SNP detection and genotyping from low coverage sequencing data on multiple diploid samples

*Si Quang Le and Richard Durbin**

Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus,

Cambridge, CB10 1SA, United Kingdom.

**Corresponding rd@sanger.ac.uk*

Abstract

Reductions in the cost of sequencing have enabled whole genome sequencing to identify sequence variants segregating in a population. An efficient approach is to sequence many samples at low coverage then to combine data across samples to detect shared variants. Here we present methods to discover and genotype single nucleotide polymorphism (SNP) sites from low coverage sequencing data, making use of shared haplotype (linkage disequilibrium) information. For each population, we first collect SNP candidates based on independent sequence calls per site. We then use MARGARITA with genotype or phased haplotype data from the same samples to collect 20 ancestral recombination graphs (ARGs). We refine the posterior probability of SNP candidates by considering possible mutations at internal branches of the 40 marginal ancestral trees inferred from the 20 ARGs at the left and right flanking genotype sites. Using a population genetic prior on tree branch length and Bayesian inference we determine a posterior probability of the SNP being real, and also the most probable phased genotype call for each individual. We present experiments on both simulation data and real data from the 1000 Genomes Project to prove the applicability of the methods. We also explore the relative tradeoff between sequencing depth and the number of sequenced samples. Software to implement the methods is available in the QCALL package from ftp://ftp.sanger.ac.uk/pub/rd/QCALL.

Introduction

Recent advances in sequencing technologies enable the sequencing of personal genomes to identify most genetic variation present in one sample (Venter et al. 2001; Levy et al. 2007; Wang et al. 2008; Wheeler et al. 2008; Kim et al. 2009). To achieve high accuracy at almost all accessible sites requires high average depth, for example the average depth in (Kim et al. 2009) is 27.8x. This high depth is expensive and limits the number of samples that can be sequenced. An alternative strategy to find sequence variants shared in a population was introduced in (Liti et al. 2009) where 70 haploid yeast samples were sequenced with only 1-4x coverage to find sequence variants. The 1000 Genomes Project is taking a similar approach and in its low coverage pilot has sequenced 179 samples at average 3.7x coverage (The 1000 Genomes Project Consortium 2010).

Several methods have been introduced to detect variants from sequencing individual genomes (Li et al. 2008; Li et al. 2009). The standard approach is to estimate the likelihood of sequencing data given possible genotypes and then convert to the probability of genotypes given data using Bayes' rule with an assumption about the prior probability of heterozygous and homozygous sequence variants. These methods work well with high coverage data but have low power and unacceptable false positive rates (FPR) when applied to individual samples with low coverage sequencing data. For example, Li et al. (2009b) reported 0.04% false positive rates per base pair (bp) for a single sample with 4x coverage data, implying cumulative false positive rates would go up to $1-(1-0.0004)^{100} \approx 4\%$ per bp, or 40 per kb, when applied to 100 independent samples. The rate of true SNPs would be expected to be approximately 6 SNPs per kb $\left(\theta \sum_{i=1}^{199} 1/i = 5.873 \times 10^{-3}\right)$ meaning that false positives would outnumber true SNP calls by ~7 to 1, giving ~87% false discovery rate (FDR). Consistent with this, when we use SAMtools (Li et al. 2009) separately on 100 samples with 4x coverage as described below, we see cumulative false positive rates of 5% per bp (see below). Moreover, genotype error rates when analyzing low coverage samples independently are, not surprisingly, high: 0.041, 0.283, and 0.030 for homozygous reference, heterozygous, and homozygous non-reference genotypes respectively.

In this paper we present two new methods to discover SNPs from low coverage sequencing data by combining data across samples, which were developed to detect SNPs in the low coverage pilot in the 1000 Genomes Project. In the first method, non linkage disequilibrium analysis (NLDA), we apply a

dynamic programming algorithm to estimate the posterior probability of k non-reference alleles in 2m chromosomes in O(m2) time for all values of k from 1 to 2m-1. Having obtained the posterior probability of k non-reference alleles in 2m chromosomes, we calculate the probability of a SNP at a site by the probability of $k>0$ given assumptions about variant frequency and allele frequency distribution. This method can be applied to the whole genome of hundreds of samples in reasonable computing time.

In the second method, linkage disequilibrium analysis (LDA), we make use of shared haplotype structure to estimate posterior probabilities of SNPs and genotypes. To do this, we build a set of possible ancestral recombination graphs (ARG) for samples using MARGARITA (Minichiello and Durbin 2006) on genotypes or phased haplotypes at previously genotyped sites. For example, we built 20 ARGs for samples in the low coverage pilot data of the 1000 Genomes Project from genotypes/phased haplotypes from the HapMap3 project (The International HapMap 3 Consortium 2010). Having built the ARGs, for each candidate SNP site we collect marginal ancestral trees inferred at the left and right flanking genotyped sites, 40 in total. We estimate the SNP posterior probability by evaluating the likelihood of the observed sequencing data for all possible mutations in the 40 trees, assuming that any sequence variant in the m samples is caused by a single mutation. Both simulated and real data show that LDA has the same SNP discovery rate as NLDA and produces lower false positive rates. However, the complexity of LDA, $O(N_A m^2 n_t)$ with number of nucleotides $N_A = 4$ and the number of trees $n_t=40$, makes LDA inapplicable to analyse the whole genome with hundreds of samples. Fortunately, we found that very few sites with low NLDA posterior probability have high LDA posterior probability, and so we adopt a strategy in which we first collect potential SNP candidates using NLDA with a threshold selected to ensure that the SNP candidate set is feasible for LDA. Then we apply LDA to the SNP candidate set and use the posterior probability of LDA to determine SNPs at a chosen threshold. We filter false positive calls by removing sites where there are 3 SNP calls within 10 bp (FW10) (Li et al. 2008). We can impute genotypes and phased haplotypes of m samples under the same LDA framework. These methods have been used to provide one of the primary call sets for the low coverage pilot of the 1000 Genomes Project (The 1000 Genomes Project Consortium 2010).

Results

We implemented QCALL as described in Methods.

NLDA and LDA comparison on simulation data

We simulated 3,000 haplotypes across a 5 Mbp region of chromosome 20 as described in Data section. Then we created five nested populations with 1,600x sequencing coverage in total, 50 samples with 32x coverage, 100 samples with 16x coverage, 200 samples with 8x coverage, 266 samples with 6x coverage, and 400 samples with 4x coverage. There are 24,289, 28,181, 31,675, 32,807, and 34,807 SNPs respectively in these simulated populations. We also simulate a population of 60 samples that have the same sequencing depths (3.7x average coverage) as the 60 CEU samples from the low coverage pilot of the 1000 Genomes Project (CEU samples are from Utah residents with Northern and Western European ancestry).

We applied both NLDA and LDA methods (see Materials and Methods) to the 400 samples with 4x coverage to understand performance of these two methods in SNP calling (Figure 1). It is clear that LDA is better than NLDA in detecting SNPs as it provides a lower false positive rate and a higher discovery rate. We applied a filter to remove sets of three or more SNP calls within 10 bases (FW10) as we found that most of the calls are false positives caused by misalignment of reads around short insertions or deletions (indels). FW10 helps to lower the false positive rates and keeps almost the same power in detecting true SNPs. In Table 1 we show the number of false positives for each sequencing strategy at a 0.99 posterior confidence level (Q20) with FW10 filter to obtain the false positives of each strategy (Table 1). We found that false positives in simulated data are mainly caused by indels, e.g., 929/942 false positives of 400 samples with 4x happen within 5 base pairs (bp) of indels (Table 2). A stronger filter FW5 that removes sets of two SNP calls within 5bp reduces the false positive number further to 510, but also removes many more true positives (overall loss of 9.8% true positives). An alternative way to filter false positives around indels would be to realign reads around indels, as is possible with DIndel (Albers et al. 2010) and GATK (McKenna et al. 2010). If these removed all false positives around indels, then we could in theory obtain a false positive rate in simulated data of about 1 per Mbp (1x5/24,804 ~ 0.0002 FDR) for 50 samples with 32x or about 2.6 per Mbp (2.6*5/29,823~0.0004 FDR) for 400 samples with 4x.

