

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.

* E-mail: Corresponding rd@sanger.ac.uk

Abstract

Reduction in the cost of sequencing has enabled whole genome sequencing of many samples from populations to identify sequence variant. 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 at the site with 0.01 posterior probability. 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 benefit of depth sequencing and the number of sequenced samples. Software to implement the methods is available in the QCall package from www.sanger.ac.uk/software/QCall/.

Author Summary

We proposed two methods coming with packages for detecting SNPs from sequencing data of multiple samples. Empirical experiments shows that these two methods help to detect SNPs with better accuracy (low false positives) than analysing samples independently. The trading between sequencing depths and sequenced samples benefits in number of detected SNPs. We can detect about 95% of SNPs with $\text{nrAF} = 1\%$ from 400 samples with 4x coverage. This proportion goes down when we increase sequencing depth and reduce sequenced samples. However too low depth would loss power in detecting SNPs with $\text{nrAF} < 0.5\%$. We applied these two methods to detect SNPs from pilot 1 of 1000 Genomes Project.

Introduction

Recent advances in sequencing technologies enable the sequencing of personal genomes to identify genetic variant presented in one sample, e.g., a Yoruba African [1], two individuals of Northwest European origin [2], a person from China [3], and a person from Korea [4]. To achieve high accuracy at almost all accessible sites requires high average depth, e.g., the average depth of AK1 is 27.8x [4]. 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 [?] where 70 haploid yeast samples were sequenced with only 1-4x coverage to find sequence variants. The 1000 Genomes Project proposes to take a similar approach and in its pilot 1 project has sequenced 180 samples at average 2-4x coverage.

Several method shave been introduced to detect variants from sequencing individual genomes [5–7]. The standard approach is to the estimate 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 when applied to many samples with low coverage sequencing data. For example, Li et. al [6] reported 0.04% false positive

rates for a single sample with 4x coverage data, implying cumulative false positive rates would go up to $1 - (1 - 0.0004)^{100} = 4\%$ per bp when applied to 100 independent samples with 4x coverage data. Our experimental results on the data simulated (see Data section) using Samtool [7] accumulative false positive rates of 11% when analyzing separately 100 samples with 4x coverage.

In this paper we present methods to discover SNPs from low coverage sequencing data by combining data cross samples that were developed to detect SNPs in Pilot 1 data in 1000 Genomes Project. We introduce two methods to detect SNPs using combined sequencing data from m samples. In the first method, non linkage disequilibrium analysis (NLDA), We designed a dynamic programming algorithm to estimate the posterior probability of k non-reference alleles in $2m$ chromosomes in $O(m^2)$ time. From that we can calculate the probability of a SNP at a site by the probability of $k \geq 0$. This method can be applied to the whole genomes with hundreds of samples with reasonable computing time. In the second method, linkage disequilibrium analysis (LDA), we make use of shared haplotype structures to estimate posterior probabilities of SNPs. We built 20 possible ancestral recombination graphs for the full set of samples population using MARGARITA [8] on genotypes or phased haplotypes at previous work genotyping sites, e.g., the HapMap project for HapMap samples [9]. Then for each SNP candidate s , we collect 40 marginal trees inferred at the left and right flanking genotyped sites and estimate the SNP posterior probability by evaluating the likelihood of the observed sequencing data for all possible mutations in the 40 trees. We assume genotypes/haplotypes of m samples are caused by a mutation. Experimental results show 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 analysis the whole genomes with hundreds of samples. Fortunately, we found that very few sites with low NLDA posterior probability could end up with high LDA posterior probability. Thus, we propose a strategy in which we first collect potential SNP candidates using NLDA with a thresholds selected to ensure that the SNP candidate set is feasible for LAD and containing as many detectable SNPs as possible. Then we apply LDA to the SNP candidate

set and use the posterior probability of LDA to determine SNPs at a chosen thresholds. We filter false positive calls by removing sites where less than half of samples containing data (s50 filter). Then we removed triple SNP calls in 10 bp (FW10) [5]. We imputed genotypes and phased haplotypes of m samples under the same LDA framework. Simulation data reveals that we obtain about 95% of SNPs with $\text{nrAF} = 1\%$ or 99% SNPs with $\text{nrAF} \geq 1\%$.

Results and discussion

NLDA and LDA comparison on simulation data

We simulated 3000 haplotypes cross 5Mbp region of chromosome 20 as described in Data section. From that we simulate five nested population samples with 1600x 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. We observed 22302, 30832, 43907, 50905, and 62445 SNPs for these populations.

We then apply NLDA and LDA (section Materials and Methods) to the 5 Mbp region and report discovery rates and false positives in order to analysis behaviours of NLDA and LDA (Figure 1).

