Haplotype-based variant detection from short-read sequencing

Erik Garrison and Gabor Marth

July 24, 2014

Abstract

With genomic variant detection methods, we can determine point-wise differences against a reference genome. Widely-used methods fail to reliably characterize the relative phase of proximal alleles. This information is essential for the functional interpretation of genomic material, and can be determined directly from the primary sequence reads. Here we propose such a method, and demonstrate that the use of haplotypes does not only improve our ability to interpret genomic information, and affords a large improvement in detection performance over existing methods. We implement our approach in the *freebayes* Bayesian variant detector. To do so, we extend the now-ubiquitous Bayesian variant detection model to allow for arbitrary genome architecture, ploidy, population structure, numbers of samples, and numbers of alleles. To further improve performance, we extend the model to incorporate an estimate of our confidence that the locus and alleles under analysis can be characterized accurately using our experimental data. These advances allow *freebayes* to outperform all existing variant detection methods at the detection of SNPs, indels, and small complex variants.

1 Motivation

While *statistical phasing* approaches are necessary for the determination of large-scale haplotype structure [Browning and Browning, 2007, Delaneau et al., 2012, Howie et al., 2011, Li et al., 2010], sequencing traces provide short-range phasing information that may be employed directly in primary variant detection to establish phase between proximal alleles. Present read lengths and error rates limit this *physical phasing* approach to variants clustered within tens to hundreds of bases, but as the cost of obtaining long sequencing traces decreases [Branton et al., 2008, Clarke et al., 2009], physical phasing methods will enable the determination of larger haplotype structure directly using only sequence information from a single sample.

Haplotype-based variant detection methods, in which short haplotypes are read directly from sequencing traces, offer a number of benefits over methods which operate on a single position at a time. Haplotype-based methods ensure semantic consistency among described variants by simultaneously evaluating all classes of alleles in the same context. Locally phased genotypes can be used to improve genotyping accuracy in the context of rare variations that can be difficult to impute due to sparse linkage information.

Similarly, they can assist in the design of genotyping assays, which can fail in the context of undescribed variation at the assayed locus. These methods can provide the direct detection of complex variants of clinical significance, such as the BLM^{Ash} allele, a complex block substitution in a helicase gene related to cancer risk [Cleary et al., 2003] or recurrent multi-nucleotide polymorphisms often found in particular cancer types [Huang et al.,

2013]. Directly detecting such alleles from sequencing data decreases the cost of secondary, manual analysis of detected variants, a significant diagnostic cost now generally accepted as necessary for the accurate reporting of non-SNP variation in clinical diagnostic contexts.

The use of longer haplotypes in variant detection can improve detection by increasing the signal to noise ratio of the genotype likelihood space that is used in analysis, provided some degree of independence between sequencing errors. This follows from the fact that the space of possible erroneous haplotypes expands dramatically with haplotype length, while the space of true variation remains constant, with the number of true alleles less than or equal to the ploidy of the sample at a given locus.

The direct detection of haplotypes from alignment data presents several challenges to existing variant detection methods. As the length of a haplotype increases, so does the number of possible alleles within the haplotype, and thus methods designed to detect genetic variation over haplotypes in a unified context must be able to model multiallelism. However, most variant detection methods establish estimates of the likelihood of polymorphism at a given loci using statistical models which assume biallelism [Li, 2011, Marth et al., 1999] and uniform, typically diploid, copy number [DePristo et al., 2011]. Moreover, improper modeling of copy number impedes the accurate detection of small variants on sex chromosomes, in polyploid organisms, or in locations with known copy-number variations, where called alleles, genotypes, and likelihoods should reflect local copy number and global ploidy.

To enable the application of population-level inference methods to the detection of haplotypes, we generalize the Bayesian statistical method described by Marth et al. [1999] to allow multiallelic loci and non-uniform copy number across the samples under consideration. We have implemented this model in FreeBayes [Garrison, 2012a]. In addition to extensions enabling haplotype-based detection, we have incorporated a model of the capacity for the alignments to characterize the locus and alleles in question into our prior probability.