Number of sequenced samples versus sequencing depth

With low coverage sequencing data, it is difficult to detect SNPs with low non reference allele frequency (the total number of non-reference allele among 2*m* haplotypes of *m* samples-nrAF) as the lower the nrAF, the smaller the chance to observe sequencing data that supports the non-reference (alternative) allele (Figure 2). This issue becomes more serious for heterozygous SNPs when they need data to

support both alleles. For example, the marginal rate of detecting singleton SNPs drops from 99% with 32x coverage data to 18% with 4x coverage data. However, for fixed sequencing budget one can sequence more samples with low coverage than at high coverage. Below we show that we can increase the total number of population variants found by decreasing coverage and increasing the number of sequenced samples.

Starting with the 24,289 SNPs in the first 50 samples (100 haplotypes), we detect 24,029 (98.9%) SNPs from 32x coverage sequencing data. The marginal discovery rate (the fraction of SNPs with a particular nrAF that are discovered) is therefore 99% for singleton SNPs (Figure 1). We miss 240 SNPs, most of which are singletons. When we sequence more samples with lower coverage, we start to miss some SNPs from the 100 haplotypes of the first 50 samples but we gain SNPs in the additional sequenced samples. For example, when we reduce the sequencing depth from 32x to 16x we lose 187 SNPs from the first 100 haplotypes, but gain 3628 new SNPs from the 100 new sequenced haplotypes. Table 2 shows how these net gains progress as we sequence more samples at lower depth. The most variants are found when we sequence 400 samples of 4x coverage.

If we look at detection power as a function of nrAF calculated from all 3,000 sequences in the simulation, 400 samples at 4x also show the best power to detect SNPs with 1% nrAF, although at lower population frequencies, 266 samples at 6x give slightly higher power (Figure 3). The strategy of 50 samples with 32x coverage shows the worst performance at low nrAF; for example it detects about 40% SNPs with 0.005 nrAF, while that of 400 samples with 4x is about 75%. The simulation results indicate that the strategy of sequencing a large number of samples with low depths (4x-6x) is better than that of sequencing a small number of samples with high depths in detecting rare SNPs. However there is no difference between these strategies in detecting high nrAF SNPs, e.g., all strategies get to 100% discovery rates for SNPs with nrAF > 5%.

CEU samples of Pilot 1 in the 1000 Genomes Project

We analyse the same 5 Mbp region on chromosome 20 (43,000,000-48,000,000) in 60 samples from the CEU population of the low coverage pilot of the 1000 Genomes Project (see subsection Data). The corresponding call set on the full genome contributed to the results from the low coverage pilot of the 1000 Genomes Project (The 1000 Genomes Project Consortium 2010). We first applied NLDA to select 61,308 SNP candidates with 1% threshold. Then we used LDA to select 16,954 SNP calls with 90% threshold (Q10). Of these calls, 31% are in HapMap2 and 67% are in dbSNP, equivalent to 33% novel

calls. The calls show a ratio between transitions (mutations between A and G, or between C and T) and transversions (mutations from A or G to C or T, or vice versa) of 2.28, which is consistent with the value of 2.30 for the final 1000 Genomes Project call set in this interval (The 1000 Genomes Project Consortium 2010), though above the genome average of approximately 2.1.

We applied NLDA and LDA to the 60 simulated samples that have the same depths as the CEU samples from the low coverage pilot of the 1000 Genomes Project. Results on these simulation data show that we are able to detect about 19,077 SNPs from 25,268 SNPs from 60 samples with 3.7x coverage data (~75%). We called 456 false positives, equivalent to $456/(5\times10^6) \sim 10^{-4}$ FPR or $456/19533 \sim 2.33\%$ FDR. We also compare the marginal discovery rate of QCALL as a function of the non reference allele frequency on simulation data and real data, using the 43 samples in the 1000 Genomes Project CEU sample for which there is HapMap2 genotype data to provide the truth for the real data calls. The power as a function of allele frequency is remarkably similar (see Figure 4).

Genotype accuracy

One advantage of the LDA method is the ability to provide more accurate genotypes estimated from low coverage data based on a local structure haplotype. For example, the NLDA genotype estimator, which generates the posterior probability of genotypes by using Bayes' rule, has an error rate about 0.424 for heterozygous SNPs. LDA, however, assigns genotypes/haplotypes for samples by averaging over sets of calls that are consistent with local haplotype structure (see Methods subsection Genotyping).

Empirical experiments on the CEU population 1000 Genomes Project data comparing with HapMap II genotypes not at HapMap 3 sites (which were used to build the ARGs) give an overall genotype false discovery rate for LDA of 2.7%, corresponding to 1.4%, 3.9%, 4.2% FDR for homozygous, heterozygous, and homozygous non-reference genotypes respectively (see Table 3). These FDRs are competitive with those of Beagle (Browning and Yu, 2009), which is another haplotype-based approach to genotype calling from likelihood data comparable to LDA (2.8% overall, 0.8%, 5.7%, 3.4% by genotype category).

For simulation data, the overall genotype FDR of QCALL drops from 2.56% to 1.94% when we increase the number of sequenced samples from 50 to 400. We believe this decrease under represents the potential of the tree based calling approach of QCALL, and is instead limited by the ability of MARGARITA to scale effectively to large sample sets, since we have noticed that for 400 samples MARGARITA, which implements a greedy algorithm, gets locked into incorrect structures. We are exploring other approaches to generating ARGs to avoid this problem.

Discussion

Detecting SNPs from multiple samples with low coverage data is an efficient approach to detect low frequency SNPs in a population. Experimental results show that QCALL with NLDA and LDA methods detects shared variants from multiple samples better than analysing individual samples independently. In particular, the genotype accuracy is substantially improved.

The probability to detect a SNP at a site depends on the number of non-reference alleles present in the sequencing samples, and the evidence in the sequencing data for the observation. The strategy of sequencing a large number of samples with low coverage increases the expected number of non-reference alleles in the sample but lowers confidence the evidence for seeing them, compared to the strategy of sequencing a small number of samples with high coverage. The best strategy for a particular nrAF is a tradeoff between the two factors. For example, at 0.005 nrAF the probability of there being at least one non-reference allele in 50 samples (100 haplotypes) is 0.3942 and so the resulting discovery rate cannot be higher than 0.3942 even at very high depth. However, the probability of there being at least one non reference allele in 400 samples (800 haplotypes) is 0.9819, and it is likely there will be more than one, so the discovery rate at 4x is about 73%. However, there is almost no difference between two strategies for high nrAF SNPs (common SNPs) as both strategies achieve near 100% power. Even when the overall power to detect variants is similar, there are circumstances in which sequencing a larger number of samples at lower depth can be preferable, such as to better characterize the allele frequency of variants, or when phenotyped samples are being sequenced for an association study and increasing the number of sequenced samples increases statistical power.

Many false positive calls are caused by short indels where sequencing reads are mapped wrongly to the reference, particularly when the indels occur at the beginning or end of the reads. Thus, we often found a set of false positives around an indel. FW10 is a simple and quite efficient method to remove the false positives as they are often very dense around the indel. However, FW10 cannot solve the problem when there are fewer than 3 false positives or the false positives are separated by more than 10 bps. An alternative solution is to realign reads around indels, as is possible with DIndel (Albers et al. 2010) and GATK (ref). LDA gives good quality SNP calls but it has two main limitations, first it is computationally expensive, and second it requires ARGs to have been previously created from genotyped data. The computational cost can be overcome by prescreening with NLDA to filter out sites without evidence of being SNPs. QCALL takes about 10 hours for one Mbp segment of 400 samples but a proportion of the 10 hours are used to prepare likelihoods of sequencing data from multiple samples.

