probability given the read data $\mathbf{r}$. Note that, unlike the single SNP genotype calling method, where we condition only on the set $\mathbf{r}_i$ of reads overlapping locus i, in the posterior decoding algorithm we take into account the *entire* set of reads:

Posterior Decoding Algorithm

Step 1. For each $i = 1, \ldots, n$, $\widehat{g}_i \leftarrow \text{argmax}_{g_i} P(g_i|\mathbf{r})$

Step 2. Return $\widehat{\mathbf{g}} = (\widehat{g}_1, \ldots, \widehat{g}_n)$

Below we detail an $O(m + nK^3)$ implementation of the posterior decoding algorithm. Since $P(g_i|\mathbf{r}) \propto P(g_i, \mathbf{r})$, for implementing the maximization in Step 1 it suffices to compute marginal probabilities $P(g_i, \mathbf{r})$ for every $i = 1, \ldots, n$ and $g_i \in \{0, 1, 2\}$. For each SNP locus i and each pair of founders (f_i, f_i') we let the *forward probability* be $\mathcal{F}^i_{f_i,f_i'} = P(\mathbf{r}_1, \ldots, \mathbf{r}_{i-1}, f_i, f_i')$, and the *backward probability* be $\mathcal{B}^i_{f_i,f_i'} = P(\mathbf{r}_{i+1}, \ldots, \mathbf{r}_n|f_i, f_i')$, respectively. Using these forward and the backward probabilities, the marginal probability $P(g_i, \mathbf{r})$ can be written as

$$P(g_i, \mathbf{r}) = P(\mathbf{r}_i|g_i) \sum_{f_i=1}^{K} \sum_{f_i'=1}^{K} \mathcal{F}^i_{f_i,f_i'} \mathcal{B}^i_{f_i,f_i'} P(g_i|f_i, f_i')$$

where $P(g_i|f_i, f_i')$ is given by:

$$P(g_i|f_i, f_i') = \sum_{\substack{h_i,h_i' \in \{0,1\} \\ h_i+h_i'=g_i}} P(h_i|f_i)P(h_i'|f_i')$$

Thus all probabilities $P(g_i, \mathbf{r})$ can be computed in $O(nK^2)$ once the forward and backward probabilities $\mathcal{F}^i_{f_i,f_i'}$ and $\mathcal{B}^i_{f_i,f_i'}$ are available.

The forward probabilities can be computed using the recurrence:

$$\mathcal{F}^1_{f_1,f_1'} = P(f_1)P(f_1') \tag{8}$$

$$\begin{aligned} \mathcal{F}^i_{f_i,f_i'} &= \sum_{f_{i-1}=1}^{K} \sum_{f_{i-1}'=1}^{K} \left(\mathcal{F}^{i-1}_{f_{i-1},f_{i-1}'} \mathcal{E}^{i-1}_{f_{i-1},f_{i-1}'} P(f_i|f_{i-1}) P(f_i'|f_{i-1}') \right) \\ &= \sum_{f_{i-1}=1}^{K} P(f_i|f_{i-1}) \sum_{f_{i-1}'=1}^{K} \mathcal{F}^{i-1}_{f_{i-1},f_{i-1}'} \mathcal{E}^{i-1}_{f_{i-1},f_{i-1}'} P(f_i'|f_{i-1}') \end{aligned} \tag{9}$$

for every $f_i, f'_i \in \{1, \ldots, K\}$ and $i = 2, \ldots, n$, where

$$\mathcal{E}^i_{f_i, f'_i} = \sum_{h_i, h'_i \in \{0,1\}} P(h_i|f_i) P(h'_i|f'_i) P(\mathbf{r}_i|G_i = h_i + h'_i) \tag{10}$$

The inner sum in equation (9) is independent of f_i, and so its repeated computation can be avoided by replacing (9) with:

$$\mathcal{C}^i_{f_{i-1}, f'_i} = \sum_{f'_{i-1}=1}^{K} \mathcal{F}^{i-1}_{f_{i-1}, f'_{i-1}} \mathcal{E}^{i-1}_{f_{i-1}, f'_{i-1}} P(f'_i|f'_{i-1}) \tag{11}$$

$$\mathcal{F}^i_{f_i, f'_i} = \sum_{f_{i-1}=1}^{K} P(f_i|f_{i-1}) \mathcal{C}^i_{f_{i-1}, f'_i} \tag{12}$$

A similar optimization can be applied when computing the backward probabilities, resulting in the following recurrence:

$$\mathcal{B}^n_{f_n, f'_n} = 1 \tag{13}$$

$$\mathcal{D}^i_{f_{i+1}, f'_i} = \sum_{f'_{i+1}=1}^{K} \mathcal{B}^{i+1}_{f_{i+1}, f'_{i+1}} \mathcal{E}^{i+1}_{f_{i+1}, f'_{i+1}} P(f'_{i+1}|f'_i) \tag{14}$$

$$\mathcal{B}^i_{f_i, f'_i} = \sum_{f_{i+1}=1}^{K} P(f_{i+1}|f_i) \mathcal{D}^i_{f_{i+1}, f'_k} \tag{15}$$

Forward and backward probabilities can thus be computed in $O(nK^3)$ by using recurrences (8), (11), and (12), respectively (13), (14), and (15), resulting in an overall runtime of $O(m + nK^3)$, where m is the number of reads, n is the number of SNPs, and K is a user selected parameter denoting the number of founders in the HMM models of haplotype diversity in the parental populations (we used $K = 7$ in our experiments).