2 Results

2.1 Small variant detection in simulated data

To assess the performance of our method, we used the population genome simulator mutatrix [Garrison, 2012b] to simulate variation in 100 samples over 100 kilobases of human chromosome 20, and the mason read simulator [Holtgrewe, 2010] to generate a simulated Illumina-like 70bp-reads at 10x depth per sample. The data were aligned with Mosaik [Lee and Strömberg, 2012], and variants were called using several popular detection methods capable of simultaneously detecting SNPs and short indels: GATK HaplotypeCaller and UnifiedGenotyper (version 2.7.4) [DePristo et al., 2011], samtools (version 0.1.19-44428cd) [Li et al., 2009], and FreeBayes (version 0.9.9.2-21-g78714b8). To assess each caller's detection performance we generated receiver-operator characteristics (ROCs) using vcfroc [Garrison, 2012c]. We provide results in terms of area under the curve (AUC) for all tested variant callers in table 1.

These results indicate that FreeBayes provides superior performance to the GATK and samtools at all assayed depths and numbers of samples. We observe that the difference in the AUC metric is dominated by both minimum distance from perfect discrimination (perfect

sensitivity and perfect specificity), in which FreeBayes consistently outperforms the other methods, and by apparent hard limitation on sensitivity imposed by the other methods. We hypothesize that the difference in performance for indels, which is larger than that for SNPs, reflects our method's detection of alleles on haplotypes, which improves the signal to noise ratio of the and effectively removes low-frequency alignment artifacts without the need for out-of-band indel and base-quality recalibration methods (we further explore this in 2.3).

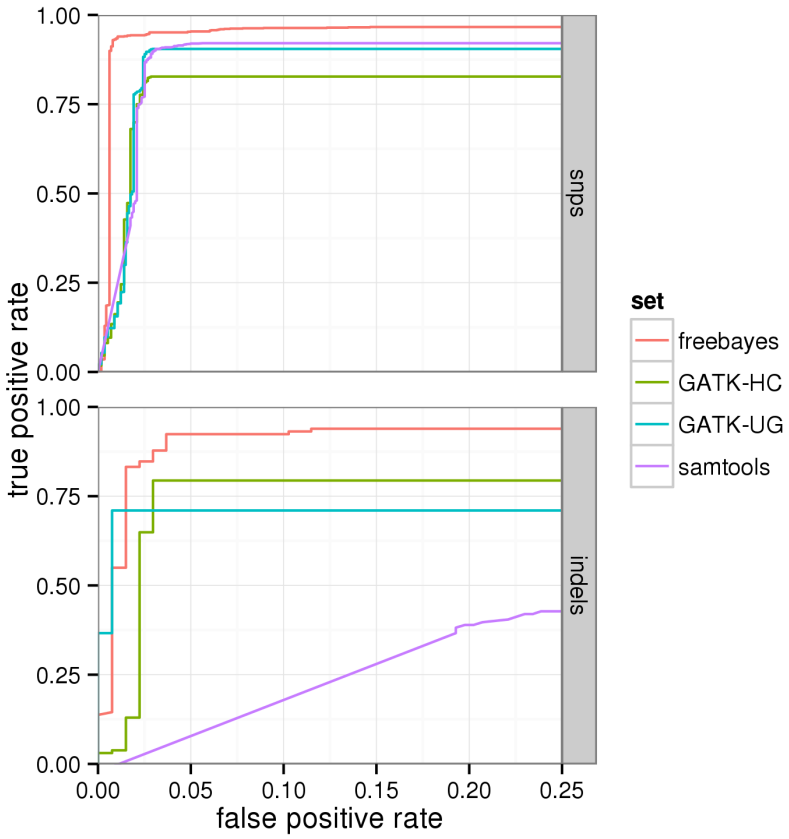

Figure 1: Receiver-operator characteristics (ROCs) for FreeBayes, GATK HaplotypeCaller and UnifiedGenotyper, and samtools on 100 samples at 10x simulated sequencing depth. FreeBayes achieves the highest area under the curve (AUC) 1, with the HaplotypeCaller and samtools each performing next-best for indels and SNPs, respectively.