To solve the requirement for genotype data, we are developing methods to add new samples into existing ARGs or build ARGs directly from sequencing data.

Although our discussion of the method and results has been in the context of full genome shotgun data, QCALL can also be used on targeted sequencing data such as from exome projects (Ng et al. 2010), given that genotype data are available from which to build ARGs with MARGARITA. Furthermore, it can be used for other types of bi-allelic variant that are in local linkage disequilibrium with SNPs, such as small insertions or deletions (indels), by limiting to two possible states rather than the four bases. For these other uses it is possible to change the prior expectation of the transition to transversion ratio from 2, which is typical for human whole genome SNPs, to for example 3.5 which is typical of coding regions, or 1 when encoding other variant types. QCALL was used for calling short indel genotypes for the 1000 Genomes Project pilot (The 1000 Genomes Project Consortium 2010).

Finally, the LDA approach we discussed here is related to other haplotype sharing imputation methods such as BEAGLE (Browning and Yu 2009) mentioned above, IMPUTE (Howie et al. 2009), or MACH (Li et al. 2010). These can all be adapted for variant calling from low coverage sequencing, and in fact both BEAGLE and MACH have been also used in the 1000 Genomes Project with the results being combined with those from QCALL to provide final consensus calls (1000 Genomes Project Consortium, 2010).

Methods

Data

All experimental results were obtained on a 5 Mbp region of chromosome 20 (43,000,001 – 48,000,000) in NCBI 36 human reference (International Human Genome Consortium Apr 2006).

Simulation data

We simulated 3,000 haplotypes using MaCs with the same population parameters provided in (Chen et al. 2009). We used Maq to simulate 51 bp paired end reads for 800 haplotypes with error parameters estimated from one Illumina lane of NA12750 (The 1000 Genomes Project Consortium 2010). We mapped the reads to Human Genome reference NCBI 36 using BWA (Li and Durbin 2009) and transform into the BAM format. We build simulated "HapMap3" sites by identifying SNPs from 10 haplotypes and selected the same number of sites as in HapMap3 data, taking the nearest site seen twice in the 10 simulated haplotypes to each true HapMap3 site. We simulated 5 sets of data with a total 1600x

coverage: 50 samples with 32x coverage, 100 samples with 16x coverage, 200 samples with 8x coverage, 266 samples with 6x coverage and 400 samples with 4x coverage. We also simulated 60 samples with 3.7x to model the data from the 60 CEU samples of the low coverage pilot in the 1000 Genomes Project.

Real data

We used the same 5 Mbp region (chromosome 20, 43-48Mbp) of the CEU population in the low coverage pilot of the 1000 Genomes Project.

Non Linkage Disequilibrium Analysis (NLDA)

Assume we have observed data $D = (d_1,...,d_m)$ of m samples at site s and likelihoods $p(d_i \mid g_i)$ for d_i given possible genotype g. $p(d_i \mid g_i)$ can be estimated using Samtools (Li et al. 2009) or GATK (McKenna et al. 2010). For example, Samtools employs the method of (Li et al. 2008) where homozygous likelihoods $p(d_i \mid g_i = aa)$ are calculated as the product of estimated base errors for non-a bases from the sequencing quality values, corrected for non-independence of errors, and heterozygous liklelihoods $p(d_i \mid g = ab)$ as $1/2^{n_a+n_b}$ times the product of estimated base errors for non-ab bases, since there is a half chance of observing an a or b (see Samtools (Li et al. 2009) and Maq (Li et al. 2008) for more detail).

Assume the haplotypes of *m* samples at a site come from bi-allelic alleles, *a* and *b*. Obviously, the posterior probability of a SNP at *s* given observed data *D*, $p(s{=}SNP|D)$ is 1 - the probability of 2*m* haplotypes being equal to the reference allele *r* at *s*.

$$p(s = SNP \mid D) = 1 - p(\mathbf{g} = (g_1,..., g_n) : g_i = rr \mid D) = 1 - \frac{p(D \mid \mathbf{g})\, p(\mathbf{g})}{\sum_{\mathbf{g}'} p(D \mid \mathbf{g}')\, p(\mathbf{g}')} \quad (1)$$

where a configuration $\mathbf{g} = (g_1,..., g_m)$ is the genotypes of *m* samples, $p(\mathbf{g})$ and $p(D \mid \mathbf{g})$ are the prior probability of $\mathbf{g}$ and the probability of *D* given genotypes $\mathbf{g}$. The prior probability of a configuration is considered as the prior probability of a mutation that results in $\mathbf{g}$, $p(k) \sim \frac{\theta}{k}$ where θ is the population scaled mutation rate and *k* is the number mutant alleles in $\mathbf{g}$. Denote n_a be the number of allele *a* in $\mathbf{g}$,

$$p(\mathbf{g}) = \begin{cases} \frac{\theta}{2}\left(\frac{1}{n_a} + \frac{1}{2m - n_a}\right)\frac{1}{C^{2m}_{n_a(\mathbf{g})}} & 2m > n_a > 0 \\ \frac{1}{2}\left(1 - \theta\sum_{i=1}^{2m-1}\frac{1}{i}\right) & \text{otherwise} \end{cases}$$

We set $\theta = 0.001$ for standard human SNP calling. It can be set as a program parameter for other uses of QCALL.

Assuming that the sequencing data is independent between samples, the probability of *D* given *m* genotypes $\mathbf{g} = (g_1,...,g_m)$, $p(D \mid \mathbf{g})$, is calculated as

$$p(D \mid \mathbf{g} = (g_1,...,g_m)) = \prod_{i=1}^{m} p(d_i \mid g_i) \quad (2)$$

The key to calculating $p(s = SNP \mid D)$ in Equation (1) is to compute the normalization factor, $\sum_{\mathbf{g}} p(D \mid \mathbf{g})\, p(\mathbf{g})$. We have

$$\sum_{\mathbf{g}} p(D \mid \mathbf{g}) = \sum_{k} p(k) \sum_{\mathbf{g}:n_a(\mathbf{g})=k} p(D \mid \mathbf{g}) = \sum_{k} p(k) Q_{m,k}$$

where $p(k) = \theta\left(\frac{1}{k} + \frac{1}{2m-k}\right)\frac{1}{C^{2m}_{k}}$ and $Q_{m,k} = \sum_{\mathbf{g}:n_a(\mathbf{g})=k} p(D \mid \mathbf{g})$ is the total probability of all possible genotype configurations $\mathbf{g}$ of *m* samples such that the number of *a* alleles in $\mathbf{g}$ is equal to *k*.

$$\begin{aligned} Q_{m,k} &= \sum_{\mathbf{g}=(g_1,...,g_m):n_a(\mathbf{g})=k} p(D \mid \mathbf{g}) \\ &= \sum_{\mathbf{g}_{m-1}:n_a(\mathbf{g}_{m-1})=k-2} p(D_{m-1} \mid \mathbf{g}_{m-1})\, p(d_m \mid g_m = aa) \\ &\quad + 2\times \sum_{\mathbf{g}_{m-1}:n_a(\mathbf{g}_{m-1})=k-1} p(D_{m-1} \mid \mathbf{g}_{m-1})\, p(d_m \mid g_m = ab) \\ &\quad + \sum_{\mathbf{g}_{m-1}:n_a(\mathbf{g}_{m-1})=k} p(D_{m-1} \mid \mathbf{g}_{m-1})\, p(d_m \mid g_m = bb) \\ &= Q_{m-1,k-2}\, p(d_m \mid g_m = aa) + 2Q_{m-1,k-1}\, p(d_m \mid g_m = ab) + Q_{m-1,k}\, p(d_m \mid g_m = bb) \end{aligned}$$

where $Q_{m-1,k}$ presents for the total probability of all figurations of *m*-1 samples such that the number of allele *a* among *m*-1 samples equals to *k*.