Results and Discussion

Datasets

We evaluated the HMM-based posterior decoding algorithm on shotgun sequencing datasets generated using three different sequencing technologies, as follows:

1. Watson 454: A set of 74.4 million reads downloaded from the NCBI SRA database (submission number: SRA000065). The reads, with an average length of ∼265 bp, were generated using the

Roche 454 FLX platform as part of James Watson's personal genome project. This is a subset of the 106.5 million 454 reads analyzed in [8]. Unless noted otherwise, the haplotype panel used to train identical HMM models for the maternal and paternal populations was obtained by phasing CEU trio genotypes from Hapmap r23a [13] using the ENT algorithm of [33] and retaining parent haplotypes from each trio. As in [8], genotype calling accuracy was assessed using the SNP genotypes determined using duplicate hybridization experiments with Affymetrix 500k microarrays (only concordant genotypes were retained in the test set).

2. NA18507 Illumina: A set of 525 million paired-end reads downloaded from the NCBI SRA database (submission number: SRA000271). These 36bp reads, which were generated using the Illumina Genome Analyzer from a Hapmap Yoruban individual identified as NA18507, are a subset of the dataset analyzed in [1]. For the analysis of this dataset the HMM models for maternal and paternal populations were trained using YRI haplotypes from Hapmap r22, excluding the haplotypes of the YRI trio that contains NA18507. As gold standard we used the genotypes published as part of Hapmap r22 for individual NA18507.

3. NA18507 SOLiD: A set of 900 million single ABI SOLiD reads generated from Hapmap individual NA18507 was kindly provided by the authors of [4]. Reads varied in length between 20 and 44 bp, and were already mapped to the reference genome. Corresponding raw reads are available for download from the NCBI SRA database (submission number: SRA000272). HMM models and gold standard genotypes were determined in the same way as for the NA18507 Illumina dataset.

Read Mapping

We mapped 454 reads on build 36.3 of the reference human genome using the NUCMER tool of the MUMmer package [34] with default parameters. We discarded alignments matching less than 90% of the reference or with 10 or more errors (mismatches or indels). We then discarded surviving reads with multiple matching positions. We mapped the Illumina reads using MAQ version 0.68 [35] with default parameters. We discarded alignments with mapping probability less than 0.9 or with sum of quality scores of mismatching bases higher than 60 (filtering was performed using the "submap" command of MAQ). SOLiD reads were mapped using the SOLiD System Analysis Pipeline Tool (Corona Lite) as described in [4]. Table 1 shows for each dataset the numbers of test SNPs, initial and mapped reads, and the average coverage per SNP after mapping.

Genotyping Accuracy

To evaluate the effects of read coverage on genotype calling for each dataset of m mapped reads we created four subsets of sizes $m/16$, $m/8$, $m/4$ and $m/2$ by picking reads at random. For each subset we called genotypes using the HMM-based posterior decoding algorithm, the binomial test of [8] (with a threshold of 0.01), and the single SNP posterior probability described under Methods. We also included in the comparison genotype calls obtained by SOAPsnp [36] and MAQ [35], two widely used LD-oblivious Bayesian methods implemented in the SAMtools package [37].[1] We measured the accuracy of each genotype calling method by computing the percentage of SNP genotype calls that match the gold standard available for each dataset. As in previous papers [1, 3, 4, 8], we separately report accuracy for homozygous and heterozygous SNPs.

Fig. 3 shows genotype calling accuracy of the compared methods for varying average mapped read coverage on the NA18507 Illumina dataset; similar results were obtained on the other two datasets. For both homozygous and heterozygous SNPs, the posterior decoding algorithm has the highest accuracy of the compared methods at every considered coverage. The improvement in accuracy is most pronounced for heterozygous SNPs and at low average coverage. This is not surprising since, as previously noted in [3, 4, 8], at low average coverage there is an increasingly high probability of leaving uncovered at least one of the alleles of a heterozygous SNP, and a minimum coverage of each called allele is required by the binomial test, SOAPsnp, and MAQ. For example, the binomial test used in [3, 8] requires that each allele be covered at least twice; in all our results we used the more relaxed requirement of covering each allele at least once. In contrast, the single-SNP posterior and the HMM-based posterior decoding algorithm do not have a minimum coverage requirement. By leveraging population allele frequencies estimated from the reference panel, the single-SNP posterior method already outperforms the binomial test, SOAPsnp, and MAQ at low average coverage. The HMM posterior decoding algorithm further improves accuracy by capturing LD information between neighboring SNPs.

