bioRxiv preprint doi: https://doi.org/10.1101/085936; this version posted November 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Large-scale Population Genotyping from Low-coverage Sequencing Data using a Reference Panel

Lin Huang[1], Petr Danecek[2], Sivan Bercovici[1], Serafim Batzoglou[1,+]
[1] Department of Computer Science, Stanford University [2] Wellcome Trust Sanger Institute
[+] To whom correspondence should be addressed. E-mail: serafim@cs.stanford.edu

Abstract

In recent years, several large-scale whole-genome sequencing projects were completed, including the 1000 Genomes Project, and the UK10K Cohorts Project. These projects aim to provide high-quality genotypes for a large number of whole genomes in a cost-efficient manner, by sequencing each genome at low coverage and subsequently identifying alleles jointly in the entire cohort. The resulting variant data are critical for the characterization of human genome variation within and across populations in the original projects, and for many downstream applications such as genome-wide association studies. The same datasets carry the potential to increase the quality of genotype calling in other low-coverage sequencing data in future sequencing projects, because the existing genotype calls capture the linkage disequilibrium structures of the cohorts they represent. In this paper we present *reference-based Reveel* (or *Ref-Reveel* in short), a novel method for large-scale population genotyping. Ref-Reveel is based on our earlier method, Reveel, which has been demonstrated to be an effective tool for variant calling from low-coverage sequencing data. Ref-Reveel leverages genotype calls from a sequenced cohort to boost the genotyping quality of new datasets, while maintaining high computational efficiency. We show that using a reference panel improves the quality of genotype calling via extensive experiments on simulated as well as real whole-genome data. Ref-Reveel is publicly and freely available at http://reveel.stanford.edu.

1 Introduction

Several large-scale whole-genome sequencing projects are completed or are actively underway, including the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2010), the UK10K Cohorts Project (http://www.uk10k.org), the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Project (CHARGE Consortium, 2009), CONVERGE Project (CONVERGE consortium, 2015), and the Haplotype Reference Consortium (http://www.haplotype-reference-consortium.org). These projects were designed to characterize human genetic variation in various populations to enable subsequent demography and association studies. For example, the 1000 Genomes Project (1000GP), which was launched in 2008, discovered >99% of single nucleotide polymorphism (SNP) variants with a frequency of >1% for multiple ancestries (The 1000 Genomes Project Consortium, 2015); later, the UK10K Cohorts Project explored rare and low-frequency variation in the UK population. To characterize human genetic variation, significant research efforts and massive resources have been expended in these projects to sequence a large number of whole genomes and to call genotypes of the sequenced genomes at polymorphic sites. Individual-level genomic data from these projects is available to the scientific community. These resources, beyond their value in the original projects and genome-wide association studies, can be used to enhance the quality of population genotyping in future genome sequencing projects. The genotypes identified capture the linkage disequilibrium (LD) structure of multiple populations. Cohorts that will be sequenced in future projects are likely to share a similar LD structure with one or more studied cohorts. Given the fact that this insight is being leveraged today in many haplotyping methods, the genotype calls from completed projects can be used as a reference panel in future genotype calling process.

While advancements in sequencing technology have enabled a sharp reduction in sequencing cost, this cost is still far from insignificant when thousands or even millions of individuals are sequenced. In this context, low-coverage whole genome sequencing of a large cohort is promising, because of its cost efficiency. Many recent sequencing projects, including the ones mentioned above, have employed the low-coverage/large-dataset strategy. For example, the 1000 Genomes Project sequenced 2,504 whole genomes

bioRxiv preprint doi: https://doi.org/10.1101/085936; this version posted November 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

at a mean depth of 7.4x; the UK10K Cohorts Project sequenced 3,781 whole genomes at a mean read depth of 7x. These experimental designs have been demonstrated to achieve reasonable genotyping accuracy. Reference-based genotype calling has the potential to further reduce the sequencing coverage needed for achieving a certain genotyping quality requirement. Thus, with the same sequencing cost, more individuals can be sequenced.

A number of population genotyping methods have been proposed to call genotypes from large-scale low-coverage whole genome sequencing data. The examples include glfMultiples+Thunder (Li et al., 2011), which employs a hidden Markov model that leverages LD information across a cohort to genotype likely polymorphic sites; SNPTools (Wang et al., 2013), which estimates genotyping likelihoods using a BAM-specific binomial mixture model and then utilizes a hidden Markov model (HMM) approach based on the statistical LD pattern model proposed in (Li and Stephens, 2003) to infer genotypes and haplotypes; and Beagle (Browning and Yu, 2009), which builds a HMM-based haplotype frequency model to capture LD pattern. These methods analyze all the sequenced samples in a cohort jointly, because calling genotypes from the data of a single low-coverage sequenced sample yields poor results. Despite the considerable success of these methods, using HMM-based models inevitably involves undesirable scalability and tendency to weaken long-distance LD, which is critical for calling the genotypes of rare variants (Huang et al., 2016). Both issues make these methods less suitable for large-scale population genotyping.

Reveel, our first generation of population genotyper, has been demonstrated to be an accurate and computationally efficient method for single nucleotide variant calling and genotyping from large-scale low-coverage sequencing data (Huang et al., 2016). Reveel infers genotypes using a summarization-maximization iterative method, which calculates the genotype probabilities using the current estimation of genotypes in the context of linkage disequilibrium in the summarization step and finds the genotypes maximizing the genotype probabilities in the maximization step. The underlying complex LD structure is leveraged by employing a simplified model, which scales linearly with the number of individuals in a cohort. Here, we present reference-based Reveel (or Ref-Reveel), a novel population genotyping method that effectively incorporates genotypes from completed projects into the Reveel framework to improve the genotyping quality of new datasets while maintaining low computational costs. Ref-Reveel infers the genotypes for a newly sequenced cohort of individuals from their genotype likelihoods, utilizing a cohort of reference individuals, for which the genotypes are known. We refer to the genomes of the newly sequenced individuals as *query genomes*, and refer to the genotypes of the previously sampled individuals as the *reference panel*. Ref-Reveel discovers likely variation sites from the query dataset, and calls genotypes at all of those sites regardless of their existence in the reference panel. We evaluate the performance of Ref-Reveel on both simulated data and real whole-genome data and demonstrate that using a reference panel substantially improves genotyping accuracy.

2 Algorithms

Our large-scale reference-based genotyping method is based on our earlier Reveel framework. Ref-Reveel infers the genotypes for a cohort of n sequenced query individuals given a background cohort of r reference individuals. In particular, we assume that prior to the analysis, the genotypes of the reference individuals are known. Ref-Reveel discovers m likely polymorphic sites from the query data using the SNP-discovery method outlined in our previous work (Huang et al. 2016) and sequentially calls the genotypes at those sites.