It is clear from Figure 1 that the discovery rate of NLDA is almost as the same as that of LDA while false positives of NLDA is much higher than that of LDA. For examples, discovery rates of NLDA at 0.75 confidence level equals to the discovery rate of LDA at 0.72 confidence level, 54.41%. However the corresponding false positive rate of NLDA is 9.07×10^{-5} , as nearly 1.5 times as that of LDA, 5.95×10^{-5} .

LDA is better than NLDA in term of false positive rates but LDA is much more expensive than NLDA in term of computation. In practice we cannot afford to apply LDA for the whole genomes of hundreds samples. Fortunately, we observed a strong correlation between posterior probabilities of NLDA and LDA. where very few sites with low NLDA posterior probability would end up with high LDA posterior probability. For example, we found only 11 sites in 400 samples with 4x coverage whose NLDA posterior

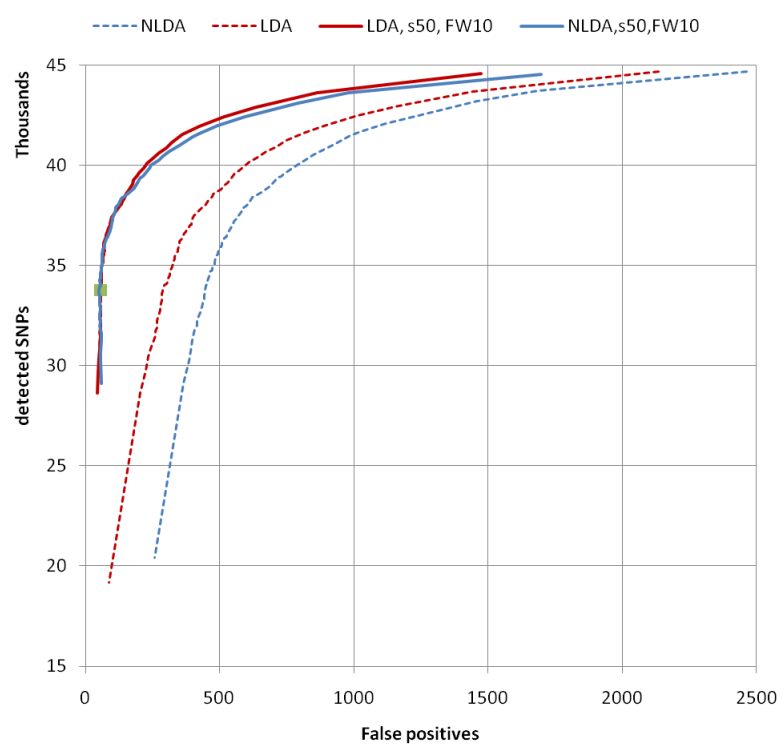


Figure 1. Discovery and false positive rates of 400 samples with 4X coverage sequencing data using LDA and NLDA