2.2 Using simulation to assess the direct detection of haplotypes

In order to facilitate our assessment of the method at determining phase between clusters of alleles, we set a mutation rate sufficient to generate many clusters of variants in these simulated samples. We then simulated reads at 20x coverage from the resulting simulated chromosomes using wgsim [Li et al., 2009], aligned the results using Mosaik [Lee and Strömberg, 2012] and ran freebayes on the resulting alignments specifying a haplotype detection length of 10bp. The results were compared to the truth set produced by mutatrix using the utility

variant detector	depth	samples	AUC SNPs	AUC indels
FreeBayes	10	100	0.9594	0.9400
GATK HaplotypeCaller	10	100	0.8155	0.7765
GATK UnifiedGenotyper	10	100	0.8907	0.7073
samtools	10	100	0.9056	0.4698

Table 1: Performance of FreeBayes, GATK HaplotypeCaller and UnifiedGenotyper, and samtools against simulated data.

vcfgeno2haplo in vcflib [Garrison, 2012c] which can construct haplotype observations of a given length from phased genotype information like that produced by mutatrix.

Our results agree with those obtained for other classes of small variants in section 2.1, showing high performance against SNPs (AUC of 0.979) and indels (AUC of 0.948). For complex variants composed between multiple small variants, direct detection provides an AUC of 0.919.

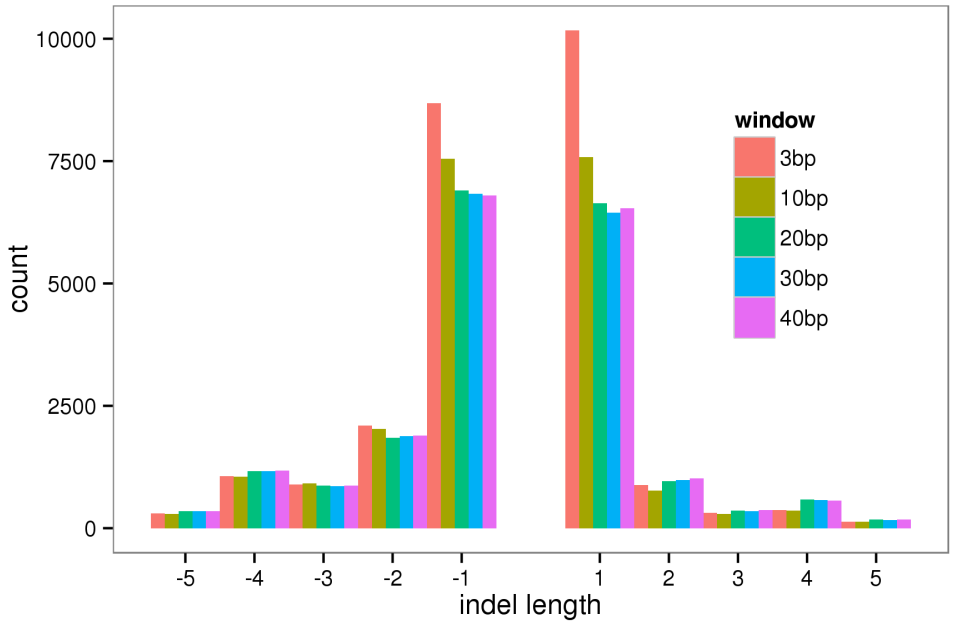

Figure 2: A known error mode of Illumina sequencing methods generates a 1bp insertion artifact that is detected by standard mapping-based variant calling methods. The artifact results in a relative over-abundance of 1bp insertions. Here, we characterize the ability of our method to remove this artifact by detecting variants in a larger detection window. As the calling window size is increased beyond 10bp, the artifact is effectively removed, and the balance between insertions and deletions at a given length normalizes.