Our method infers the genotypes from a cohort sequenced at low-coverage using a two-step iterative method. In the first step, given the current estimation of genotypes **G**, our method calculates the genotype probabilities **P**. As the samples are sequenced at low coverage, the mapped reads have limited power to estimate the genotype probabilities **P** with a reasonable confidence. We therefore utilize the non-random association of alleles at different markers, that is, the linkage disequilibrium, to leverage evidence from additional sights to establish the probability of the genotypes. While the evidence provided by the mapped reads at marker i is limited, when allele A and marker i and allele B at marker j have strong association, observing allele B implies an increased chance of observing A. In the second step, the method infers the genotypes **G** by maximizing the current estimation of genotype probabilities; the predicted genotypes are

bioRxiv preprint doi: https://doi.org/10.1101/085936; this version posted November 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

then used to refine the genotype probabilities in the following iteration. We update **P** and **G** in a synchronous manner, that is, we first calculate the new values for **P** for every marker in every sample, and then overwrite all the old values with the new values before updating **G**. The two steps are iterated until the genotypes call converge. The pseudocode of the Ref-Reveel algorithm is shown in Algorithm 1.

Formally, we define the estimated genotype at a marker evaluated in a sample (denoted as target) as g_{target}. Let S be a set of markers that have strong LD with the evaluated marker. We define a vector $\boldsymbol{g}_{S,\text{target}}$ composed of the estimated genotypes at the markers in set S in the evaluated sample. In the first step of the ℓth iteration, we estimate the genotype probability $p^{(\ell)}_{\text{target},h}$ by calculating the probability that the target site exhibits a certain genotype h given $\boldsymbol{g}_{S,\text{target}}$ in the $(\ell-1)^{\text{th}}$ iteration and read alignments at the target site, that is $\Pr\left\{g^{(\ell-1)}_{\text{target}} = h \middle| \boldsymbol{g}^{(\ell-1)}_{S,\text{target}}, \text{alignments}\right\}$, where h can be $\{0,1,2\}$, representing homozygous reference, heterozygous, homozygous alternate, respectively. Using the Bayes' rule, we compute this conditional probability as follows, in which we do not need to calculate the denominator as it is identical to all the h's:

$$p^{(\ell)}_{\text{target},h} \propto \Pr\left\{\text{alignments} \middle| g^{(\ell-1)}_{\text{target}} = h, \boldsymbol{g}^{(\ell-1)}_{S,\text{target}}\right\} \cdot \Pr\left\{g^{(\ell-1)}_{\text{target}} = h \middle| \boldsymbol{g}^{(\ell-1)}_{S,\text{target}}\right\} \quad (1)$$

The first term can be simplified as $\Pr\left\{\text{alignments} \middle| g^{(\ell-1)}_{\text{target}} = h\right\}$ because of conditional independence. The probability $\Pr\{\text{alignments} | g_{\text{target}} = h\}$ is essentially the genotype likelihoods at the target site, which is pre-calculated from sequencing read alignments. The conditional probability in the second term is identical across all the samples in the cohort. This is important because we can divide the computation of genotype probability of all the samples at a certain marker into two $O(n)$ computations. First, we calculate $\Pr\left\{g^{(\ell-1)} = h \middle| \boldsymbol{g}^{(\ell-1)}_{S}\right\}$ for $h = \{0,1,2\}$ by accessing every sample once. Second, we calculate $p^{(\ell)}_{\text{target},h}$ by simply retrieving the $\Pr\left\{g^{(\ell-1)} = h \middle| \boldsymbol{g}^{(\ell-1)}_{S} = \boldsymbol{g}^{(\ell-1)}_{S,\text{target}}\right\}$ value and the genotype likelihood value. We simplify Equation (1) as

$$p^{(\ell)}_{\text{target},h} \propto \Pr\{\text{alignments} | g_{\text{target}} = h\} \cdot \Pr\left\{g^{(\ell-1)} = h \middle| \boldsymbol{g}^{(\ell-1)}_{S}\right\} \quad (2)$$

The genotype probabilities for all the h's are then used to update the genotype in the second step:

$$g^{(\ell)}_{\text{target}} \leftarrow \arg\max_{h} p^{(\ell)}_{\text{target},h} \quad (3)$$

Our experiments on simulated data show that, the vast majority of loci converge within ten iterations; applying additional iterations yields little improvement in genotyping accuracy.

2.1 Feature selection

Ideally, the feature vector $\boldsymbol{g}_S$ classifies the genotype at the evaluated marker g perfectly, meaning that for every specific genotype vector $\boldsymbol{g}_S$, all the samples with this vector have the same g at the evaluated marker. We introduce a minimum entropy feature selection technique. The heuristic of constructing the feature vector is as follows. We first estimate the LD between the evaluated marker and all other markers using a metric sim described later. All the markers are sorted based on their sim values from strong LD to weak LD. Then, we select the most informative sites, that is, the first k markers in the ranked list (by default,

Input: genotype likelihoods of n query genomes at m markers
genotypes of r reference genomes
Do for $T = 1,2,\cdots$
Evaluate pairwise LD
Do for each LD estimation metric
For each marker, (re-)pick a feature vector
Initialize genotypes of query genomes $\mathbf{G}^{(0)}$
Do for $R = 1,2,\cdots$
Calculate genotype probability $\mathbf{P}^{(\ell)}$ for each marker in each query genome
Find genotypes $\mathbf{G}^{(\ell)}$ that maximize $\mathbf{P}^{(\ell)}$
For each marker, combine the resulting genotypes with multiple LD metrics using the final hypothesis obtained from an AdaBoost classifier
Output genotypes of n query genomes at m markers

Algorithm 1. *Overview of Ref-Reveel.*

bioRxiv preprint doi: https://doi.org/10.1101/085936; this version posted November 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

k=3-5, depending on $n+r$). Although each of these markers is one of the most informative features independently, the set may not be the most informative feature vector; when the chosen sites strongly correlate with each other, the information gained from selecting additional sites after the first one is limited. Hence, we assess the entropy of the initial feature vector using the reference panel. When the entropy is high, we select an additional marker from a ranked list to replace a marker in the feature vector as long as such a replacement reduces the entropy. The procedure is repeated until the entropy is sufficiently small. The entropy used in feature selection is a weighted average over all branches resulting from the split based on $\boldsymbol{g}_S$. Let $\#\{A\}$ be the number of reference samples exhibiting event A. The entropy given the current feature selection can be written as

$$E = -\sum_{\boldsymbol{g}_S} \frac{\#\{\boldsymbol{g}_S\}}{r} \sum_{h=0}^{2} \frac{\#\{g = h \wedge \boldsymbol{g}_S\}}{\#\{\boldsymbol{g}_S\}} \log \frac{\#\{g = h \wedge \boldsymbol{g}_S\}}{\#\{\boldsymbol{g}_S\}} \tag{4}$$

The feature selection process focuses on the most informative markers based on their LD, regardless of their genetic distance from the evaluated marker. The same feature vector provides additional value by guiding towards a good initial guess of genotypes. Let c and d be the number of reads supporting reference and alternate alleles at the evaluated marker. Since the summation of c and d is less likely adequate for genotype calling, we borrow the read counts from the sites in set S. Similarly, let $\boldsymbol{c}_S$ and $\boldsymbol{d}_S$ be the read counts at all the sites in set S. We use $C = c + \sum \boldsymbol{c}_S$ and $D = d + \sum \boldsymbol{d}_S$ as our initial estimate.