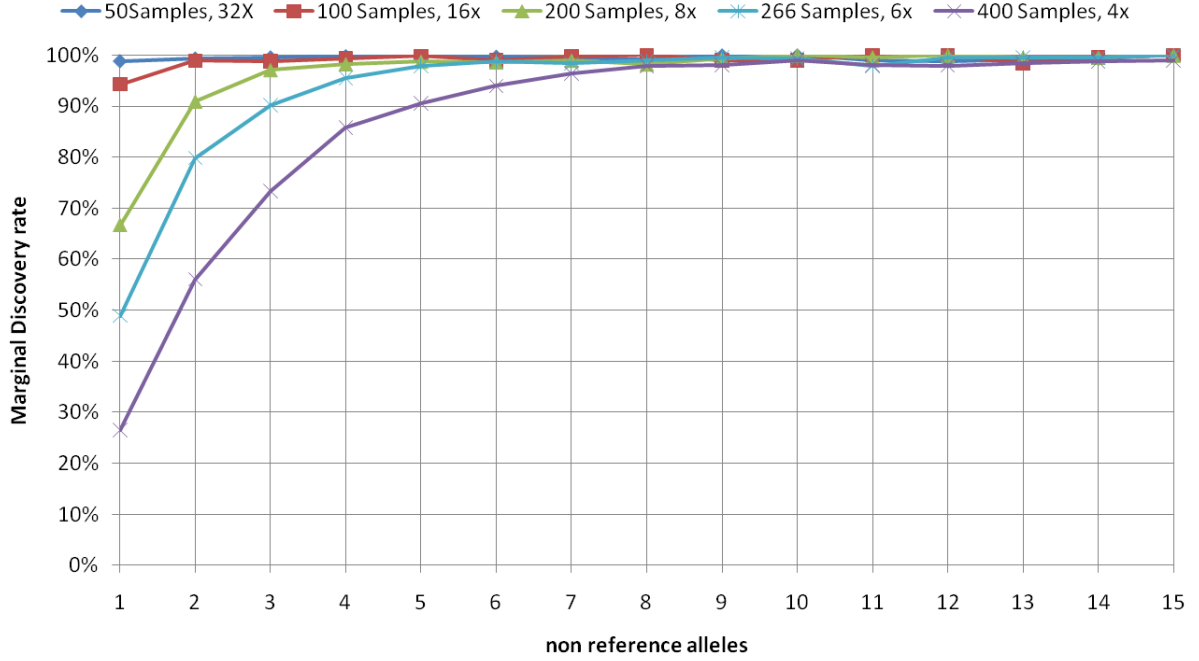


Figure 2. Marginal discovery rate from different populations.

probability is smaller than 0.5% and LDA posterior probabilities is bigger than 50%. This experiment result leads us to the strategy, in which we use NLDA to select SNP candidates from the whole genomes and use LDA to refine SNP posterior probabilities. In practice, we set the NLDA threshold at 0.1% to select about 40000 SNP candidates per mega base on average and then apply LDA to these candidates to collect SNPs with LDA confidence level.

To reduce false positives, we removed clustered SNPs (defined as 3 or more SNPs in within 10bp) (FW10), and remove SNPs for which less than half of samples are covered by any reads (s50). These filters work very well to drop the number of false positives while keeping the power in detecting SNPs. For example, the false positive would drop to about 56/5 mega bases (10^{-5}) with the power to detect about 33754/62445 (54%) SNPs when applied to SNPs with 0.75 LDA confidence level (the square point in Figure 1).

Obviously missing SNPs at any population samples have low nrAF because most of sequencing data

supports for NCBI 36 human reference and **against** very few sequencing data coming from rare non-reference alleles. (Figure 2). Marginal discovery rates are lowest at singletons or doubleton, specially with low coverage data. For example, the marginal rate of singleton SNPs drops from 99% when we have 32x coverage data to 27% with 4x coverage data. It is obvious as we would highly see no data coming from the singleton with 4x coverage. Even we observe some data coming from the singleton, the data (information) may be too weak to pass a confidence level. To level down confidence would bring more low nrAF SNPs but raise false positives caused by sequencing or mapping errors. How the marginal discovery rate of SNPs with $\text{nrAF} = 1\%$ would reach about 98-99% at any population samples.

Number of sequenced samples Venus sequencing depth

To determine the efficient way to find variants in a population, we applied the methods to these 5 population samples.

We start with 22302 SNPs of the first 50 samples (100 haplotypes) where we detect 22056 (98.9%) SNPs from 32x coverage sequencing data. At this depth, we have essentially the same power to find singleton SNPs at sample levels. The marginal discovery rate jumps immediately up to 99% for singleton SNPs (Figure 2). We miss 246 SNPs that are mainly singletons. This result indicates we could find nearly all SNPs of sequenced samples from 32x coverage data. However sequencing all genomes of a large human population with 32x coverage is impractical while sequencing a small a proportion of the population is unable to detect SNPs outside of the sequenced samples. For example, if we sequence only the first 50 samples of the 400 samples with 32x coverage, then we will obviously miss 40143 SNPs which come from the other 350 samples.

When we increase the number of sequenced samples and decrease the sequencing depth, we miss more SNPs from the first 100 haplotypes (50 samples) but we gain SNPs from added sequenced samples. For example, we miss another 281 SNPs of the first 100 haplotypes when reducing sequencing depth from 32x to 16x and obtain 8103 new SNPs from the 100 added sequenced haplotypes. Experiment results show