2.3 Using haplotype-based variant detection to improve the signal to noise ratio of candidate variants

The fluorescence-based imaging utilized by Illumina sequencing machines is susceptible to errors generated by air bubbles introduced into the flowcell in which the sequencing reaction

takes place. Bubble errors tend to manifest themselves as high-quality 1bp insertions in sequencing traces derived from spots in the affected regions of the sequencing flowcell. These errors are randomly distributed with respect to reference position, but their high frequency in some sequencing runs means that they will spuriously be detected by single-position mapping-based variant detectors when they overlap positionally. We can observe the presence of this error because it causes a preponderance of 1bp insertions over deletions. Typically, 1bp insertions are discoverable in human genomes at a slightly lower frequency than deletions, and thus this error process can be observed by inspection of the indel length-frequency distribution.

To assess the ability of our haplotype-based method to overcome this characteristic error, we detected variants in the previously described AFR191 sample set using a number of different haplotype lengths. The indel detection results (figure 2) indicate that this error mode can be effectively removed from small variant calls by increasing the detection window size to 10bp or greater.

As we increase the length of detected haplotypes, we increase the number of possible erroneous haplotypes without increasing the number of true haplotypes. This effect results in an improved signal to noise ratio for detected variants at larger haplotype sizes. As such, increasing window size in our algorithm allows us to exclude likely insertion artifacts from consideration, as the recurrance of an erroneous haplotype diminishes rapidly with haplotype length. We hypothesize that this effect dominates the improvement in specificity yielded by assembly methods. However, if window sizes are fixed, as is the case in the existing implementations of such methods, sensitivity to rare variation will suffer (discussed in section 2.8).

2.4 Using haplotype-based variant detection to understand genotyping array design failure

Variant calls generated during the pilot phase of the 1000 Genomes Project [1000 Genomes Project Participants, 2012] were used to design a genotyping array, (the Illumina OMNI2.5). Subsequently, many of the alleles on this array (approximately 10%) were found to be putatively monomorphic in the same set of samples, suggesting they resulted from variant detection error.

We investigated these loci using whole-genome calls in the low-coverage cohort in Phase I of the 1000 Genomes Project. We ran freebayes using a haplotype window of 10 base pairs. On comparison with the monomorphic array loci, we found that approximately 90% of the array-monomorphic loci overlap non-SNP or non-biallelic variation in these samples within 10bp of the target SNP, whereas the opposite is true of polymorphic loci— greater than 90% of loci assayed as polymorphic overlap biallelic SNPs.

We observe that many of the apparent failures in variant detection are actually caused by an inability of methods to assess local clusters of variation. The accurate design of genotyping arrays and their use in cross-validation of sequencing-based genotyping performance thus requires information about local haplotypes structure.

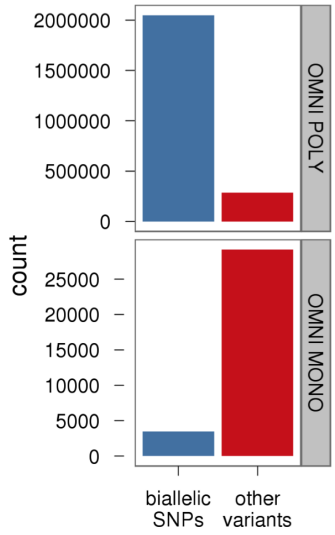

Figure 3: The Omni 2.5 genotyping array includes a number of alleles which consistently report as non-polymorphic (monomorphic) in the 1000 Genomes cohort in which they were originally detected. By detecting variants using our method at a 10bp variant calling window, we demonstrate that more than 90% of the apparently monomorphic loci are not biallelic SNPs, and thus the array design does not match the local variant structure in these samples. By using a haplotype-based approach, groups designing genotyping arrays can avoid this common error mode.