Using this recursion, we can calculate $Q_{m,k}$ from the individual genotype likelihoods $p(d_i \mid g_i)$ in $O(m^2)$ steps by dynamic programming. Having obtained $Q_{m,k}$, we can easily estimate $\sum_{\mathbf{g}} p(D \mid \mathbf{g}) p(\mathbf{g})$ in Equation (1)

Linkage Disequilibrium Analysis (LDA)

First we give an informal description, then the technical details. We assume that genetic variants are caused by a single mutation on a coalescent tree during evolution. Figure 5 shows an example of an ancestral tree at some site for 4 samples, s_1, s_2, s_3, and s_4. Assuming for the moment that this tree is correct, and that we know the ancestral base value and the position of a mutation on the tree, for example the mutation from A to C shown in Figure 5, we can infer the base for each haplotype at the site, and hence the genotypes of the individuals. Given the genotypes we can calculate the likelihood of the sequencing data *D* given the tree, the root value and the mutation. Since we do not know the root value and the mutation site we integrate over them, weighting the mutation probabilities by the expected branch length under a population genetic prior, and then we average over a sample of trees to provide an estimate of the total likelihood of data *D* given that there was a mutation. Conditional on there being a mutation, we marginalize over genotypes to generate genotype posteriors.

To find the coalescent trees, we use MARGARITA (Minichiello and Durbin 2006) to estimate ancestral recombination graphs (ARGs) from known genotypes or phased haplotypes at samples at sites genotyped on SNPs. We prefer phased haplotypes to unphased genotypes because MARGARITA works better with phased data. However it can work on unphased genotype data, or a mixture. For example, for the low coverage pilot of the 1000 Genomes Project we used phased haplotypes from HapMap 3 for most samples, but genotypes for a few samples for which phased haplotypes were not available. For the simulation data we used phased haplotypes at a subset of sites selected to correspond to HapMap 3 sites as described above.

MARGARITA (Minichiello and Durbin 2006) can only handle a limited number of SNPs (markers) in terms of running time and memory, and therefore we cut the whole genome into 1Mbp segments. To make ARGs consistent at the ends of the 1 Mbp segments, we expanded 0.5 Mbp at each end of each 1 Mbp segment, so MARGARITA was run across overlapping intervals of 2Mb. We kept 20 ARGs for further analysis as a compromise between QCALL's accuracy and running time. A higher number of

ARGs does not improve the performance of QCALL much but increases running time linearly. For example, MARGARITA takes on average approximately 8 hours to build 20 ARGs for 400 samples on one 2Mb segment.

MARGARITA only gives trees at the sites that were used to build it. We approximated coalescent trees at candidate SNP sites *s* by the trees *T* at the left and right flanking genotyped sites. Let Δ and $\overline{\Delta}$ be the two cases where there is one and no mutation at *s*. We compute the probability of a mutation at s given *D* by Bayes' rule:

$$p(\Delta \mid D,T) = \frac{p(D \mid \Delta,T)\,p(\Delta \mid T)}{p(D \mid \Delta,T)\,p(\Delta \mid T) + p(D \mid \overline{\Delta},T)\,p(\overline{\Delta} \mid T)} \tag{3}$$

where the priors $p(\Delta \mid T) = \theta \sum_{i=1}^{2m} \frac{1}{i}$ and $p(\overline{\Delta} \mid T) = 1 - p(\Delta \mid T)$ are derived from standard neutral population genetics theory.

We start solving Equation (3) by estimating the probability of *D* given no mutation, $p(D \mid \overline{\Delta},T)$. To handle the situation where there are errors in the reference sequence, we set

$$p(D \mid \overline{\Delta},T) = \sum_{r} p(D \mid \overline{\Delta},T,r)\,p(r) \ ,$$

where *r* is the true (ancestral) unmutated reference

$$p(r) = \begin{cases} 1-\varepsilon & r = \text{reference allele of NCBI 36} \\ \varepsilon/3 & \text{otherwise} \end{cases} \tag{4}$$

where ε is the error rate in the observed reference, which we set to $\varepsilon = 2\times10^{-5}$ based on empirical experiments in the 1000 Genomes Project. Given true base *r* and no mutation at *s*, all genotypes of *m* samples must be *rr*, leading to

$$p(D \mid \overline{\Delta},T) = \sum_{r} p(r) \prod_{i=1}^{m} p(d_i \mid g_i = rr)$$

To estimate $p(D \mid \Delta,T)$, we scan all possible mutations on trees of T, and integrate the probabilities of D given these mutations weighted by a prior distribution over mutations. Let us start with reference *r,*

$$p(D \mid \Delta, T) = \sum_r p(r)\, p(D \mid \Delta, T, r) \\ = \sum_r p(r) \sum_{\mathbf{t}_k} p(D \mid \Delta, \mathbf{t}_k, r) p(\mathbf{t}_k) \tag{5}$$

where $\mathbf{t}_k$ is a tree at flanking site of *s*. We assume trees $\mathbf{t}_k$ are independent and have the same prior probability, $p(\mathbf{t}_k) = 1/|T|$.

To estimate $\sum_{\mathbf{t}_k} p(D \mid \Delta, \mathbf{t}_k, r)$, we scan all possible mutations in $\mathbf{t}_k$ such that the reference *r* must exist among *m* genotypes. We also consider the case where *r* is not represented in the *m* observed samples and was caused by a mutation outside $\mathbf{t}_k$.

$$p(D \mid \Delta, \mathbf{t}_k, r) = \mu \sum_{a \neq r} p(a,r)\, p(D \mid g_i = aa : i = 1..m) \\ + (1-\mu) \sum_{e \in \mathbf{t}_k} \frac{1}{2} p(e \mid \mathbf{t}_k) \sum_{a \neq r} \left(p(a,r)\, p(D \mid e_{ar}) + p(r,a)\, p(D \mid e_{ra}) \right) \tag{6}$$

where μ is the prior probability of an external mutation from a to r, set to that of a mutation at a leaf branch of $2m+1$ leaves, p(a,r) is the prior probability of a mutation from a to r, $p(e \mid \mathbf{t}_k)$ is the prior probability of a mutation happening on edge e, and $p(D \mid e_{ar})$ ($p(D \mid e_{ra})$) is the probability of data given a mutation from a to r (r to a) at edge e.

The first part of Equation (6) allows for an external mutation from *a* to *r* outside $\mathbf{t}_k$ and the second part handles the case where a mutation happens at an edge in $\mathbf{t}_k$. μ is set proportional to $\frac{1}{2m+1}$ and normalized with $\sum_{i=1}^{2m+1} \frac{1}{i}$. The prior probability of a mutation from a to r, $p(a,r)$, can be set to allow for an arbitrary transition to transversion ratio. For standard genome wide calls we set this to be 2.0

$$p(a,r) = \begin{cases} 4/24 & (ar) \in \text{transition} = \{(AG),(GA),(CT),(TC)\} \\ 1/24 & \text{otherwise (transversion)} \end{cases}$$

$p(D \mid g_i = aa : i = 1..m)$ is simply estimated as

$$p(D \mid g_i = aa : i = 1..m) = \prod_{i=1}^{m} p(d_i \mid g_i = aa)$$

The prior probability of a mutation happening at *e* is set such that the more recent mutations have lower prior probability.

$$p(e) \propto \left(\frac{1}{n_a} + \frac{1}{2m - n_a} \right) \frac{1}{N(n_a)}$$

where n_a is the number of haplotype *a* at the leaves when mutation $a \rightarrow b$ happens at edge *e* and $N(n_a)$ is the number of possible mutations in $\mathbf{t}_k$ that result in n_a haplotype *a* at the leaves. We normalize $p(e \mid \mathbf{t}_k)$ such that $\sum_e p(e \mid \mathbf{t}_k) = 1$.