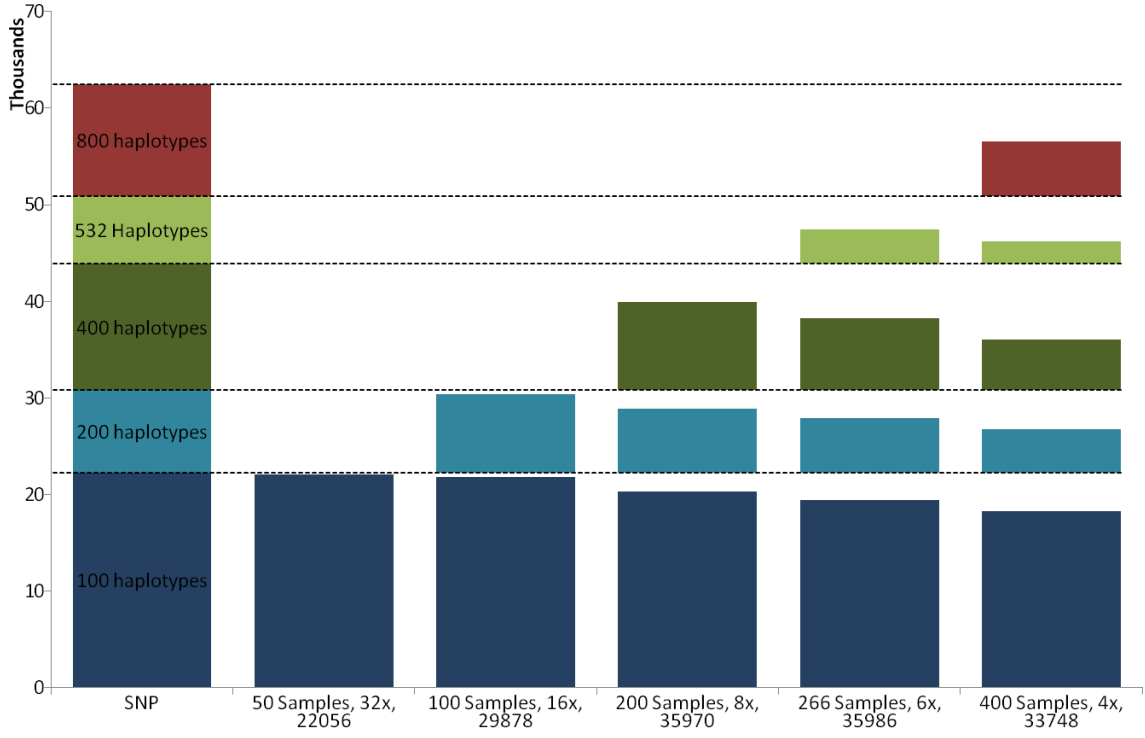


Figure 3. Number of detected SNPs with different sequencing strategy. The first column (SNP) indicates for the number of SNPs for each populations. Each of other columns presents for the number of detected SNPs, e.g., we detect 35970 SNPs from 200 samples with 8x, 20321 from the first 100 haplotypes, 6559 from the next 100 haplotypes and 9090 SNPs from the last 200 haplotypes.

that new detected SNPs always outnumber missing SNPs (Figures 2) when we lower the sequencing depth from 32x coverage.

The trading of the sequencing depth to the number of sequenced samples gains SNPs until sequencing depth reaches to 6x coverage. For example, we gain 7822 SNPs when decreasing 32x coverage to 16x coverage, and 6092 SNPs when decreasing 16x coverage to 8x coverage. We start losing power in detecting SNPs when lowering the sequencing depth from 6x to 4x coverage. That is caused by the low discovery rate of 4x coverage data for SNPs with low nrAF ($< 0.4\%$)(Figures 2 and 4). For example, we detect 21.2% SNPs with $\text{nrAF} = 0.1\%$ from 400 samples with 4x coverage, about 2.3% lower than that from 266

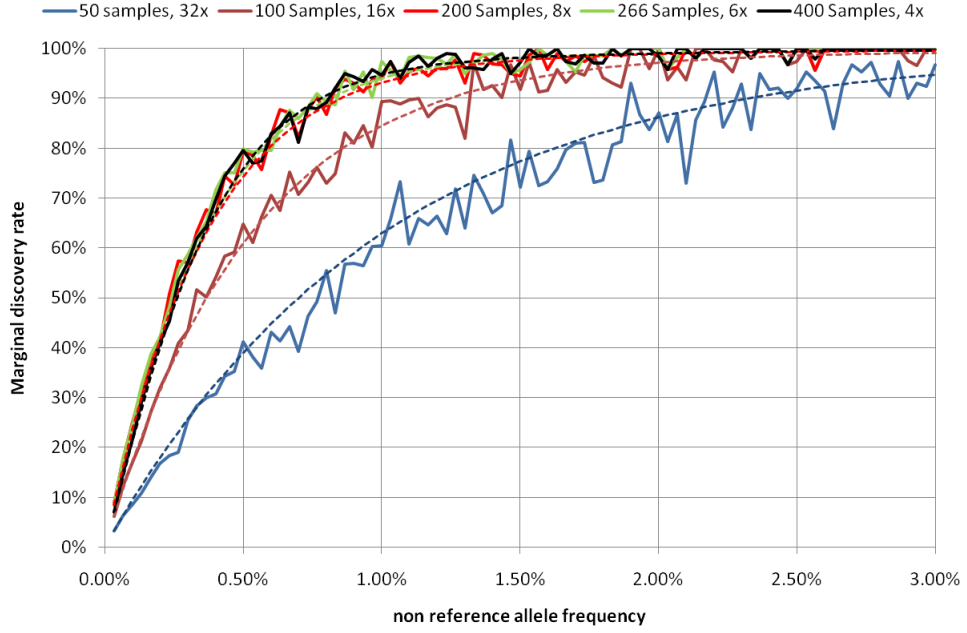


Figure 4. SNP detection power with different strategy. The dash lines present for estimations based on sampling with marginal discovered rates shown in Figure 2 and the continuous lines indicate the empirical results from a 3000 haplotype population.