2.5 The importance of accurately modeling copy number variations on sex chromosomes

Our method is currently the only variant detector in common use which provides the ability to call males and females jointly on chromosome X with correct copy number. To evaluate the benefits of this approach, we detected variants in chromosome X for 191 low-coverage 1000 Genomes samples of African ancestry using FreeBayes both with and without copy-number awareness. Comparison of our results to the genotyping array calls (excluding cases of likely array failure due to non-SNP, non-biallelic variation as described in section 2.4) indicates that when calling without copy-number awareness, our genotyping error rate was 7.28%, whereas when calling with awareness of copy-number, the genotyping error rate is only 3.55%. The relatively high error rate is typical in the case of low-coverage data. The difference in overall error rate suggests that there is substantial benefit to directly modeling copy number within the variant detection process.

2.6 Comparing to other methods in low-coverage sequencing data

In the testing phase of the 1000 Genomes Project, participating groups submitted callsets based on 191 samples of African ancestry (AFR191). Results are characterized in figure 2. Unlike other haplotype-based and assembly metods, the approach described in this paper (BC2) provides sensitivity to known variants equivalent to mapping-based methods (BCM,

call set	BC	BCM	BI1	SI1	UM	BC2	BI2	OX1	SI2	OX2	Union	2/9	3/9	4/9	BC cons
SNPs [K]	459	512	481	480	491	495	362	452	252	101	621	548	518	487	543
Omni poly [%]	91.6	98.9	96.5	95.2	97.6	97.4	88.4	87	83.1	44.6	99.3	98.9	98.6	97.6	98.7
Hapmap [%]	94.5	99.4	98	95.6	98.9	98.3	93.6	90.3	91.1	53.7	99.4	99.4	99.3	99	98.6
Omni mono [%]	1.39	1.63	0.29	0.62	0.77	0.56	0.14	1.1	0.72	0.1	3.73	0.97	0.67	0.48	0.65

Table 2: Performance of various variant detection pipelines tested as part of the 1000 Genomes Project. Sets are Boston College; non-haplotype-based method (BC), haplotype-based method described in this paper (BC2), Baylor College of Medicine (BCM), Broad Institute GATK UnifiedGenotyper (BI1), Sanger Institute Samtools (SI1), University of Michigan GlfMultiples (UM), Broad Institute GATK HaplotypeCaller (BI2), Oxford Platypus (OX1), Sanger SGA (SI2), Oxford Cortex (OX2). Union: combination of all variants detected in component methods. 2/9, 3/9, 4/9: voting-based consensus results. BC cons: haplotype-based ensemble method.

center	specificity	sensitivity	caller	optimality	AUC
Oxford Cortex	98	27	OX2	73.02739	0.2646
Pindel	90	52	Pindel	49.03060	0.4680
BC	83	66	BC	38.01316	0.5478
Broad assembly	80	67	BI2	38.58756	0.5360
Sanger	76	69	SI1	39.20459	0.5244
Broad mapping	65	74	BI1	43.60046	0.4810
Oxford Platypus	60	55	OX1	60.20797	0.3300

BC1, SI1, UM). Furthermore, the method's ability to characterize haplotypes in loci which appeared to be monomorphic on the Omni genotyping array allows for discrimination against known artifacts as good as the best mapping-based detection pipelines. Thus we achive a result which is nearly equivalent in sensitivity to the most-sensitive mapping-based method (BCM) and of a similar specificity to that achieved by assembly methods (OX2, SI2 BI2).

2.7 Indel detection performance

2.8 Sensitivity to low-frequency variation

Current methods for haplotype-based variant detection rely on assembly methods, which can be applied globally [Iqbal et al., 2012] or locally [Albers et al., 2011]. These methods remove reference bias from the analysis of short-read sequencing data, but the generation of assemblies of large genomes requires pruning of low-frequency kmer observations. While low-frequency kmers are often generated by sequencing error, in many cases they represent true variation, and thus this pruning reduces the sensitivity of existing assembly methods to true low-frequency variants. In many contexts it is important to accurately and sensitively assess low-frequency variation, such as in experiments involving large numbers of samples, in the detection of sub-clonal mutations in cancer, and in pooled sequencing projects such as viral and metaganomic studies. Our method does provide direct description of haplotypes, but because these haplotypes are generated only where multiple variations segregate an observed haplotype from the reference, it maintains sensitivity to low-frequency variants.