Let $\mathbf{g} = (g_1, ..., g_m)$ be the genotypes of *m* samples that result from mutation $a \rightarrow r$ (or $r \rightarrow a$) at edge *e*,

$$p(D \mid e_{ar}) = \prod_{i=1}^{m} p(d_i \mid g_i)$$

Merging Equations (5) and (6), we have

$$\begin{aligned} p(D \mid \Delta, T) = p(D \mid \Delta, \mathbf{t}_k, r) &= \mu \sum_{a \neq r} p(a, r)\, p(D \mid \mathbf{g} = \mathbf{aa}) \\ &+ \frac{(1-\mu)}{2} \sum_{r, a \neq r, \mathbf{t}_k, e \in \mathbf{t}_k} p(r)\, p(\mathbf{t}_k)\, p(e \mid \mathbf{t}_k) \big(p(a, r)\, p(D \mid e_{ar}) + p(r, a)\, p(D \mid e_{ra}) \big) \end{aligned} \quad (7)$$

We note that the complexity of computing $p(D \mid \Delta, T)$ in Equation (7) is $O(N_A m^2 n_t)$ where the number of nucleotides, $N_A = 4$, and number of flanking trees, n_t =40.

Genotyping

Let $\mathbf{g} = (g_1, ..., g_m)$ be the genotypes of *m* samples at *s*. Given a mutation at *s*, we calculate the posterior probability for g_i as follows:

$$p(g_i = ab \mid D, \Delta, T) = \sum_r p(r)\, p(g_i = ab \mid D, \Delta, T, r)$$

where *r* is the reference allele and $p(r)$ is the prior probability estimated as in Equation (4). $p(g_i = ab \mid D, \Delta, T, r)$ is given by

$$p(g_i = ab \mid D, \Delta, T, r) = \frac{p(D, g_i = ab \mid \Delta, T, r)}{\sum_{a'b'} p(D, g_i = a'b' \mid \Delta, T, r)}$$

where

$$p(D, g_i = ab \mid \Delta, T, r) = \begin{cases} 0 & a \neq b, a \neq r, b \neq r \\ \sum_{\mathbf{t}_k} p(\mathbf{t}_k)\, p(D, g_i = ab \mid \Delta, \mathbf{t}_k, r) & \text{otherwise} \end{cases}$$

Let E_{ik}^j be the set of edges in $\mathbf{t}_k$ where *j* (*j*=0,1 or 2) haplotype(s) of sample *i* are mutants caused by a mutation at $e \in E_{ik}^j$. $p(D, g_i = ab \mid \Delta, \mathbf{t}_k, r)$ is estimated under following cases:

If $a = b = r$, then $p(D, g_i = aa \mid r = a, \mathbf{t}_k, \Delta)$ is the sum of probabilities of all possible mutations from a to x on edge $e \in E_{ik}^0$ or from x to a on edge $e \in E_{ik}^2$.

$$p(D, g_i = aa \mid r = a, \mathbf{t}_k, \Delta) = \sum_{e \in E_{ik}^0} p(e \mid \mathbf{t}_k) \sum_{x \neq a} p(a, x) p(D \mid e_{ax}) + \sum_{e \in E_{ik}^2} p(e \mid \mathbf{t}_k) \sum_{x \neq a} p(x, a) p(D \mid e_{xa})$$

If $a = b \neq r$, then $p(D, g_i = aa \mid r \neq a, \mathbf{t}_k, \Delta)$ is the sum of probabilities of mutations from a to r on edges outside $\mathbf{t}_k$ or on edge $e \in E_{ik}^0$ or from r to a on edge $e \in E_{ik}^2$.

$$p(D, g_i = aa \mid r \neq a, \mathbf{t}_k, \Delta) = \mu p(a, r)\, p(D \mid \mathbf{g} = \mathbf{aa}) + (1 - \mu)\left[\sum_{e \in E_{ik}^0} p(e \mid \mathbf{t}_k)\, p(a, r)\, p(D \mid e_{ar}) + \sum_{e \in E_{ik}^2} p(e \mid t_k)\, p(r, a)\, p(D \mid e_{ra}) \right]$$

If $a \neq b$, then $p(D, g_i = ab \mid r, \mathbf{t}_k, \Delta)$ is the sum of a mutation from a to b or b to a on edge $e \in E_{ik}^1$. r must be either a or b.

$$p(D, g_i = ab \mid r, \mathbf{t}_k, \Delta) = \sum_{e \in E_{ik}^1} \frac{1}{2} \left[p(a, b)\, p(D \mid e_{ab}) + p(b, a)\, p(D \mid e_{ba}) \right]$$

Having obtained posterior genotype probabilities $p(g_i = ab \mid D, \Delta, T)$, we determine the genotype of sample i as the maximum likelihood genotype:

$$g_i = \arg\max_{ab} \left\{ p(g_i = ab \mid D, \Delta, T) \right\}$$

Haplotype Phasing

Let $\mathbf{h} = (h_1,...,h_{2m})$ be the $2m$ haplotypes of m samples at site s. We compute the posterior probability of $h_i = a$ given a mutation Δ, observed data D, marginal coalescent trees T as:

$$p(h_i = a \mid D, T, \Delta) = \sum_r p(r)\, p(h_i = a \mid D, T, \Delta, r)$$

where r is the reference allele and $p(r)$ is the prior probability estimated as in Equation (4).

$p(h_i = a \mid D, T, \Delta, r)$ is calculated as

$$p(h_i = a \mid D, T, \Delta, r) = \frac{p(D, h_i = a \mid T, \Delta, r)}{\sum_b p(D, h_i = b \mid T, \Delta, r)}$$

where

$$p(D \mid h_i = a, T, \Delta, r) = \sum_{\mathbf{t}_k} p(\mathbf{t}_k)\, p(D, h_i = a \mid \mathbf{t}_k, \Delta, r)$$

Denote $E_{i,k}$ be the set of edges in $\mathbf{t}_k$ such that a mutation $e \in E_{ik}$ results in h_i.

$$p(D, h_i = a \mid r = a, \mathbf{t}_k, D, \Delta) = \sum_{e \in E_{ik}} p(e \mid \mathbf{t}_k) \sum_{x \neq a} p(x, a)\, p(D \mid e_{xa}) + \sum_{e \notin E_{ik}} p(e \mid \mathbf{t}_k) \sum_{x \neq a} p(a, x)\, p(D \mid e_{ax})$$

and

$$p(D, h_i = a \mid r \neq a, \mathbf{t}_k, D, \Delta) = \mu p(D \mid \mathbf{h} = \mathbf{a}) +$$
$$(1 - \mu)\left[\sum_{e \in E_{ik}} p(e \mid \mathbf{t}_k)\, p(r, a)\, p(D \mid e_{ra}) + \sum_{e \notin E_{ik}} p(e \mid \mathbf{t}_k)\, p(a, r)\, p(D \mid e_{ar}) \right]$$

Having obtained $p(h_i = a \mid D, T, \Delta)$, we determine the haplotype of sample *i* as the maximum likelihood allele.

$$h_i = \arg\max_a \{ p(h_i = a \mid D, T, \Delta) \}$$

Issue with singletons and haplotype phasing

Singletons are a special case where a mutation happens at leaf branches. For each singleton, there are two possible mutations at leaf branches resulting in the same genotype configuration (Figure 6). This results in an equal posterior probability for both alleles at the singleton. Thus, we cannot phase singletons. In practice, when our genotype calls indicate there is a singleton (all homozygous except one

that is heterozygous) we only give a genotype call for the heterozygous sample and do not attempt to phase it.