2.2 Estimation of linkage disequilibrium

We conducted multiple rounds of the iterative method described above. Each round starts with feature selection, using the output genotypes of the previous round as initial genotypes. The previous round genotype calls are also used for LD estimation.

The LD is best estimated using the phased haplotypes of a cohort. In the initial round, however, both the linkage phase as well as the genotypes are missing. We need to effectively estimate LD based on the genotype likelihoods at every marker in each sample. Three metrics were proposed in (Huang et al., 2016) using the number of read supporting reference and alternate alleles. The read counts are not always available in practice unless the raw read alignments are accessible. Many large-scale projects provide genotype likelihoods instead. To reuse the metrics introduced in our previous work, we calculate the *effective* reference and alternate counts from the genotype likelihoods. We precompute a table of $\Pr\{c, d|g\}$ for every possible combination of reference and alternate counts c, d given an estimated error rate using the revised MAQ model (Li, 2010). For every target site, we find a combination of $\hat{c}, \hat{d}$ that minimizes the Manhattan distance between $\Pr\{\hat{c}, \hat{d}|g\}$ and the input genotype likelihoods at the target site over all the g's.

With $\hat{c}, \hat{d}$, we move on to calculate the metrics. Let $X_{i,t}$ be the event that at least one read at marker i of sample t supporting alternate allele, that is, $\hat{d}_{i,t} \geq 1$. We categorize samples using the following relation predicates: $X_{i,t} \wedge X_{j,t}$, $\neg X_{i,t} \wedge X_{j,t}$, $X_{i,t} \wedge \neg X_{j,t}$, and $\neg X_{i,t} \wedge \neg X_{j,t}$. The samples with $\hat{d}_{i,t} \geq 1$ and $\hat{d}_{j,t} \geq 1$, that is $X_{i,t} \wedge X_{j,t}$, are the evidence of LD; we therefore use the number of such samples as the numerator of our metric. A naïve denominator is the total number of samples. This number, however, can easily be dominated by the number of samples with $\hat{d}_{i,t} = 0$ and $\hat{d}_{j,t} = 0$, that is $\neg X_{i,t} \wedge \neg X_{j,t}$, especially when markers i and j are rare variants. We therefore subtract this dominant number from the denominator. The metric can be written as follows, where the sign function $\mathrm{sgn}(x) = 0$ if $x = 0$, and $\mathrm{sgn}(x) = 1$ if $x > 0$.

$$sim_1(i,j) = \frac{\sum_t \min\{\mathrm{sgn}(\hat{d}_{i,t}), \mathrm{sgn}(\hat{d}_{j,t})\}}{\sum_t \max\{\mathrm{sgn}(\hat{d}_{i,t}), \mathrm{sgn}(\hat{d}_{j,t})\}} \tag{5}$$

Because sequencing reads exhibit sequencing errors, a very small $\hat{d}_{i,t}$ value does not offer strong evidence of the existence of alternate alleles in the genotype. On the other hand, observing multiple reads supporting alternate alleles at both markers i and j in a sample increases the probability that alleles at those two markers are non-randomly associated. The insight drives our second metric:

$$sim_2(i,j) = \frac{\sum_t \min\{\hat{d}_{i,t}, \hat{d}_{j,t}\}}{\sum_t \max\{\hat{d}_{i,t}, \hat{d}_{j,t}\}} \tag{6}$$

bioRxiv preprint doi: https://doi.org/10.1101/085936; this version posted November 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The third metric is defined as follows. In comparison to the earlier proposed metrics, the last metric generates a score that increases rapidly as both samples exhibit more reads that support alternate alleles.

$$sim_3(i,j) = \frac{\sum_t \min\{(\hat{d}_{i,t})^2, (\hat{d}_{j,t})^2\}}{\sum_t \max\{(\hat{d}_{i,t})^2, (\hat{d}_{j,t})^2\}} \tag{7}$$

In the subsequent rounds, the phasing information is hidden. We use a composite LD estimator Δ as proposed previously (Schaid, 2004), where A and B represent the major alleles of two markers, and a and b represent the minor alleles:

$$\Delta = 2p_{aabb} + p_{aaBb} + p_{Aabb} + {}^1\!/_2\, p_{AaBb} - 2p_a p_b \tag{8}$$

The number of rounds can be specified by users. In our experiments, we apply three rounds because we have found that the p's in Equation (8) stabilize within three rounds.

2.3 Bias reduction

When the initial genotype estimates are biased towards the homozygous reference, the iterative method using Equations (2) and (3) generates biased genotypes as well. Given the biased initial estimate, $\Pr\left\{g^{(\ell-1)} \neq 0 \middle| \boldsymbol{g}_S^{(\ell-1)}\right\}$ tends to be a very small number. Such bias can hardly be adjusted by the iterative method. To counter, we leverage the alternate allele frequency (AF) of the evaluated marker across the cohort and rewrite Equation (2) as:

$$p_{\text{target},h}^{(\ell)} \propto \Pr\{\text{alignments} | g_{\text{target}} = h\} \cdot \Pr\left\{g^{(\ell-1)} = h \middle| \boldsymbol{g}_S^{(\ell-1)}, \text{AF}\right\} \tag{9}$$

The conditional probability $\Pr\left\{g^{(\ell-1)} = h \middle| \boldsymbol{g}_S^{(\ell-1)}, \text{AF}\right\}$ is calculated using a noisy-MAX gate (Zagorecki and Druzdzel, 2013), which favors the largest h value, that is, homozygous alternate. As a result, this estimation is biased towards homozygous alternate. The probability $\Pr\{g^{(\ell-1)} = h | \text{AF}\}$ for a certain marker can be calculated in $O(1)$ time.

$$\Pr\left\{g^{(\ell-1)} = h \middle| \boldsymbol{g}_S^{(\ell-1)}, \text{AF}\right\} = \sum_{\forall u,v:\max\{u,v\}=h} \Pr\left\{g^{(\ell-1)} = u \middle| \boldsymbol{g}_S^{(\ell-1)}\right\} \cdot \Pr\{g^{(\ell-1)} = v | \text{AF}\} \tag{10}$$

We use Equations (2) and (9) alternatingly in the iterations of our algorithm for calculating $p_{\text{target},h}^{(\ell)}$.