samples with 6x coverage. The out performance of 400 samples with 4x to others starts from SNPs with $\text{nrAF} > 0.4\%$. Specially, we detect 94.7% SNPs with $\text{nrAF} = 1\%$ from 400 samples with 4x coverage, about 0.7% (1.6%) higher than 266 samples with 6x coverage (200 samples with 8x coverage).

CEU samples of Pilot 1 in 1000 Genomes Project

We used the same 5Mbp regions (chromosome 20:43M-48M) in CEU population (60 samples+Trios) to validate the methods. We first applied NLDA to select 61308 SNP candidates with 1% threshold. Then we used LDA to select 16954 SNPs with the 90% threshold. Table presents statistics of SNP calls that are competitive to the results of other groups in 1000 Genomes Project (personal communications). The full report of these methods on Pilot 1 in 1000 Genomes Project can be found at [10].

Materials and Methods

Data

All experimental results were obtained on a 5 Mbp region of chromosome 20 (43Mbp-48Mbp) in NCBI 36 human reference.

Simulation data

We simulated 3000 haplotypes using MaCs [11] with population parameters inferred from [12]. Recombination hot spots were incorporated as identified in HapMap2 [13]. We used Maq [5] to simulate 50bp reads for 800 haplotypes with error parameters estimated from one Illuminate lane of NA12750 [10]. We mapped the reads to Human Genome reference NCBI 36 using BWA [14] and refine the mapping using GATK [?]. We generated likelihood (GLF) using Samtool. We build simulated "HapMap3" sites by identifying SNPs from 10 additional simulated haplotypes. Then for each true SNP in HapMap3, we selected the closest simulated "HapMap3" SNPs and considered these selected sites as "HapMap3" sites for 3000 simulated haplotypes. We simulated 5 sets data with the total depth 1600x, 50 samples with 32x coverage, 100 samples with 16x coverage, 200 samples with 8x coverage, 266 sasmples with 6x coverage and 400 samples with 4x coverage.

Real data

We used the same 5 Mbp regions (chromosome 20, 43M-48 region) of CEU population in Pilot 1 of 1000 Genomes Project. The mother of Trios, NA12892, was included with 4x coverage data while the father and child (NA12878 and NA12891) were imputed without sequencing data. The data was downloaded from the web site of 1000 Genomes Project and likelihood of observed data given possible genotypes (GLF) is generated using Samtool.

SNPs	dbSNP	HapMap2	HapMap3	ts	tv	ts/tv	NovelSNP
16954	11368	5258	2785	11784	5169	2.280	31949

Table 1. Statistics on Chr20:43-48M, CEU with Trios. ts: transitions ($A \leftrightarrow G, C \leftrightarrow T$), tv: transversions ($A, G \leftrightarrow C, T$).

Non Linkage Disequilibrium Analysis (NLDA)

We now have observed data $D = \{d_1, \dots, d_m\}$ of m sample at site s and obtained probabilities of data d_i given possible genotype g , $p(d_i|g)$, using Samtool. We need to estimate the SNP probability of s given observed data D , $p(s = SNP|D)$. Obviously $p(s = SNP|D)$ is the probability where at least one haplotype among $2m$ haplotypes is different from human reference allele r inferred from NCBI 36. The probability is equivalent to the complement when $2m$ haplotypes are all identical to reference allele r , $p(s = SNP|D) = 1 - p(g_1 = g_2 = \dots = g_m = rr|D)$ where g_i is the genotype of sample i^{th} . Denote $\mathbf{g} = (g_1, \dots, g_m)$

$$p(s = SNP|D) = 1 - p(\mathbf{g} = \mathbf{rr}|D) = 1 - \frac{p(D|\mathbf{g})p(\mathbf{g})}{\sum_{\mathbf{g}'} p(D|\mathbf{g}')p(\mathbf{g}')} \quad (1)$$

where $p(\mathbf{g})$ and $p(D|\mathbf{g})$ are the prior probability of \mathbf{g} and the probability of D given genotypes \mathbf{g} .

Let a and b be the two alleles at site s . The prior probability of genotype $p(\mathbf{g})$ is approximated as

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

where n_a is the number of allele a in \mathbf{g} and $N(n_a)$ is the number of possible genotype combinations \mathbf{g} with n_a alleles [?]. The population mutation rate, θ , is and set to 0.001 for human [15].