Fig. 4(a) shows the accuracy achieved by the HMM posterior decoding algorithm when varying the average mapped read coverage for all three datasets. Genotyping accuracy achieved on the NA18507 Illumina reads matches that observed on Watson 454 reads for homozygous SNPs, and is only slightly lower for heterozygous SNPs. The accuracy achieved on the NA18507 SOLiD reads is consistently lower than that achieved for the other two datasets over the tested range of average coverages. We found that this difference

[1] Unfortunately we could not compare our method with similar tools developed by the members of the 1000 genomes project. A beta version of Thunder [38] has only recently been made publicly available, while Qcall [39] has not yet been released.

is due to a bias towards the reference allele during color-to-base translation for reads mapped with Corona Lite. This bias is likely to induce incorrect heterozygous calls for some homozygous non-reference SNPs and homozygous reference calls for some heterozygous SNPs. The presence of this bias can be observed in Fig. 4(b), which shows the distribution of reference allele coverage ratios (i.e., ratios between the number of reference allele calls and the total number of mapped reads covering a locus) for heterozygous SNPs in the Watson 454, NA18507 Illumina, and NA18507 SOLiD datasets. In the absence of allele call biases, the average of reference allele coverage ratios over heterozygous SNPs should be close to 50%. We found that this was indeed the case for both the Watson 454 and NA18507 Illumina datasets (with averages of 51.39% and 51.02%, respectively) but not for the NA18507 SOLiD dataset (for which the average ratio is 63.02%). Fig. 5 shows the concordance of genotypes called by HMM posterior decoding on the NA18507 Illumina dataset for groups of SNPs with varying rates of local recombination, respectively minor allele frequency, both estimated from the YRI panel of Hapmap. The percentage of SNPs in each group is also plotted using dashed lines. For both homozygous and heterozygous SNPs concordance is relatively stable over the entire range of local recombination rates (see Fig. 5(a)), dropping below 96% only for heterozygous SNPs in regions with local recombination rate of over 10 cM/Mb. The effect of minor allele frequency is more pronounced (see Fig. 5(b)), with heterozygous SNPs concordance dropping to 83% for SNPs with minor allele frequency below 0.05. However, the overall accuracy is not affected too much since only 2% of heterozygous SNPs of NA18507 have an estimated allele frequency in this range.

To assess the effect of the size of the reference panel on genotyping accuracy, we conducted additional experiments on the Watson 454 reads using $N = 242$ CEU haplotypes available in Hapmap3. Similar to experiments with varying read coverage, we generated subsets of approximately $N/16$, $N/8$, $N/4$, and $N/2$ randomly selected reference haplotypes, and compared the accuracy achieved by running the HMM posterior algorithm using these subsets to that obtained using all N reference haplotypes. Fig. 6(a) gives the genotype call concordance obtained for different panel sizes. The results suggest that no significant improvement is achieved by increasing the reference panel size beyond 60-90. Thus – in contrast to methods for imputing untyped SNPs, which continue to benefit from increasing the panel size to several hundreds of haplotypes [40] – highly accurate genotype calling from sequencing data is possible with relatively small reference panels.

Since our algorithm computes a posterior probability for each SNP genotype, further increases in calling accuracy can be obtained at the expense of leaving uncalled a small percentage of SNP genotypes with low posterior probability. Such "no-calls" are commonly used in microarray-based genotyping for SNPs for

which hybridization signals are ambiguous. Fig. 6(b) shows the tradeoffs achievable between the concordance and call rate when running the HMM posterior decoding algorithm on the full set of Watson 454 reads. Over all SNPs, concordance with the duplicate Affymetrix genotypes reaches 99.4% at a no-call rate of only 6%.

Conclusions

In this paper we introduced a statistical model for multi-locus genotyping that integrates shotgun sequencing data with LD information extracted from a reference panel. Although finding the multi-locus genotype with maximum posterior probability under the integrated model is NP-Hard, experimental results suggest that a simple posterior decoding algorithm produces highly accurate genotype calls even from low-coverage sequencing data. Compared to current LD-oblivious genotype calling methods, our method allows researchers to achieve a desired accuracy target with reduced sequencing costs. For example, genotype calling accuracy achieved at 5-6× average coverage by a previously proposed binomial test is matched by the HMM-based posterior decoding algorithm using less than 1/4 of the reads.

While a full comparison of sequencing and microarray based genotyping in the context of GWAS is beyond the scope of this paper, experimental results on three publicly available datasets generated using the 454, Illumina, and ABI SOLiD sequencing platforms suggest that at a mapped coverage depth of 5-6× our algorithm achieves an accuracy that is comparable to that of microarray platforms. Concordance rates reported for microarrays often exceed 99.9% (see, e.g., [41]), and are even higher for methods that integrate hybridization signals with LD information [21]. However, due to cost constraints, microarrays typically assay only a fraction of the SNPs represented in reference panels. For example, the next generation of Illumina microarrays is expected to assay only 5 million of the estimated 35 million SNPs generated by the 1000 genomes project [42]. Genotypes for the untyped SNPs would have to be inferred based solely on LD information, and even the best imputation methods have error rates of 5-6% [20], or 2-3% when leaving 10% of SNPs uncalled. Since the majority of SNPs must be imputed, this results in an overall accuracy below that achieved by the HMM posterior algorithm on the Watson 454 dataset.