Results from the experiments described in 2.6 demonstrate that our method, while acting as a form of local assembly, does not incur the same sensitivity penalties seen in both local and global assembly methods. We assess this using the count of minor alternate alleles as reported by each caller (figure 5). These results indicate that both global and local assembly

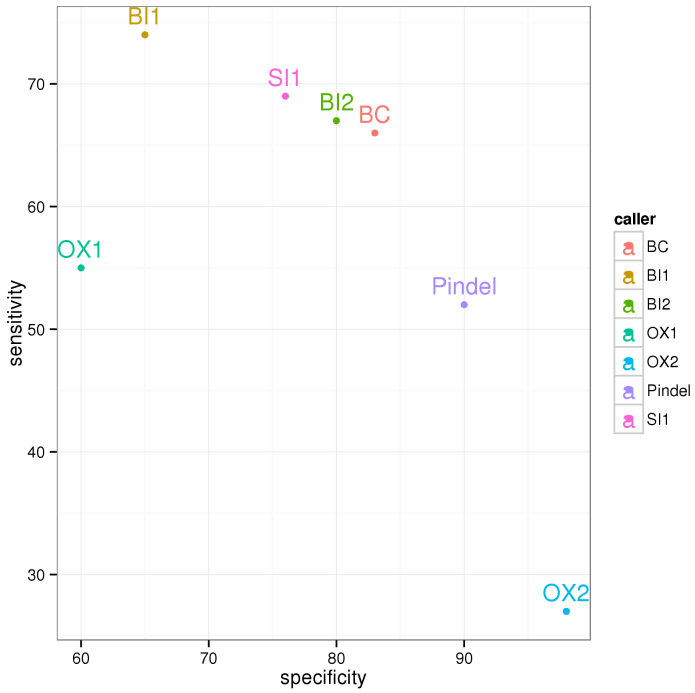

Figure 4: Performance of indel detection methods in 1000 Genomes project on the AFR191 sample set as assesed via high-depth resequencing validation. Sets are Boston College FreeBayes (BC), Broad Institute GATK UnifiedGenotyper (BI1), Sanger Institute Samtools (SI1), Broad Institute GATK HaplotypeCaller (BI2), Oxford Platypus (OX1), Oxford Cortex (OX2).

methods suffer significant decrease in sensitivity to low-frequency variants, although the effect is less severe for local assembly. In this test our method performs as well or better than the GATK UnifiedGenotyper, which is purely mapping-based, the GATK HaplotypeCaller, which uses local assembly, and the string graph assembler (SGA) which is a global, reference-free assembly approach.

2.9 Haplotype-based consolidation of small variant calls

Ensemble methods have been shown to provide superior performance to component inference methods in many contexts [Opitz and Maclin, 1999]. We hypothesize that ensemble approaches to variant detection from short-read sequencing may provide improved performance in the context of variant detection. While ensemble approaches have already been successfully applied to SNPs in large-scale resequencing projects [**?**], their application to other variant classes is problematic because detectors can output the same allele in slightly different ways. In the 1000 Genomes Phase I integrated callset, we find 181,567 cases in which incorrect description of small variants results in an "impossible" haplotype, such as where a small variant is phased inside of of a deletion, or multiple deletions overlap within 50 base pairs. We can avoid such errors by using an approach that establishes the local haplotype structure around variants prior to using statistical phasing approaches to estimate