2.4 Boosting the performance

The iterative algorithm is applied based on three metrics sim_i separately, resulting in three genotype matrices. In this process, we apply either Equation (9) or Equation (2) in each iteration. In the last iteration, however, we apply both Equations (2) and (9) to achieve three additional genotype matrices. In total we use six applications of our iterative algorithm to achieve six genotype matrices $\mathbf{G}_i$ and $\mathbf{G}_i{}'$, where $i = 1,2,3$. Each of these matrices independently provides decent genotype calls.

We further boost the genotyping quality by training AdaBoost classifiers (Freund and Schapire, 1995). The six applications of our iterative algorithm are used as weak classifiers, denoted as H's. For every marker we train an AdaBoost classifier with r labelled training examples $(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_r, y_r)$, where $\boldsymbol{x}_i$ is all the known information regarding the ith reference genome, and the class label y_i is the genotype of the reference genome at the marker. We apply our iterative algorithm to the reference genomes assuming uniform genotype likelihoods to achieve the output genotype $H_j(\boldsymbol{x}_i)$, in which we use the feature vectors selected for genotyping the query genomes, and calculate $\Pr\left\{g^{(\ell-1)} = h \middle| \boldsymbol{g}_S^{(\ell-1)}\right\}$ in Equations (2) and (9) using the genotypes of the reference genomes. The pseudocode of training AdaBoost classifiers is shown in Appendix 1. The final hypotheses resulted from AdaBoost classifiers are used to combine $\mathbf{G}_i$ and $\mathbf{G}_i{}'$ of the query genomes.

2.5 Filters for quality control

To limit the false positives (defined in Table 1) resulting from the genotype calling process, we measure two ratios at every marker. Let θ_X be the probability that allele X exists at a marker in a sample. The function $f(\theta_X) = a\theta_X/(1 + a - \theta_X)$ with $a = 5 \times 10^{-6}$ introduced in (Huang et al, 2016) is a monotonically increasing function with $f(0) = 0$ and $f(1) = 1$; its derivative is also monotonically increasing. This function has an important property that $f(\theta_X)$ approaches 1 if and only if θ_X is very close

bioRxiv preprint doi: https://doi.org/10.1101/085936; this version posted November 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 1. *Metrics used for genotyping performance assessment.*

metric	expression	metric	expression
accuracy	$(n_{\text{hom,hom}} + n_{\text{het,het}} + n_{\text{homnon,homnon}})/n_{\bullet,\bullet}$	TP	$n_{\text{het,het}} + 2n_{\text{homnon,homnon}} + n_{\text{homnon,het}}$
het-accuracy	$(n_{\text{het,het}} + n_{\text{homnon,homnon}})/(n_{\text{het},\bullet} + n_{\text{homnon},\bullet})$	FP	$n_{\text{hom,het}} + 2n_{\text{hom,homnon}} + n_{\text{het,homnon}}$
sensitivity	$\text{TP}/(\text{TP} + \text{FN})$	FN	$n_{\text{het,hom}} + 2n_{\text{homnon,hom}} + n_{\text{homnon,het}}$
precision	$\text{TP}/(\text{TP} + \text{FP})$		

$n_{x,y}$: the number of sites where the truth genotype is x and the outcome is y. $\bullet$: wildcard, representing any genotypes.

to 1. Let score_X be the summation of $f(\theta_X)$ over all the samples at the marker. We measure the ratio of score_X between major and minor alleles, and the ratio of allele frequencies between major and minor alleles. Because of the property of $f(\theta_X)$, score_X can be used to approximate the number of samples in which we observe strong evidence for the existence of allele X. Thus, the two ratios described above should not be significant different. If at a marker these two ratios are highly inconsistent, that marker is labelled as invariant site in the outputs.

The false positive rate can be further reduced by applying a few more hard filters. The following markers are also labelled as invariants: (1) the markers that fail the Hardy-Weinberg equilibrium, using Pearson's chi-squared test; (2) the sites that are called as common variants by our algorithm but not reported in the integrated call set of the 1000GP Phase 3; (3) the sites at which SAMtools calls multiple complex variants.

3 Simulations

We conducted extensive simulations to demonstrate the effectiveness of Ref-Reveel on query datasets with a sample size ranging from 100 to 1000, and a reference panel with 7500 diploid samples.

3.1 Experimental setup

We simulated variants in 25,000 European haplotypes using *COSI* (Schaffner 2005) with parameters from the best-fitting model for a 1-cM region. Using the human genome build GRCh37 chr20:43,000,000-44,000,000 as the reference genome, we generated 25,000 simulated haplotypes, which were randomly partitioned into two sets. The first set contained 15,000 haplotypes, forming the reference panel. We assumed the genotypes of this panel were all known. The remaining 10,000 haplotypes were used to create simulated query samples. We paired every two subsequent haplotypes in this set to generate 5,000 query samples. To generate the reads, for each query sample, we randomly selected a sample from the 1000GP Phase 3 dataset and extracted the read lengths and positions. A set of simulated reads were then created for the query sample using the same positions and lengths. Finally, sequencing errors were simulated and injected into the generated reads. Once all reads were generated, a subsequent alignment was applied to the reads to map them back to the reference genome using BWA-MEM. The genotype likelihoods of these query samples were computed using SAMtools.

3.2 Results

We compared the performance of our earlier Reveel method and Ref-Reveel using three query datasets, which included 100, 500, and 1000 query samples. The performance was evaluated at the discovered polymorphic sites; the sites were identical for both Reveel and Ref-Reveel. A few measurements were used for performance assessment, as summarized in Table 1. As shown in Figure 1(a)-(c), using the reference panel resulted in significant improvements in genotyping accuracy across the entire allele frequency spectrum. We attribute the improved performance to the fact that the use of the reference panel facilitates an improved feature selection regardless of allele frequencies. The genotyping accuracy of the dataset with the 100 query samples improved the most. The genotyping accuracy at all the variants discovered by Reveel/Ref-Reveel increased from 99.5143% to 99.7775%, corresponding to a reduction in the number of wrongly inferred variants from 2,172 to 995 across the cohort. Indeed, leveraging prior information when inferring the genotypes of a small cohort enables more adequate approximation of the LD model. On the other hand, the dataset with 1,000 query samples already contained sufficient informative markers for LD estimation. While representing a more modest increase in accuracy, the use of the additional reference panel increased genotyping accuracy from 99.8845% to 99.9222%. Appendix 2 illustrates that use of a reference

bioRxiv preprint doi: https://doi.org/10.1101/085936; this version posted November 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

panel simultaneously improved both sensitivity and precision for every allele frequency category, meaning that incorporating reference panels reduced both false positives and false negatives.