In ongoing work we are exploring efficient algorithms for LD-based haplotype reconstruction from paired shotgun sequencing reads. We also plan to empirically compare our method with [38] and [39].

Authors contributions

IIM and YW conceived the study. JD, JK, SD, and YH implemented the methods and conducted the experiments. JD and IIM drafted the manuscript. All authors participated in the development of the methods, data analysis, and manuscript revision. All authors have read and approved the final manuscript.

Acknowledgments

JD, JK, and IIM were supported in part by NSF awards IIS-0546457, IIS-0916948, and DBI-0543365. YW was supported in part by NSF award IIS-0953563. SD and YH were supported in part by NSF award CCF-0755373.

References

1. Bentley *et al* D: **Accurate Whole Human Genome Sequencing using Reversible Terminator Chemistry**. *Nature* 2008, **456**:53–59.
2. Drmanac *et al* R: **Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays**. *Science* 2009, **327**(78):78–81.
3. Levy *et al* S: **The Diploid Genome Sequence of an Individual Human**. *PLoS Biology* 2007, **5**(10):e254+.
4. McKernan *et al* K: **Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding**. *Genome Research* 2009, **19**:1527–1541.
5. Pushkarev D, Neff N, Quake S: **Single-molecule sequencing of an individual human genome**. *Nature Biotechnology* 2009, **27**(9):847–850.
6. Schuster *et al* S: **Complete Khoisan and Bantu genomes from southern Africa**. *Nature* 2010, **463**(18):943–947.
7. Wang *et al* J: **The diploid genome sequence of an Asian individual**. *Nature* 2008, **456**:60–65.
8. Wheeler *et al* D: **The complete genome of an individual by massively parallel DNA sequencing**. *Nature* 2008, **452**:872–876.
9. The 1000 Genomes Project Consortium: **The 1000 Genomes Project Consortium** [http://www.1000genomes.org/].
10. Snyder M, Du J, Gerstein M: **Personal genome sequencing: current approaches and challenges**. *Genes & Development* 2010, **24**:423–431.
11. Bashir A, Bansal V, Bafna V: **Designing deep sequencing experiments: detecting structural variation and estimating transcript abundance**. *BMC Genomics* 2010, **11**:385.
12. Wendl M, Wilson R: **Aspects of coverage in medical DNA sequencing**. *BMC Bioinformatics* 2008, **9**:239.
13. The International HapMap Consortium: **A second generation human haplotype map of over 3.1 million SNPs**. *Nature* 2007, **449**:851–861.
14. Howie BN, Donnelly P, Marchini J: **A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies**. *PLoS Genet* 2009, **5**(6):e1000529.
15. Kennedy J, Măndoiu I, Paşaniuc B: **Genotype Error Detection and Imputation using Hidden Markov Models of Haplotype Diversity**. *Journal of Computational Biology* 2008, **15**(9):1155–1171.
16. Li Y, Abecasis GR: **Mach 1.0: Rapid Haplotype Reconstruction and Missing Genotype Inference.** *American Journal of Human Genetics* 2006, **79**:2290.
17. Marchini J, Howie B, Myers S, McVean G, Donnelly P: **A new multipoint method for genome-wide association studies by imputation of genotypes**. *Nature Genetics* 2007, **39**:906–913.

18. Stephens M, Scheet P: **Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation.** *American Journal of Human Genetics* 2005, **76**:449–462.

19. Wen X, Nicolae DL: **Association studies for untyped markers with TUNA**. *Bioinformatics* 2008, **24**:435–437.

20. Marchini J, Howie B: **Genotype imputation for genome-wide association studies.** *Nature reviews. Genetics* 2010, **11**(7):499–511.

21. Browning B, Yu Z: **Simultaneous Genotype Calling and Haplotype Phasing Improves Genotype Accuracy and Reduces False-Positive Associations for Genome-wide Association Studies**. *The American Journal of Human Genetics* 2009, **85**(18):847–861.

22. Nyholt DR, Yu CE, Visscher PM: **On Jim Watson's APOE status: genetic information is hard to hide**. *European Journal of Human Genetics* 2008, **17**(2):147–149.

23. Applied Biosystems: **SOLiD 4 System product description** [https://products.appliedbiosystems.com/].

24. Burton PR, Hansell AL, Fortier I, Manolio TA, Khoury MJ, Little J, Elliott P: **Size matters: just how big is BIG?: Quantifying realistic sample size requirements for human genome epidemiology**. *Int. J. Epidemiol.* 2009, **38**:263–273.

25. Ewing B, Green P: **Base-calling of automated sequencer traces using phred. II. Error probabilities.** *Genome Research* 1998, **8**(3):186–194.

26. Ghahramani Z, Jordan M: **Factorial Hidden Markov Models**. *Mach. Learn.* 1997, **29**(2-3):245–273.

27. Fine S, Singer Y, Tishby N: **The Hierarchical Hidden Markov Model: Analysis and Applications**. *Mach. Learn.* 1998, **32**:41–62.