Acknowledgments

We thank James Stalker, Thomas Keane and David Craig for making the .bam files, Heng Li for his significant help with Maq and samtools, Gary Chen for MaCs, the 1000 Genomes Project Consortium for their data, the HapMap3 Consortium for providing genotype calls and Gilean McVean group for phasing them, and Goncalo Abecasis and Richard Durbin groups for comments and feedback. Funding for this project was provided by Microsoft Research and Wellcome Trust grant WT089088/Z/09/Z to R.D.

FIGURE LEGENDS

Figure 1: Discovery and false positive rates of QCALL for 400 samples with 4.0x coverage sequencing data. LDA and LDA,FW10 stand for linkage disequilibrium analysis without FW10 and with FW10. The same notation is applied to NLDA.

Figure 2: SNP discovery power for different sequencing strategies, all using 1,600x data, plotted as a function of the number of non-reference alleles present in the sequenced samples.

Figure 3: SNP discovery power for different sequencing strategies as a function of the non-reference allele frequency in the population. The continuous lines show empirical results from the simulation with the allele frequency estimated from all 3,000 simulated haplotypes, and the dashed lines present calculations based on sampling with marginal discovery rates per sample from Figure 2 and the.

Figure 4: Marginal discovery rates as a function of non reference allele count in 43 samples, from the CEU simulation, and from 1000 Genomes Project data evaluated at HapMap 2 sites not in HapMap 3, on the 43 sequenced samples overlapping HapMap 2.

Figure 5: An illustrative example of a coalescent tree for 4 samples (8 haplotypes). Given a value at the root, A in this example, and a mutation from A to C in this example, we can infer genotypes for the 4 samples and hence compute the probability of data *D* conditional on this configuration. We estimate the likelihood of D given a tree $\mathbf{t}$, $p(D \mid \mathbf{t})$, by summing over all possible root values and mutations in $\mathbf{t}$.

Figure 6: Two mutations at two edges of a singleton (edges connected to haplotypes 4th or 8th) lead to the same genotype configuration

FIGURES

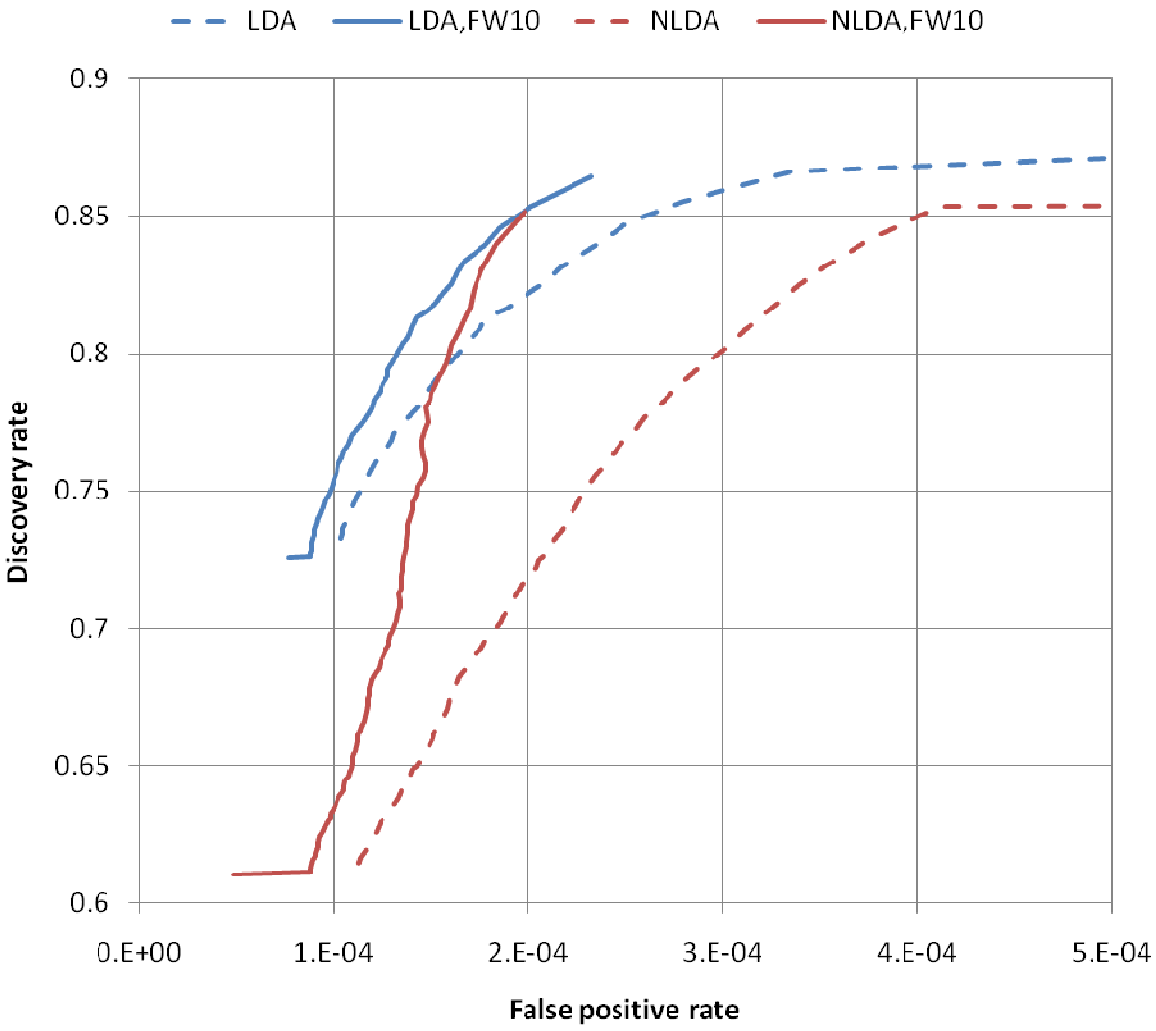

Figure 1

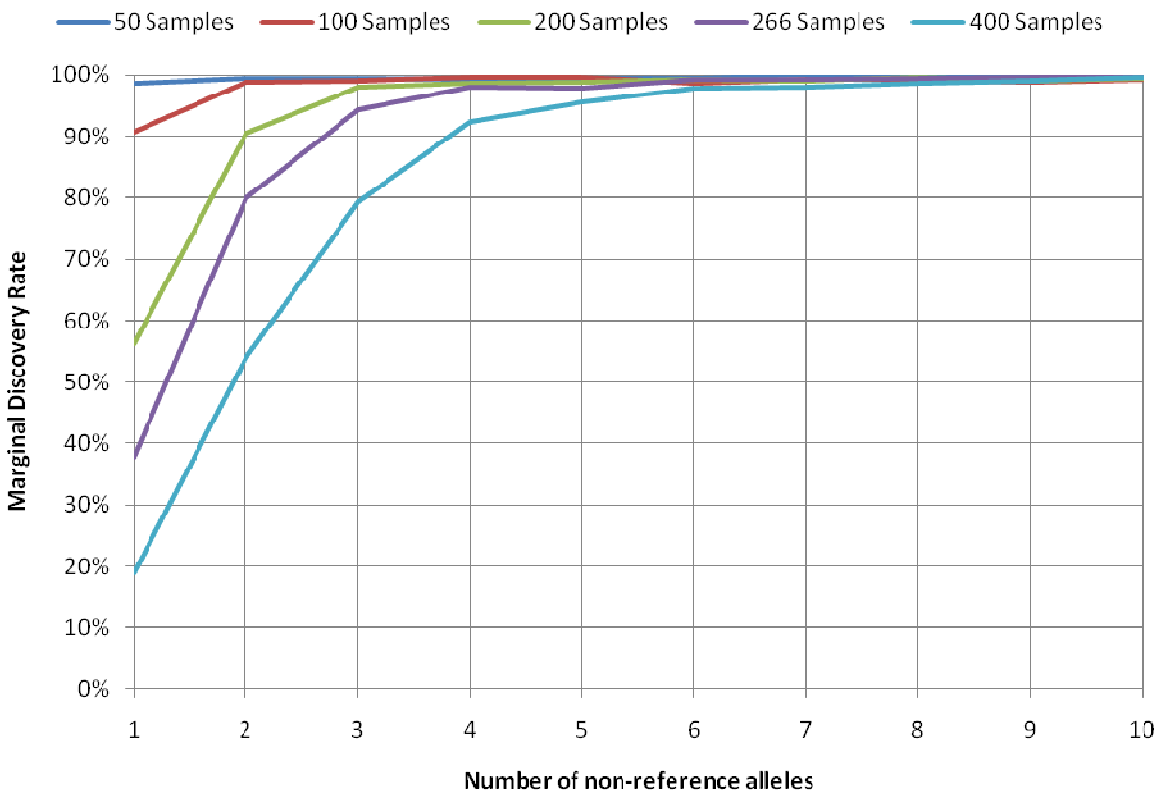
50 Samples
100 Samples
200 Samples
266 Samples
400 Samples
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Marginal Discovery Rate
1
2
3
4
5
6
7
8
9
10
Number of non-reference alleles