We also measured the genotyping accuracy at the sites where the truth genotypes were heterozygous and homozygous alternate, that is het-accuracy in Table 1. Again, we focused on the polymorphic sites discovered by Reveel/Ref-Reveel (see Figure 1(d)-(f)). Similar to the results described above, we observed significant improvements across the entire allele frequency spectrum. The mean het-accuracy across all the variants increased by 0.70%, 0.39%, 0.19% for n = 100, 500, 1000, respectively; more notably, the mean het-accuracy at rare variants increased by 1.32%, 1.17%, 0.77% for the three datasets.

When we applied Ref-Reveel, parsing the reference VCF file and incorporating the reference haplotypes into the genotyping process introduced additional computation overhead. As a result, the running time of Ref-Reveel was 4.44, 2.78, and 2.41 times that of Reveel for the datasets with 100, 500, and 1000 query samples, respectively. The ratio of the running time of Ref-Reveel in comparison to Reveel scaled well with the increase of sample size.

4 Improved Genotyping Performance using Reference Panels

4.1 Genotype Calling UK10K Samples using Reveel

As the first step of our whole genome experiment, we created a reference panel by applying Reveel[1] to the genotype likelihoods of 3,910 UK10K samples from two British cohorts, the Avon Longitudinal Study of Parents and Children and TwinsUK, to infer the genotypes of these samples. This input dataset included 61,897,468 unfiltered sites across the whole genome. The input genotype likelihoods were calculated from low-coverage sequencing data (average read depth 7x) using SAMtools and BCFtools by the UK10K Consortium.

We performed this experiment in parallel on Sanger Institute's computing farm. To do so, we first partitioned the whole genome into 4,720 non-overlapped genomic segments. The chromosomes were chunked by marker numbers; the maximum number of markers per segment was set to 12,000. Then, Reveel was applied to each segment separately. A total of 61,167,575 sites were genotyped by Reveel across the entire genome; the other 729,893 sites were identified by Reveel as invariant reference alleles across the studied cohort. Reveel produced genotype likelihoods at common and low-frequency variation sites and genotypes at rare and invariant sites[2]. The genotype likelihoods at common and low-frequency variation

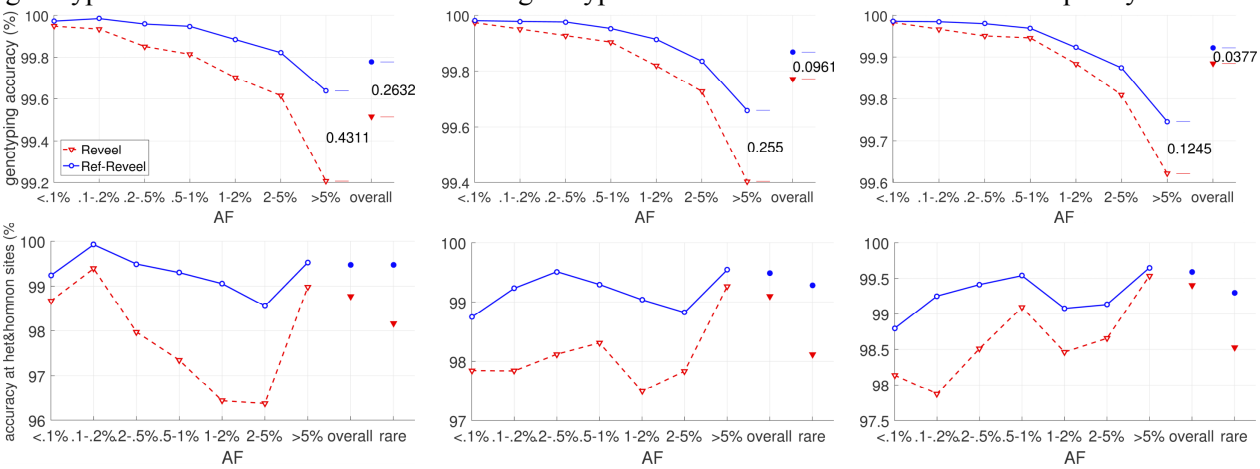

Fig. 1. *Comparison of the performance of Reveel and Ref-Reveel using simulated datasets*. The figures in rows 1 and 2 show the genotyping accuracy at all the variants discovered by Reveel/Ref-Reveel and the genotyping accuracy at the heterozygous and homozygous alternate sites respectively. The figures in column 1-3 show the performance for the n = 100, 500, 1000 cases respectively. The variants were grouped according to their AFs. Besides reporting the overall performance and the performance in each AF group, we also reported the het-accuracy at rare variants.

[1] In terms of functionality, Reveel and Ref-Reveel without reference are equivalent.

[2] Throughout this paper, we define common, low-frequency, and rare variants as variants with minor allele frequency (MAF) ≥5%, 1-5%, and <1%, respectively.

bioRxiv preprint doi: https://doi.org/10.1101/085936; this version posted November 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

sites, which were roughly 13% of all the genotyped sites, were fed into Beagle (version 3.3.2) for the final refinement. The refined genotypes were merged with the genotypes at rare and invariant sites. The total running time for the combined pipeline, when applied on the above dataset, was 85,211 CPU hours, out of which Reveel consumed 49,856 CPU hours (59% of the total running time); the memory usage was 18.6 GB.

We compared the panel (labelled as *R*) with the haplotypes of 3,781 UK10K samples reported by the UK10K Consortium (labelled as *10K*). The *10K* panel was created as follows by the UK10K Consortium (The UK10K Consortium, 2015). The SNP calls in the *10K* panel were made using SAMtools and BCFtools, then recalled using the GATK (version 1.3-21) UnifiedGenotyper, and filtered using the GATK VariantRecalibrator followed by GATK ApplyRecalibration. The missing and low confidence genotypes were then refined using Beagle 4 (rev909). The UK10K Consortium reported 42,001,233 polymorphic sites that passed the filter; the haplotypes at these sites were included in the panel.

We measured the quality of these two reference panels on 66 TwinsUK samples that were both low-coverage whole-genome sequenced and high read-depth exome sequenced. The genotype calls from high read-depth exome sequencing reported by TwinsUK were used as benchmarks, including 160,119 sites. We evaluated genotyping accuracy, sensitivity and precision as defined in Table 1 at the polymorphic sites discovered from the exome sequencing data.