28. Kimmel G, Shamir R: **A block-free hidden Markov model for genotypes and its application to disease association**. *Journal of Computational Biology* 2005, **12**:1243–1260.

29. Rastas P, Koivisto M, Mannila H, Ukkonen E: **Phasing genotypes using a Hidden Markov model**. In *Bioinformatics Algorithms: Techniques and Applications*, Wiley 2008, preliminary version in *Proc. WABI 2005*:355–373.

30. Schwartz R: **Algorithms for Association Study Design Using a Generalized Model of Haplotype Conservation**. In *Proc. CSB* 2004:90–97.

31. Baum L, Petrie T, Soules G, Weiss N: **A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains**. *Annals of Mathematical Statistics* 1970, **41**:164–171.

32. Lyngsø R, Pedersen C: **The consensus string problem and the complexity of comparing hidden Markov models**. *Journal of Computer Systems Science* 2002, **65**(3):545–569.

33. Gusev A, Mandoiu I, Pasaniuc B: **Highly Scalable Genotype Phasing by Entropy Minimization**. *IEEE/ACM Trans. on Computational Biology and Bioinformatics* 2008, **5**(2):252–261.

34. Kurtz *et al* S: **Versatile and open software for comparing large genomes**. *Genome Biology* 2004, **5**(2):R12.

35. Li H, Ruan J, Durbin R: **Mapping short DNA sequencing reads and calling variants using mapping quality scores**. *Genome Research* 2008, **18**:1851–1858.

36. Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K, Wang J: **SNP detection for massively parallel whole-genome resequencing**. *Genome Research* 2009, **19**:1124–1132.

37. Li H, Handsaker B, Wysoker A, Fennell T, Ruan *et al* J: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009, **25**(16):2078–2079.

38. Li Y, Abecasis G: **Thunder (beta version)** Oct 2010, [http://genome.sph.umich.edu/wiki/Thunder].

39. Quang L, Durbin R: **QCALL**. *in preparation.*

40. Kennedy J, Mandoiu I, Pasaniuc B: **GEDI: Scalable Algorithms for Genotype Error Detection and Imputation**. Tech. Rep. 0911.1765, Cornell University arXiv e-print 2009, [http://arxiv.org/abs/0911.1765].

41. Hong H, Su Z, Ge W, Shi L, Perkins R, Fang H, Xu J, Chen J, Han T, Kaput J, Fuscoe J, Tong W: **Assessing batch effects of genotype calling algorithm BRLMM for the Affymetrix GeneChip Human Mapping 500 K array set using 270 HapMap samples**. *BMC Bioinformatics* 2008, **9**(Suppl 9):S17.

42. Illumina: **Empowering GWAS for a new era of discovery** [http://www.illumina.com/documents/products/technotes/technote_empower_gwas.pdf].

Figures

Figure 1

HF-HMM model for multilocus genotype inference.

Figure 2

Schematic state diagram for the HMMs M and M' used in the reduction of the consensus string problem to MMGPP.

Figure 3

Genotype calling accuracy of compared methods for homozygous (a) and heterozygous (b) SNPs of the NA18507 Illumina dataset.

Figure 4

HMM posterior decoding accuracy (a) and distribution of reference allele coverage ratios for heterozygous SNPs (b) on the Watson 454, NA18507 Illumina, and NA18507 SOLiD datasets.

Figure 5

Effect of local recombination rate (a) and minor allele frequency (b) on concordance of genotypes called by the HMM posterior decoding algorithm on the NA18507 Illumina dataset.