Figure 2

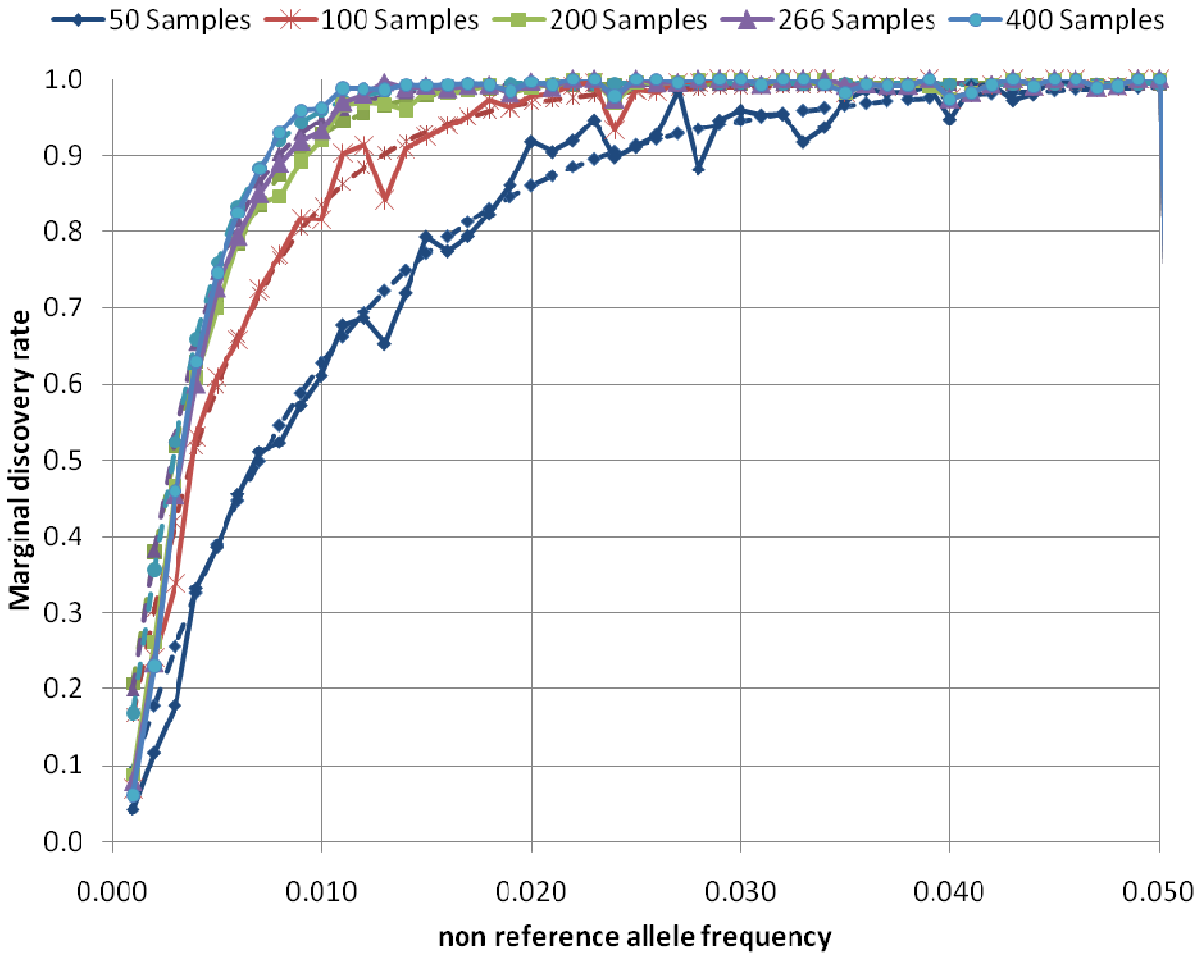
50 Samples
100 Samples
200 Samples
266 Samples
400 Samples
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Marginal discovery rate
0.000
0.010
0.020
0.030
0.040
0.050
non reference allele frequency

Figure 3

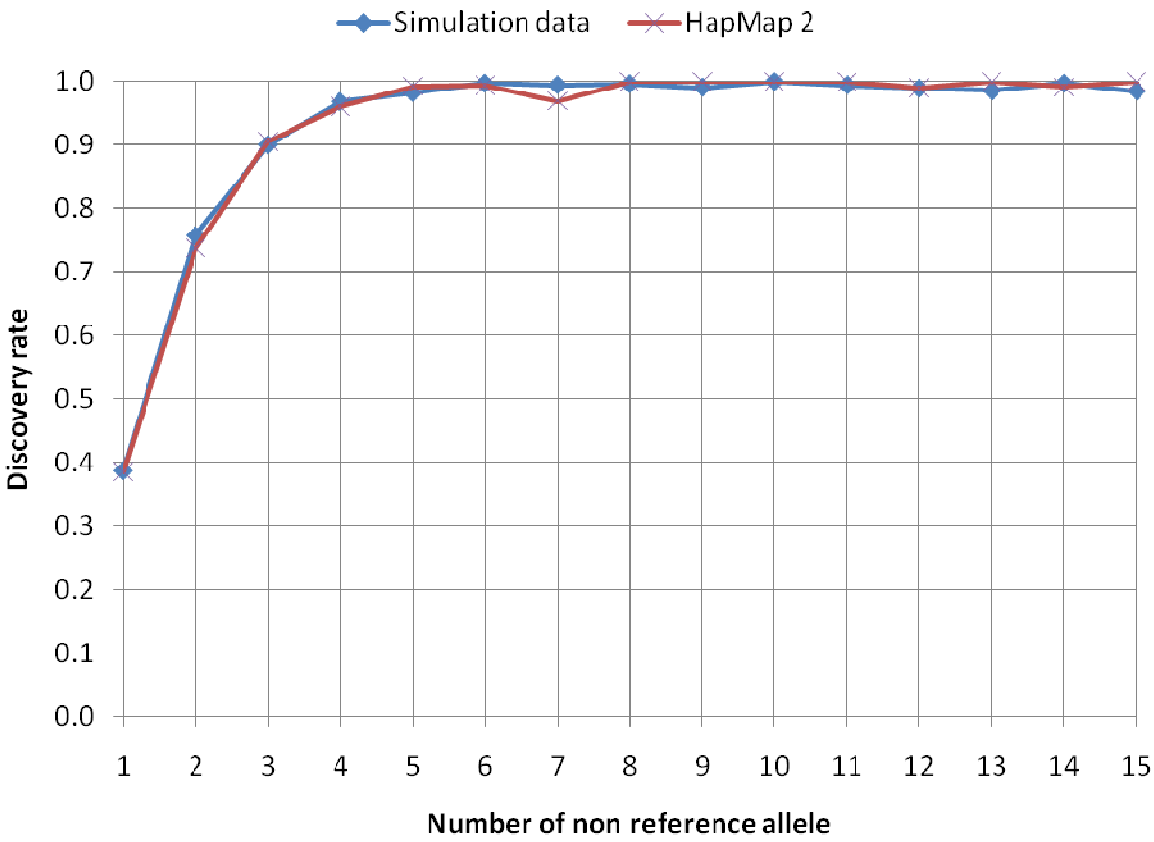
Simulation data
HapMap 2
Discovery rate
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Number of non reference allele