Figure 2 clearly shows that panel *R* achieved higher quality in comparison to the *10K* panel. A closer examination shows that panel *R* had moderately lower precision and significantly higher sensitivity than panel *10K*. The result was expected, because when Reveel was used to generate a reference panel, its SNP discovery parameter was set to a very low value to infer as many likely variation sites as possible. Despite its higher false positive rate, we used this parameter setting as our experiments show that a panel with high SNP discovery rate benefited the downstream reference-based genotype calling (see Section 4.2).

To compare the Reveel-called panel and panel *10K* at the same precision level, we applied four hard filters described in Section 2.5 to reduce the false positive rate of panel *R*. Across the whole genome 2,310,217 sites (3.78%) were removed by these filters. As shown in Figure 2(c), the sensitivity of the filtered Reveel-called panel was higher than that of panel *10K*.

The superiority of panel *R* over panel *10K* was observed across the whole AF spectrum, but the advantage at the sites with MAF<0.1% and MAF≥1% was particularly significant (see Figure 2(d)). For rare variants, Reveel had higher genotyping accuracy because of its unique capability of identifying the most informative sites in a way that is less sensitive to genetic distance. Other methods, by contrast, implicitly weaken the association between remote sites when they build their models. For high AF variants, we attributed the high genotyping accuracy of Reveel to its another important feature, that is, Reveel provided high-quality genotype probabilities at high AF sites. Applying the hard filters to panel *R* cancelled the advantage at the very rare sites, implying that a good amount of very rare variants was removed by the hard filters. This could weaken the power of panel *R* as a reference panel. Applying the hard filters also reduced the accuracy at common variants. Even though, the filtered *R* panel still had higher quality than panel *10K* at those sites.

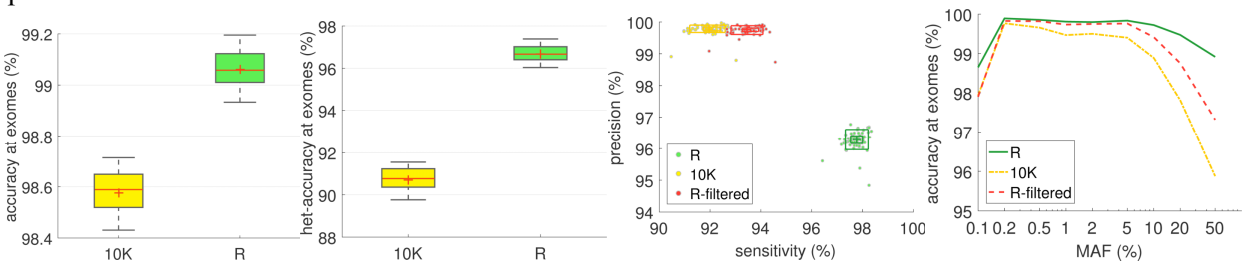

Fig. 2. *Quality of two UK10K reference panels.* We showed the aggregated (A) accuracy and (B) het-accuracy of 66 TwinsUK samples using boxplots, in which the central marks are the mean, the red lines are the median, the edges of the box are the 25th and 75th percentiles, and the whiskers span 9th to 91st percentiles. We also showed the sensitivity and precision using rangefinder boxplots (C). The dots are the performance of individuals; the inner and outer boxes span the 25th-75th and the 9th-91th percentiles respectively. The average accuracy across the whole genome as a function of minor allele frequency is shown in (D).

bioRxiv preprint doi: https://doi.org/10.1101/085936; this version posted November 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 2. *Genotyping accuracy of Ref-Reveel on the whole genome*. We evaluated the genotyping accuracy of four settings of Ref-Reveel at the sites where the CG data reported heterozygous and homozygous alternate and Ref-Reveel discovered SNPs.

method	overall (%)	S_{common} (%)	S_{loF} (%)	S_{rare} (%)	$S_{\text{invariant}}$ (%)
R-baseline	98.7267	99.1241	98.4575	97.4269	80.5800
R-10K	98.8818	99.2576	98.8222	97.9519	79.7389
R-R	99.0280	99.3671	99.0418	98.1839	81.4276
R-1000GP	98.9069	99.3089	98.7196	97.7574	79.5758

4.2 Applying Ref-Reveel to CEU samples using reference panels

Experimental setup

We demonstrated the improved performance of Ref-Reveel by applying our method on real low-coverage data, contrasting our calls against an orthogonal validation set. In particular, we applied Ref-Reveel on 99 CEU samples from the 1000GP Phase 3. The BAM files corresponding to the low-coverage sequencing data were retrieved from the 1000GP website (retrieved on 01/13/2016). Sequentially, we applied SAMtools (version 1.2)' mpileup command to generate the genotype likelihoods. In our previous work, we established Reveel outperforms state-of-the-art methods in SNP discovery and genotype calling. For the analysis here, we will use the application of reference-free Reveel as the baseline for our comparison, denoted as *R-baseline*.

We assessed the performance of Ref-Reveel using three reference panels. The first two reference panels were panel *10K* and panel *R,* as described above. The third reference panel, denoted as *1000GP*, contained the haplotypes of 2,405 non-CEU 1000GP samples from the integrated call set reported by the 1000GP Phase 3 (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/). The resulting genotype calls using these three reference panels were labelled as *R-10K*, *R-R*, and *R-1000GP*, respectively.

We compared Ref-Reveel against three state-of-the-art pipelines. In the first pipeline, we combined SNPTools (Wang et al., 2013) and Beagle (Browning and Browning, 2009), denoted as *S+B*. To estimate the genotype likelihoods at polymorphic sites, we used SNPTools (v1.0)' bamodel→poprob commands. We then applied Beagle 4 (r1399) to infer the final set of genotypes. For the second pipeline, we combined GATK (DePristo et al., 2011; McKenna et al., 2010) with Beagle, denoted as *G+B*. Namely, Beagle 4 (r1399) used the genotype likelihoods generated by GATK UnifiedGenotyper (v3.3) to infer the final set of genotypes. Finally, our third pipeline was glfMultiples+Thunder (Li et al., 2011), denoted as *g+T*. While Thunder is computationally intensive, it has demonstrated improved genotyping accuracy when applied to the output of glfMultiples. All methods were applied using default parameters, unless otherwise specified.

Performance

To assess performance, we contrasted the generated calls against the genotypes reported in the Complete Genomics (CG) dataset (retrieved on 01/13/2016). A total of 63 samples, out of the initial 99 samples described above, were reported in the CG dataset. We measured genotyping accuracy at all the sites where the CG data reported heterozygous and homozygous alternate, that is, het-accuracy defined in Table 1.