Figure 6

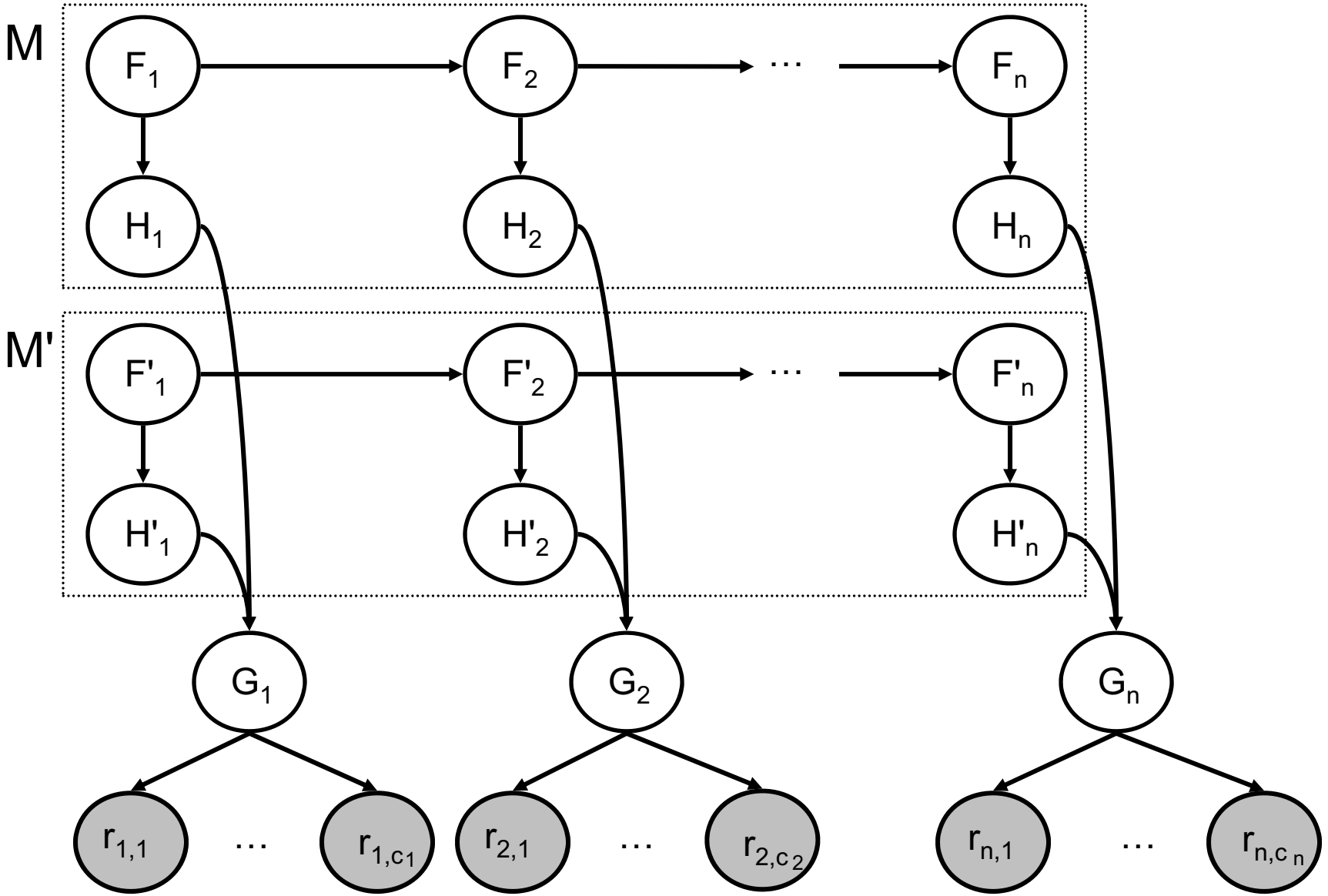
Effect of the reference panel size (a) and tradeoff between concordance and calling rate (b) for genotypes called by the HMM posterior decoding algorithm on the Watson 454 dataset.