With assumption that sequencing data of m samples are independent, probability of D given m genotypes $\mathbf{g} = (g_1, \dots, g_m)$, $p(D|\mathbf{g})$, is calculated as

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

The key to calculate $p(s = SNP|D)$ in Equation 1 is to compute the normalization factor, $\sum_{\mathbf{g}'} p(D|\mathbf{g}')p(\mathbf{g}')$.

We have

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

where

$$p(k) = \theta \left(\frac{1}{k} + \frac{1}{2m-k} \right) \frac{1}{N_k}$$

Obviously,

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

Thus, we estimate $Q_{m,k}$ in $O(m^2)$ using a dynamic programming algorithm.

Linkage Disequilibrium Analysis (LDA)

We assume haplotypes of m samples caused by a mutation on a coalescent tree during evolution. Thus, if we know the tree and the mutation, we could rebuild the haplotypes of samples. Figure 5 shows an example of the marginal ancestral tree with 4 samples, s_1 , s_2 , s_3 , and s_4 . If we know the marginal ancestral tree and we know a mutation from A to C at the branch, we could rebuild the haplotypes of 4 samples. Then we could compute the probability of D given the mutation. We estimate likelihood of data D given coalescent tree \mathbf{t} by summing likelihood of D given possible mutations on \mathbf{t} .

To find the coalescent trees, we collected known phased haplotypes of samples and use MARGARITA to estimate ancestral recombination graphs (ARGs). For example, we used HapMap3 phased genotypes to estimate ARGs for samples of Pilot 1 in 1000 Genomes Project. Since the MARGARITA cannot handle

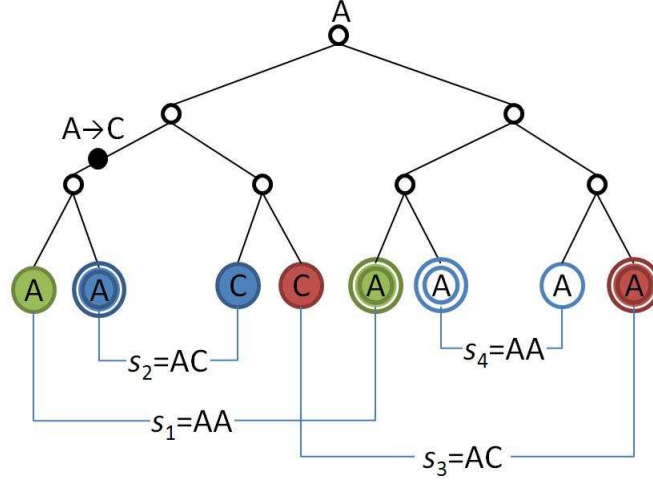


Figure 5. A toy sample for linkage disequilibrium analysis with 4 samples (8 haplotypes) and a coalescent tree. Given a mutation, we build genotypes for 4 samples and then compute the probability of D given the mutation. We estimate $p(D|\mathbf{t})$ given a tree by scanning all possible mutation in \mathbf{t} .

with large number of SNPs (markers) in term of running time and memory, we cut the whole genome into 1M bp segments. To make ARGs be consistent at sites of ends of 1Mb segments, we expanded 0.5 Mbp at each ends of a 1Mbp segment. We kept 20 ARGs for further analysis. In short, we used 2Mbp segments to estimate 20 ARGs for the central 1 Mbp segments. MARGARITA took a week for 400 samples on each segment.

We approximated coalescent trees of s by marginal coalescent trees T at flanking sites of s with the assumption that there is no recombination between the flanking sites. T contains 20 or 40 trees depending on whether there are one or two flanking sites around s . Denote Δ and $\bar{\Delta}$ be the cases where there is one and none mutation at s . We compute the probability of a mutation at s given D by Bayes' rule:

$$p(\Delta|D, T) = \frac{p(D|\Delta, T)p(\Delta|T)}{p(D|\Delta, T)p(\Delta|T) + p(D|\bar{\Delta}, T)p(\bar{\Delta}|T)} \quad (3)$$

where

$$p(\Delta|T) = \theta \sum_{i=1}^{2m-1} \frac{1}{i}; p(\bar{\Delta}|T) = 1 - p(\Delta|T)$$

We start solving Equation 3 by estimating probability of D given no mutation, $p(D|\bar{\Delta}, T)$. Let r be the correct reference at s , we approximate the prior probability of r

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

where ϵ is the error rate of NCBI 36. Empirical experiments with 1000 Genomes Project suggested that error rate ϵ would be higher than 10^{-5} and we set $\epsilon = 2 \times 10^{-5}$ in our experiments.