The performance of *R-baseline, R-10K*, *R-R*, and *R-1000GP* as evaluated across the entire human genome is outlined in Table 2. We focused on the sites that were both reported by the CG data and genotyped by Ref-Reveel. Besides measuring the overall performance, we further divided the sites into four categories: common variants, denoted as S_{common}; low frequency variants, denoted as S_{loF}; rare variants, denoted as S_{rare}; and finally invariant sites, denoted as $S_{\text{invariant}}$, for sites that do not appear in the integrated call set of the 1000GP Phase 3. The performance was evaluated for each variant category separately.

Given the above dataset, Ref-Reveel discovered 25,261,018 likely polymorphic sites, and reported invariant alleles at remaining sites. Compared to the baseline, using the UK10K samples as references improved the genotyping accuracy. The observed improvement in genotyping accuracy and sensitivity was not surprising since the British cohort and the CEU samples exhibited low genetic divergence (Eyheramendy et al., 2015; The 1000 Genomes Project Consortium, 2015). As such, we expected that indeed the UK10K reference panel would provide a proper approximation of the LD structure in the CEU

bioRxiv preprint doi: https://doi.org/10.1101/085936; this version posted November 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

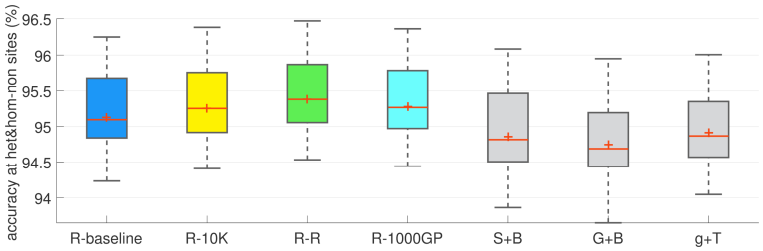

Fig. 3. *Genotyping performance for different methods and reference panels on chromosome 20*. For each method or setting, the genotyping accuracy at heterozygous and homozygous alternate sites of 63 samples that were studied by both 1000GP Phase 3 and CG was represented by a boxplot.

population. Notably, *R-R* exhibited a higher genotyping accuracy in comparison to the *R-10K*, although both reference panels originated from the UK10K dataset. We attributed the observed result to the improved genotyping quality of panel *R*. Using a reference panel improved the performance across the entire allele frequency spectrum compared to *R-baseline*. At the common variants, using panel *R* reduced the genotyping error rate from 0.88% to 0.63%, corresponding to the genotypes of 7,847 markers per sample corrected; at the low-frequency variants, we observed the reduction from 1.54% to 0.96%, corresponding to the genotypes at 1,089 markers per sample corrected; at the rare variants, the error rate was reduced from 2.57% to 1.82%, corresponding to the genotypes at 878 markers per sample corrected. It is important to note that while the *1000GP* reference panel originated from a more heterogeneous set of reference populations, when that panel was utilized for genotype calling, compared to the baseline we observed a 21.1%, 17.0%, 12.8% reduction in the genotyping error rate at common, low-frequency, and rare variants, respectively. This could imply that the LD structure of the variants with sufficiently high AF were likely to be captured by this reference panel, and that with the AF decrease the LD structure became less likely to be captured.

We contrasted the performance of Ref-Reveel against state-of-the-art methods over chromosome 20, which was roughly 2% of the whole genome; assessing the performance of the alternative methods across the entire genome was computational prohibitive. The performance was evaluated at all the sites where the CG data reported either heterozygous or homozygous alternate on this chromosome. As shown in Figure 3, Ref-Reveel, with or without a reference panel, outperformed previous state-of-the-art methods. The SNPTools+Beagle and GATK+Beagle pipelines performed worse than any setting of Ref-Reveel in terms of het-accuracy. The third pipeline, glfMultiples+Thunder, achieved higher accuracy in comparison to the two previous ones, yet exhibited a lower performance in comparison to the *R-baseline* results.

Running time

As shown in Table 3, *R-baseline* had the lowest CPU running time of 7.36 CPU days when evaluating calls across the entire genome, and a total of 134 minutes when evaluating calls on chromosome 20 alone. Incorporating the reference panels roughly doubled the running time. When panel *R* was used, Ref-Reveel's computational time was 12.79 CPU days on the whole genome, and 353 minutes on chromosome 20. This computational overhead is fairly practical, even when a single core is used. Given the fact that Ref-Reveel is parallelizable in a straightforward manner (by analyzing independent genomic regions in parallel). A cluster can be used to reduce the end-to-end running time to less than a day.

Among the state-of-the-art methods, SNPTools+Beagle was the only one that had efficiency comparable with that of Ref-Reveel. According to our experiment on chromosome 20, GATK+Beagle was 4.2 times slower than *R-R*, whereas glfMultiples+Thunder was 62.1 times slower. Extrapolating from these results, one can estimate that the running time of these two pipelines on the entire genome will be approximately 53 CPU days and 794 CPU days respectively. We conclude that Ref-Reveel provides substantial accuracy and efficiency improvements in population genotyping, and enables the accurate and efficient genotyping of a sequenced cohort using a previously genotyped reference panel cohort.

Table 3. *Running time*. The running time was measured on a 2.40GHz Intel Xeon processor.

Method	R-baseline	R-10K	R-R	R-1000GP	S+B	G+B	g+T
chromosome 20 (mins)	134	323	353	275	407	1464	21924
whole genome (CPU days)	7.36	13.02	12.79	11.55	-	-	-

bioRxiv preprint doi: https://doi.org/10.1101/085936; this version posted November 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Acknowledgements

TwinsUK is funded by the Wellcome Trust, Medical Research Council, European Union, the National Institute for Health Research (NIHR)-funded BioResource, Clinical Research Facility and Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust in partnership with King's College London.

References

Arthur R et al. (2016) Rapid genotype refinement for whole-genome sequencing data using multi-variate normal distributions. Bioinformatics, 32(15):2306–2312.

Browning BL and Browning SR. (2009) A unified approach to genotype imputation and haplotype phase inference for large data sets of trios and unrelated individuals. The American Journal of Human Genetics, 84:210–223.

Browning BL and Yu Z. (2009) Simultaneous genotype calling and haplotype phase inference improves genotype accuracy and reduces false positive associations for genome-wide association studies. The American Journal of Human Genetics, 85:847–861.

CONVERGE consortium. (2015) Sparse whole-genome sequencing identifies two loci for major depressive disorder. Nature, 523:588–591.

CHARGE Consortium. (2009) Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium: Design of prospective meta-analyses of genome-wide association studies from five cohorts. Circ Cardiovasc Genet., 2:73-80.

DePristo MA et al. (2011) A framework for variation discovery and genotyping using next-generation dna sequencing data. Nature Genetics, 43:491–498.

Eyheramendy S et al. (2015) Genetic structure characterization of Chileans reflects historical immigration patterns. Nature Communications, DOI: 10.1038/ncomms7472.