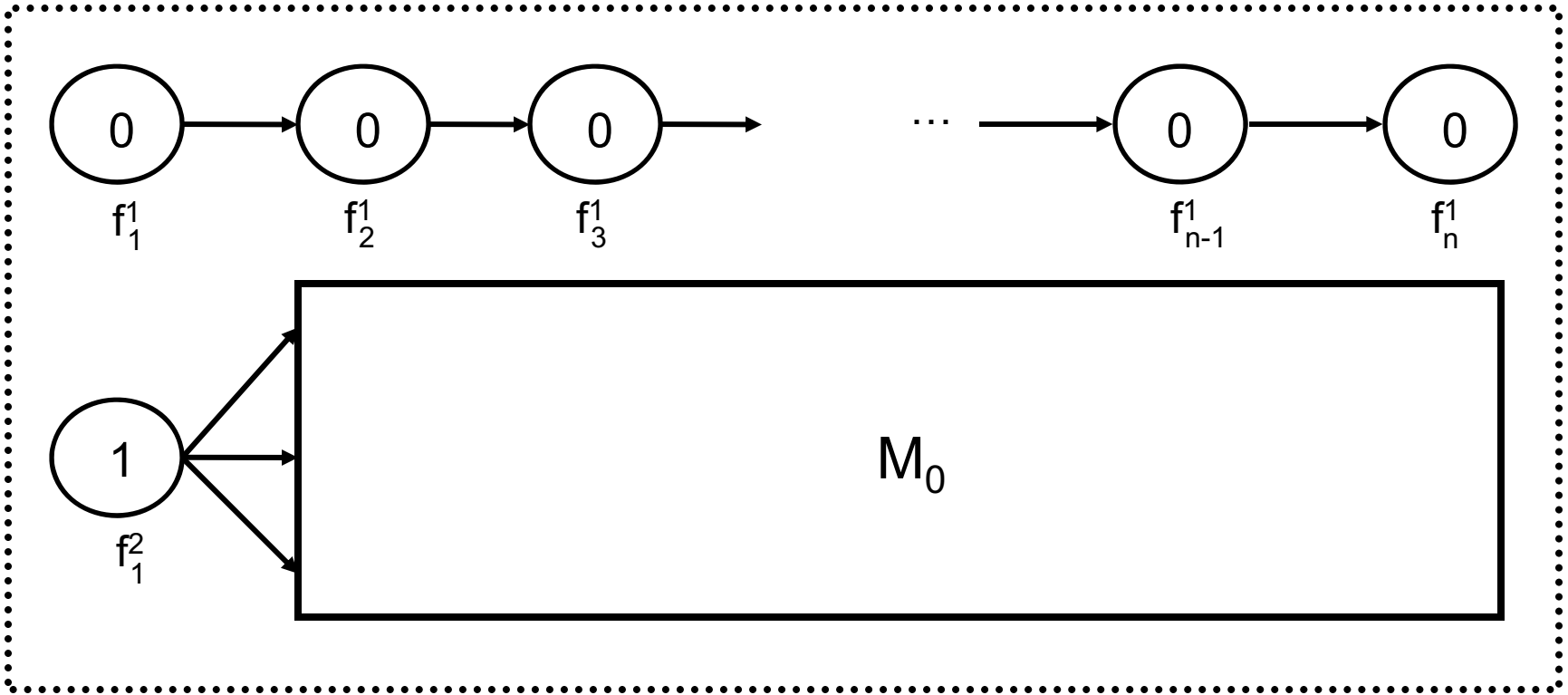
Tables

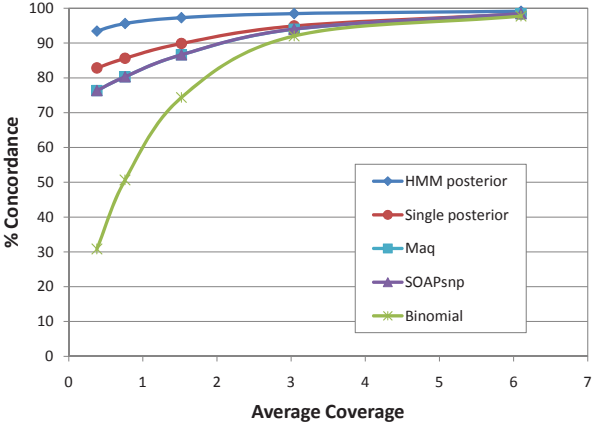
Table 1

Summary statistics for the three datasets used in evaluation.

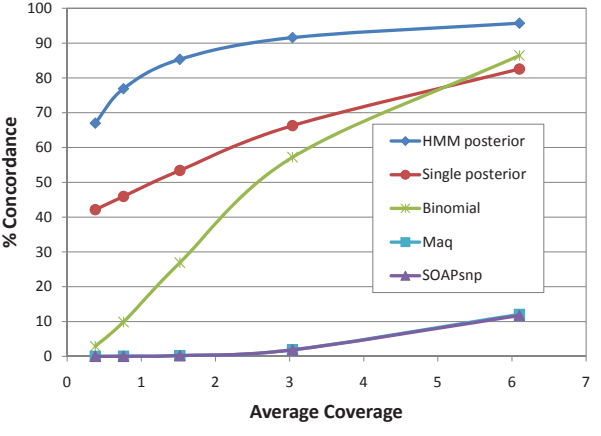
Dataset	Test SNPs	Raw Reads	Raw Sequence	Mapped Reads	Avg. Mapped SNP coverage
Watson 454	$443K$	$74.2M$	$19.7Gb$	$49.8M$ (67%)	5.85×
NA18507 Illumina	$2.85M$	$525M$	$18.9Gb$	$397M$ (78%)	6.10×
NA18507 SOLiD	$2.85M$	$2.45G$	$75Gb$	$900M$ (37%)	9.85×


M
F1
F2
…
Fn
H1
H2
Hn
M'
F'1
F'2
…
F'n
H'1
H'2
H'n
G1
G2
Gn
r1,1
…
r1,c1
r2,1
…
r2,c2
rn,1
…
rn,cn

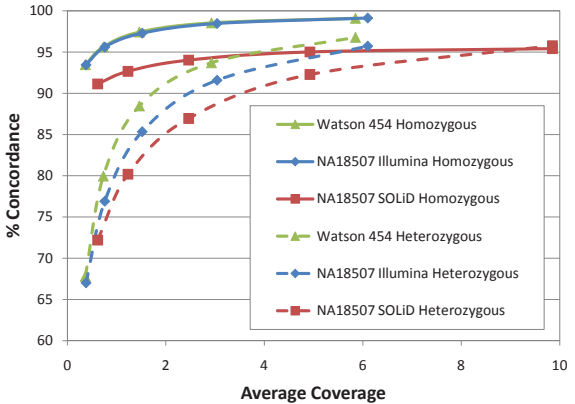

0
0
0
…
0
0
f^1_1
f^1_2
f^1_3
f^1_{n-1}
f^1_n
1
f^2_1
M_0


% Concordance
Average Coverage
HMM posterior
Single posterior
Maq
SOAPsnp
Binomial

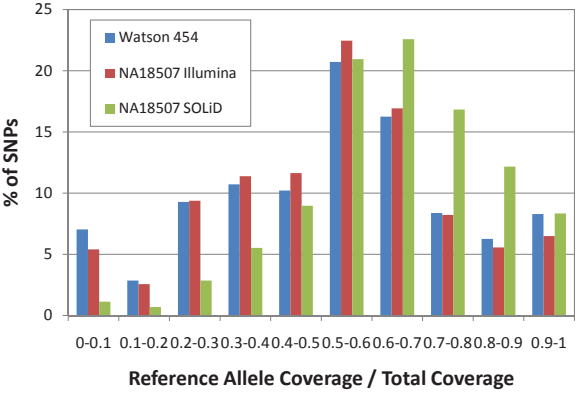
(a)