$$p(D|\bar{\Delta}, T) = \sum_r p(D|\bar{\Delta}, T, r)p(r)$$

Given reference allele r and non-mutation at s , all genotypes of m samples must be rr . Thus,

$$p(D|\bar{\Delta}, T, r) = \sum_i^m p(d_i|rr)$$

and,

$$p(D|\bar{\Delta}, T) = \sum_r p(r) \sum_i^m p(d_i|rr)$$

To estimate $p(D|\Delta, T)$, we scan all possible mutations on trees of T , and sum up probabilities of D given these mutations into $p(D|\Delta, T)$. Let start with reference r ,

$$\begin{aligned} p(D|\Delta, T) &= \sum_r p(D|\Delta, T, r)p(r) \\ &= \sum_r p(r) \sum_{\mathbf{t}_k \in T} p(D|\Delta, \mathbf{t}_k, r)p(\mathbf{t}_k) \end{aligned} \quad (5)$$

where we assume trees $\mathbf{t}_k \in T$ are independent and have the same prior probability $p(\mathbf{t}_k) = \frac{1}{|T|}$

To estimate $p(D|\Delta, \mathbf{t}_k, r)$ we scan all possible mutations in \mathbf{t}_k such that r must exist among $2m$ genotypes. We also consider a mutation outside \mathbf{t}_k that lead to reference r .

$$\begin{aligned} p(D|\Delta, T, r) &= \mu \sum_{a \neq r} p(a, r) p(D|\mathbf{g} = \mathbf{aa}) \\ &+ (1 - \mu) \sum_{e \in \mathbf{t}_k} \frac{1}{2} \sum_{a \neq r} [p(a, r) p(D|e_{ar}) + p(r, a) p(D|e_{ra})] \end{aligned} \quad (6)$$

where the prior probability of the unseen mutation is considered as

$$\mu = \frac{1}{2m+1} \left(\sum_{i=1}^{2m+1} \frac{1}{i} \right)^{-1}$$

The first part of Equation 6 presents for unseen mutation from the allele a of m samples to reference r at some branches outside \mathbf{t}_k . Prior probability of a unseen mutation is set as that of mutation at a leaf branch, μ . The prior probability of a mutation from a to r is set under the constraint: ration between transition ($A \leftrightarrow G$ and $C \leftrightarrow T$) and Transversion ($A, G \leftrightarrow C, T$) is about to 2. $p(D|\mathbf{g} = \mathbf{aa})$ is estimated as Equation 2.

The second part of Equation 6 presents for the seen-mutation between a and r at a branch $e \in \mathbf{t}_k$. The prior probability of a mutation happening at e is set as

$$p(e|\mathbf{t}_k) \propto \left[\frac{1}{n_a} + \frac{1}{2m - n_a} \right] \frac{1}{N(n_a)}$$

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

Call $\mathbf{g} = (g_1, \dots, g_m)$ genotypes of m samples caused by mutation $a \rightarrow r$ at edge e ,

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

Merge Equations 5 and 6,

$$\begin{aligned} p(D|\Delta, T) &= \mu \sum_r p(r) \sum_{a \neq r} p(a, r) p(D|\mathbf{h} = \mathbf{a}) \\ &+ \frac{1-\mu}{2} \sum_{r, \mathbf{t}_k \in T, a \neq r} p(r) p(\mathbf{t}_k) p(e|\mathbf{t}_k) [p(a, r) p(D|e_{a,r}) + p(r, a) p(D|e_{r,a})] \end{aligned} \quad (7)$$

It is obviously that complexity of computing Equation 7, $p(D|mut, T)$, is $O(N_A m^2 n_t)$ where $N_A = 4$, the number of nucleotides.

Genotyping

Let $\mathbf{g} = (g_1, g_2, \dots, g_m)$ be genotypes of m samples. Given a mutation at s , we estimate genotype g_i for sample i^{th} as follows:

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

where r is the reference and $p(r)$ is estimated from Equation 4. $p(g_i = ab|D, \Delta, T, r)$ is inferred by using Bayes' rule,

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

We have

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

Denote 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 $e \in E_{ik}^j$. $p(D, g_i = ab|r, \mathbf{t}_k, \Delta)$ is estimated under following cases:

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

$$\begin{aligned} p(D, g_i = aa|r \neq a, \mathbf{t}_k, \Delta) &= \mu p(a, r)p(D|\mathbf{g} = \mathbf{aa}) \\ &+ \sum_{e \in E_{ik}^0} p(e|\mathbf{t}_k)p(a, r)p(D|e_{a,r}) + \sum_{e \in E_{ik}^2} p(e|\mathbf{t}_k)p(r, a)p(D|e_{r,a}) \end{aligned}$$

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

Having obtained $p(g_i = ab|D, \mathbf{t}_k, \Delta)$, we set genotype g_i of sample i^{th}

$$g_i = \arg \max_{ab} \{p(g_i = ab|D, \mathbf{t}_k, \Delta)\}$$

Haplotype Phasing

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

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

where $p(r)$ is the prior probability of reference r (Equation 4). We infer $p(h_i = a|D, \Delta, T, r)$ using Bayes' rule.