Figure 4

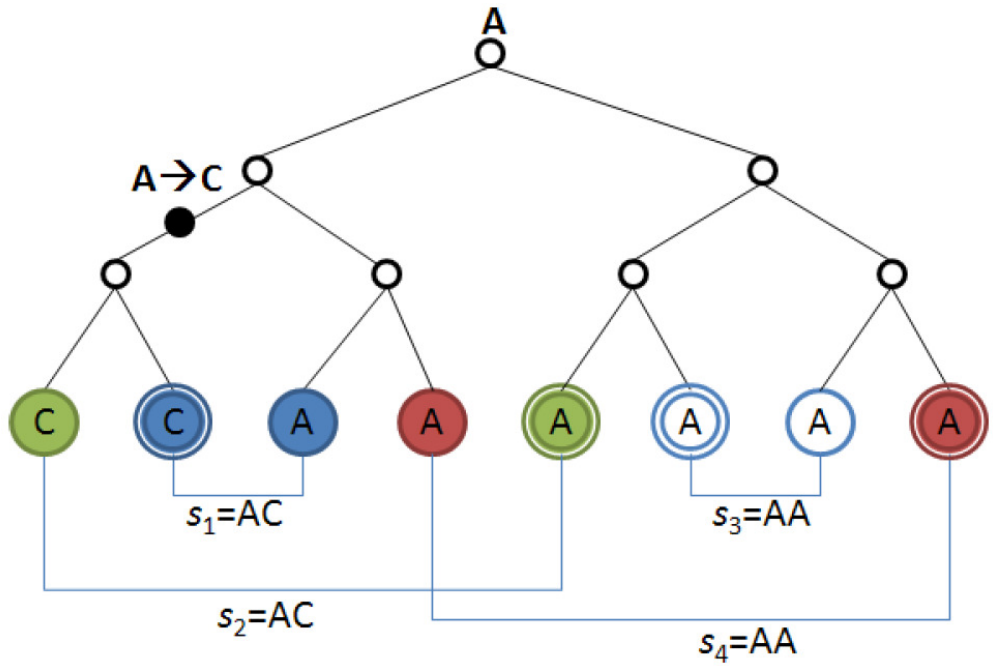

Figure 5

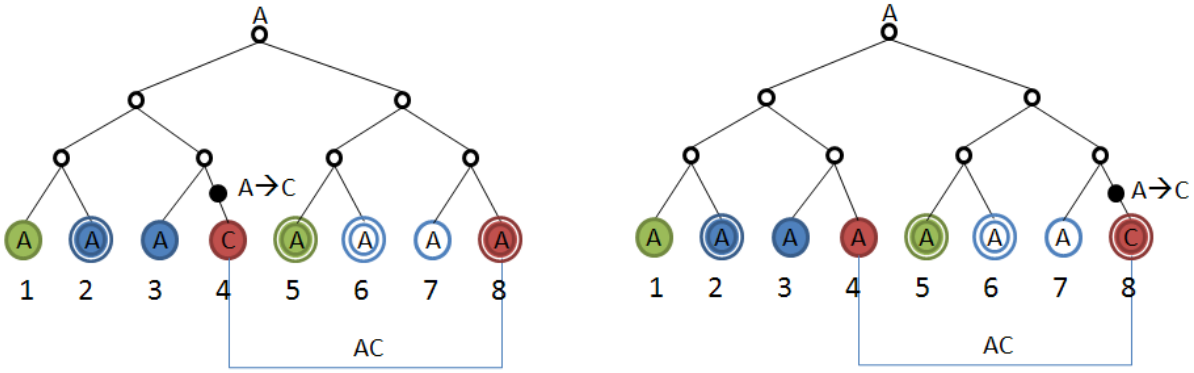

Figure 6

TABLES

Table 1: Distribution of false positives as a function of distance from the nearest indel.

Distance	50 samples	100 samples	200 samples	266 samples	400 samples
0	419	457	468	473	472
1	275	313	320	323	334
2	49	71	81	80	77
3	18	23	28	28	34
4	3	4	5	5	5
5	5	5	4	3	7
6	1		1	1	1
7	2	1	1	2	2
8	1		1	1	1
9				1	1
14				1	1
60					1
>200	1	3	7	4	6
Total	774	877	916	922	942

Table 2: Discovery rates with different sequencing strategies

	Samples	50	100	200	266	400
	#SNP	32x	16x	8x	6x	4x
100 Haps	24289	24029	23842	23438	23267	23148
200 Haps	28181		27470	26521	26156	25942
400 Haps	31674			28877	28251	27891
532 Haps	32807				28793	28353
800 Haps	34807					28880

Table 3: Average genotype error rates according to HapMap2 genotypes of 5 Mbp on chromosome 20 (20:43000000-48000000) of 43 samples that are the overlapped samples between the 60 CEU samples and HapMap2 samples. Error rates are calculated on 2711 QCALL sites that are in HapMap2 but not in HapMap3.

		QCALL			
		Hom	Het	Hom-nonref	error rate
HapMap2	Hom	55277	766	28	0.014
	Het	876	32107	411	0.038
	Hom-Nonref	334	681	23141	0.042

REFERENCES

Browning BL, Yu Z. 2009. Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. *Am J Hum Genet* **85**(6): 847-861.

Chen GK, Marjoram P, Wall JD. 2009. Fast and flexible simulation of DNA sequence data. *Genome Res* **19**(1): 136-142.

Howie BN, Donnelly P, Marchini J. 2009. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**(6): e1000529.

International Human Genome Consortium. Apr 2006. Build 36, hg18

Kim JI, Ju YS, Park H, Kim S, Lee S, Yi JH, Mudge J, Miller NA, Hong D, Bell CJ et al. 2009. A highly annotated whole-genome sequence of a Korean individual. *Nature* **460**(7258): 1011-1015.

Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G et al. 2007. The diploid genome sequence of an individual human. *PLoS Biol* **5**(10): e254.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**(14): 1754-1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**(16): 2078-2079.

Li H, Ruan J, Durbin R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18**(11): 1851-1858.

Liti G, Carter DM, Moses AM, Warringer J, Parts L, James SA, Davey RP, Roberts IN, Burt A, Koufopanou V et al. 2009. Population genomics of domestic and wild yeasts. *Nature* **458**(7236): 337-341.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M et al. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*.

Minichiello MJ, Durbin R. 2006. Mapping trait loci by use of inferred ancestral recombination graphs. *Am J Hum Genet* **79**(5): 910-922.

Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA et al. 2010. Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* **42**(1): 30-35.

The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population scale sequencing. *Nature* **Accepted**.

The International HapMap 3 Consortium. 2010. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**(7311): 52-58.

Venter JC Adams MD Myers EW Li PW Mural RJ Sutton GG Smith HO Yandell M Evans CA Holt RA et al. 2001. The sequence of the human genome. *Science* **291**(5507): 1304-1351.

Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, Guo Y et al. 2008. The diploid genome sequence of an Asian individual. *Nature* **456**(7218): 60-65.

Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT et al. 2008. The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**(7189): 872-876.