% Concordance
Average Coverage
HMM posterior
Single posterior
Binomial
Maq
SOAPsnp

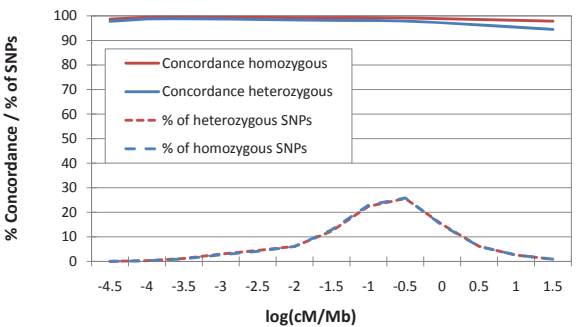
(b)


% Concordance
Average Coverage
Watson 454 Homozygous
NA18507 Illumina Homozygous
NA18507 SOLiD Homozygous
Watson 454 Heterozygous
NA18507 Illumina Heterozygous
NA18507 SOLiD Heterozygous

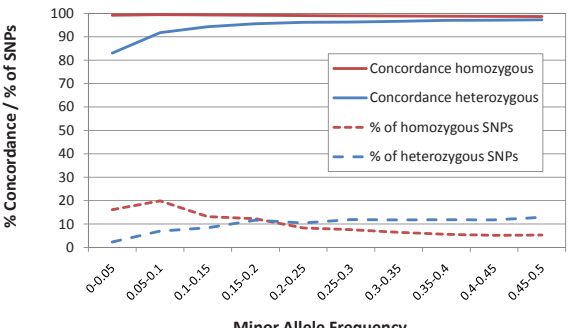
(a)


% of SNPs
Watson 454
NA18507 Illumina
NA18507 SOLiD
0-0.1
0.1-0.2
0.2-0.3
0.3-0.4
0.4-0.5
0.5-0.6
0.6-0.7
0.7-0.8
0.8-0.9
0.9-1
Reference Allele Coverage / Total Coverage

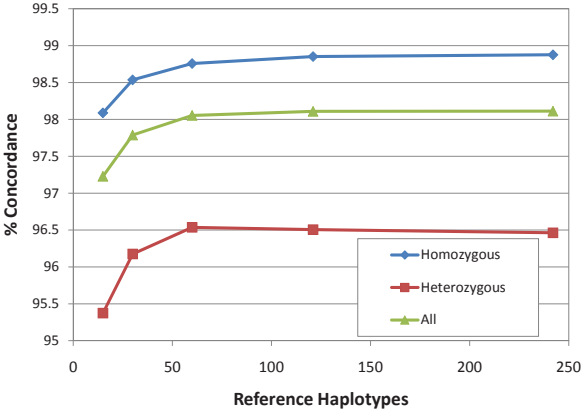
(b)


% Concordance / % of SNPs
100
90
80
70
60
50
40
30
20
10
0
Concordance homozygous
Concordance heterozygous
% of heterozygous SNPs
% of homozygous SNPs
-4.5
-4
-3.5
-3
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
log(cM/Mb)

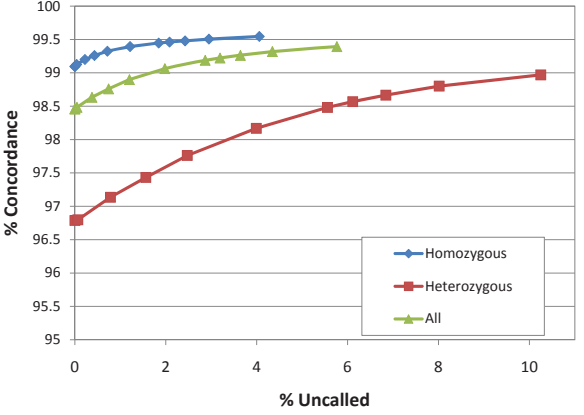
(a)


% Concordance / % of SNPs
100
90
80
70
60
50
40
30
20
10
0
Concordance homozygous
Concordance heterozygous
% of homozygous SNPs
% of heterozygous SNPs
0-0.05
0.05-0.1
0.1-0.15
0.15-0.2
0.2-0.25
0.25-0.3
0.3-0.35
0.35-0.4
0.4-0.45
0.45-0.5
Minor Allele Frequency

(b)


% Concordance
99.5
99
98.5
98
97.5
97
96.5
96
95.5
95
0
50
100
150
200
250
Reference Haplotypes
Homozygous
Heterozygous
All

(a)


% Concordance
100
99.5
99
98.5
98
97.5
97
96.5
96
95.5
95
0
2
4
6
8
10
% Uncalled
Homozygous
Heterozygous
All

(b)