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

where

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

Denote R_{ik} be the set of edges in which a mutation happen at $e \in R_{ik}$ results in h_i .

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

and

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

Having obtained $p(h_i = a|D, T, \Delta)$ for 4 alleles A, C, G , and T , h_i is considered as the allele with the maximum posterior probability,

$$h_i = \arg \max_a p(h_i = a|D, T, \Delta)$$

Issue with singleton and haplotype phasing

Singleton is a special case where a mutation happens at leaf branches. For each singleton genotype figuration, there are two possible mutations at leaf branches resulting in the figuration (Figure 6). That leads to the equal posterior probability for two alleles at singleton sites. Thus, the phased method cannot

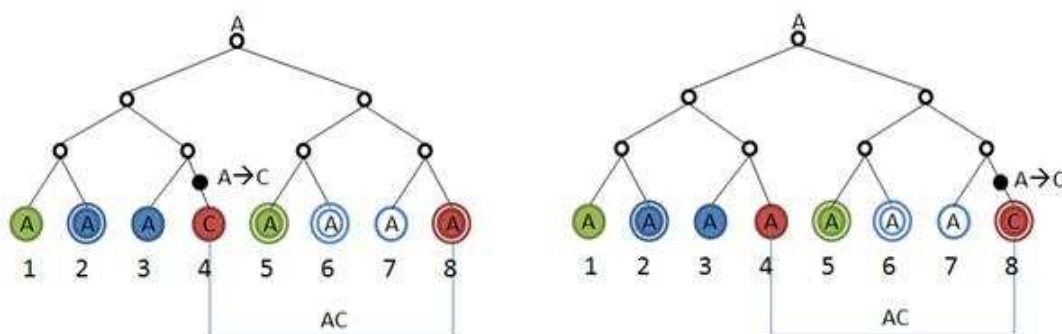


Figure 6. Two mutations at a singleton site lead to the sample genotyping configuration

be applied for singletons. In practice, we consider a sites as singleton when none or only one sample contains non-reference alleles. For these sites, we used genotyping results.

Acknowledgments

We thank James Stalker, Thomas Keane and David Craig for making the .bam files, Heng Li for his significant help with Maq, Gary Chen for MaCs, Eric Banks and Mark A. DePristo for GATK, the HapMap 3 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 and the Wellcome Trust.

References

1. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, et al. (2007) The diploid genome sequence of an individual human. PLoS Biol 5: e254.
2. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, et al. (2008) The complete genome of an individual by massively parallel dna sequencing. Nature 452: 872–876.

3. Wang J, Wang W, Li R, Li Y, Tian G, et al. (2008) The diploid genome sequence of an asian individual. *Nature* 456: 60–65.
4. Jong-Il Kim YS A highly annotated whole-genome sequence of a korean individual : Abstract : *Nature* .
5. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research* 18: 1851-1858.
6. Li R, Li Y, Fang X, Yang H, Wang J, et al. (2009) SNP detection for massively parallel whole-genome resequencing. *Genome Research* 19: 1124-1132.
7. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25: 2078-2079.
8. Minichiello JM, Durbin R Mapping trait loci by use of inferred ancestral recombination graphs. *Am J Hum Genet* 79: 910922.
9. 1000 Genomes Project Consortium HapMap3 (2009) Hapmap phase 3 draft 2 release. Technical report, <http://www.hapmap.org/>.
10. 1000 Genomes Project Consortium (2008) Meeting report: A workshop to plan a deep catalog of human genetic variation. Technical report, http://www.1000genomes.org/bcms/1000_genomes/Documents/1000Genomes-MeetingReport.pdf.
11. Chen GK, Marjoram P, Wall JD (2009) Fast and flexible simulation of DNA sequence data. *Genome Research* 19: 136-142.
12. Plagnol V, Wall JD (2006) Possible ancestral structure in human populations. *PLoS Genet* 2: e105.

13. The International HapMap Consortium HapMap2 A second generation human haplotype map of over 3.1 million snps. *Nature* 449: 851-861.
14. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25: 1754-1760.
15. Human Genome Sequencing Consortium international (2004) Finishing the euchromatic sequence of the human genome. *Nature* 431: 931–945.

Figure Legends

Tables