Freund Y and Schapire RE. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55(1), 119–139.

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

Huang J et al. (2015) Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. Nature Communications, DOI: 10.1038/ncomms9111.

Huang L et al. (2016) Reveel: large-scale population genotyping using low-coverage sequencing data. Bioinformatics, 32(11):1686–1696.

Li H. (2010) Mathematical Notes on SAMtools Algorithms. https://software.broadinstitute.org/gatk/media/docs/Samtools.pdf.

Li H. (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics, 27(21): 2987–2993.

Li Y et al. (2011) Low-coverage sequencing: implications for design of complex trait association studies. Genome Research, 21:940–951.

Li N and Stephens M. (2003) Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics, 165(4):2213-2233.

McKenna A et al. (2010) The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. Genome Research, 20(9):1297–1303.

Reich DE et al. (2001) Linkage disequilibrium in the human genome. Nature, 411:199–204.

Schaffner SF et al. (2005) Calibrating a coalescent simulation of human genome sequence variation. Genome Research, 15:1576–1583.

Schaid DJ. (2004) Linkage disequilibrium testing when linkage phase is unknown. Genetics, 166(1):505–512.

Schapire RE, Singer Y. (1999) Improved boosting algorithms using confidence-rated predictions. Machine Learning, 37(3):297–336.

The 1000 Genomes Project Consortium. (2010) A map of human genome variation from population-scale sequencing. Nature, 467:1061–1073.

The 1000 Genomes Project Consortium. (2015) A global reference for human genetic variation. Nature, 526:68–74.

The UK10K Consortium. (2015) The UK10K project identifies rare variants in health and disease. Nature, 526:82–90.

Wang Y et al. (2013) An integrative variant analysis pipeline for accurate geno-type/haplotype inference in population NGS data. Genome Research, 23(5):833-842.

Zagorecki A and Druzdzel MJ. (2013) Knowledge engineering for bayesian net-works: how common are noisy-max distributions in practice? IEEE Transactions on Systems, Man, and Cybernetics: Systems, 43(1):186–195.

bioRxiv preprint doi: https://doi.org/10.1101/085936; this version posted November 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Appendix 1 Supplementary Algorithm

Input: genotypes of r reference genomes
feature vectors picked for genotyping the query genomes in Algorithm 1

// calculate $H_j(\boldsymbol{x}_i)$ for all the reference genomes, all the markers, all the weak learners simultaneously by applying our
// iterative algorithm
Do for $T = 1,2,\cdots$
  Do for each LD estimation metric
    Retrieve feature vectors used in Algorithm 1
    Do for $R = 1,2,\cdots$
      Calculate genotype probability $\mathbf{P}^{(\ell)}$ for each marker based on the retrieved feature vectors using the genotypes of the reference genomes: $p^{(\ell)}_{\text{target},h} \propto \Pr\left\{g^{(\ell-1)} = h \middle| \boldsymbol{g}_S^{(\ell-1)}\right\}$ if R is odd and $p^{(\ell)}_{\text{target},h} \propto \Pr\left\{g^{(\ell-1)} = h \middle| \boldsymbol{g}_S^{(\ell-1)}, \text{AF}\right\}$ if R is even
      Find genotypes $\mathbf{G}^{(\ell)}$ that maximize $\mathbf{P}^{(\ell)}$

// train AdaBoost classifiers
For each marker
  Initialize $D_i = 1/r$ for all $i = 1,\cdots,r$
  Do for $j = 1,\cdots,6$
    Calculate the error of weak learner H_j: $\epsilon_j = \sum_{i:H_j(\boldsymbol{x}_i)\neq y_i} D_i$
    If $\epsilon_j > 1/2$
      Set $\alpha_j = \frac{1}{2}\ln\left(\frac{1-\epsilon_j}{\epsilon_j}\right)$
      For each i
        If $H_j(\boldsymbol{x}_i) \neq y_i$
          $D_i \leftarrow D_i \cdot \exp(\alpha_j)$
        Else
          $D_i \leftarrow D_i \cdot \exp(-\alpha_j)$
      Normalize D_i for all i

Output the final hypothesis $H_{\text{final}} = \arg\max_{y\in\{0,1,2\}} \sum_{j:H_j(\boldsymbol{x})=y} \alpha_j$

Algorithm A1. *Training AdaBoost classifiers.*

Appendix 2 Supplementary Table

Table A1. *Precision and sensitivity when apply Reveel and Ref-Reveel to simulated datasets*. We grouped the polymorphic sites according to their allele frequencies, and measured the efficiency of each group as defined in Table 1.

(A) 100 query samples

AF category	Reveel		Ref-Reveel	
	sensitivity (%)	precision (%)	sensitivity (%)	precision (%)
overall	99.2723	99.3161	99.6690	99.6864
<.1%	99.2361	98.6878	99.6528	99.2393
.1-.2%	99.6950	99.5648	99.9564	99.8695
.2-.5%	98.9782	99.5672	99.7315	99.8656
.5-1%	98.4503	99.6744	99.5619	99.9121
1-2%	97.9543	99.5679	99.4169	99.6282
2-5%	97.5855	99.5382	99.0290	99.6462
>5%	99.4071	99.3320	99.7009	99.7248

bioRxiv preprint doi: https://doi.org/10.1101/085936; this version posted November 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(B) 500 query samples

AF category	Reveel		Ref-Reveel	
	sensitivity (%)	precision (%)	sensitivity (%)	precision (%)
overall	99.4584	99.4332	99.7070	99.6529
<.1%	98.8586	96.8487	99.4518	97.4307
.1-.2%	98.9201	99.4049	99.6345	99.6259
.2-.5%	99.0330	99.6986	99.7498	99.8380
.5-1%	99.0635	99.7961	99.5915	99.8529
1-2%	98.4943	99.6909	99.4056	99.7361
2-5%	98.5467	99.3533	99.2119	99.5310
>5%	99.5693	99.4860	99.7445	99.7160

(C) 1000 query samples

AF category	Reveel		Ref-Reveel	
	sensitivity (%)	precision (%)	sensitivity (%)	precision (%)
overall	99.6451	99.5871	99.7699	99.7127
<.1%	99.1392	98.4541	99.5153	98.4730
.1-.2%	98.8904	99.6012	99.6394	99.6668
.2-.5%	99.2103	99.7953	99.6997	99.9029
.5-1%	99.4962	99.8567	99.7362	99.8927
1-2%	99.0546	99.7863	99.4214	99.8108
2-5%	99.0913	99.4451	99.4337	99.6001
>5%	99.7314	99.6699	99.8065	99.7914

Appendix 3 Additional Experiments

The description of additional experiments is available at http://reveel.stanford.edu/supplementary.pdf