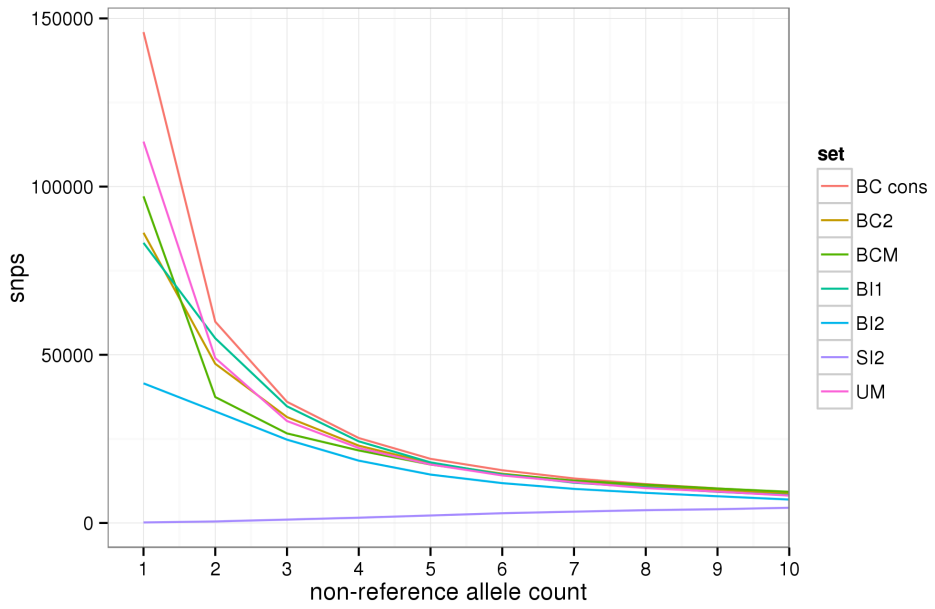

Figure 5: Sensitivity to low-frequency variants of various detection methods, as assessed in 191 samples of African ancestry in the 1000 Genomes low-coverage cohort. BC2 is FreeBayes, BI1 is the GATK UnifiedGenotyper, BI2 is the GATK HaplotypeCaller, and SI2 is the global assembler SGA.

large-scale haplotypes.

3 Methods

3.1 Definitions

At a given genetic locus we have n samples drawn from a population, each of which has a copy number or multiplicity of m within the locus. We denote the number of copies of the locus present within our set of samples as $M = \sum_{i=1}^{n} m_i$. Among these M copies we have K distinct alleles, $b_1, \ldots, b_K$ with allele counts $c_1, \ldots, c_K$ and frequencies $f_1, \ldots, f_K$. Each individual has an unphased genotype G_i comprised of k_i distinct alleles $b_{i_1}, \ldots, b_{k_i}$ with corresponding genotype allele counts $c_{i_1}, \ldots, c_{i_{k_i}}$ and genotype allele frequencies $f_{i_1}, \ldots, f_{i_{k_i}} : f_i = c_i/k_i$. G_i may be equivalently expressed as a multiset of alleles $B_i : |B_i| = m_i$. For the purposes of our analysis, we assume that we cannot accurately discern phasing information outside of the haplotype detection window, so our G_i are unordered and all G_i containing equivalent alleles and frequencies are regarded as equivalent. Assume a set of s_i sequencing observations $r_{i_1}, \ldots, r_{i_{s_i}} = R_i$ for each sample in our set of n samples such that there are $\sum_{i=1}^{n} |R_i|$ reads at the genetic locus under analysis. We use q_i to denote the mapping quality, or probability that the read r_i is mis-mapped against the reference.

3.2 A Bayesian approach

To genotype the samples at a specific locus, we could simply apply a Bayesian statistic relating $P(G_i|R_i)$ to the likelihood of sequencing errors in our reads and the prior likelihood of specific genotypes. However, this maximum-likelihood approach limits our ability to incorporate information from other individuals in the population under analysis, which can improve detection power.

Given a set of genotypes $G_1, \ldots, G_n$ and a set of observations observations $R_1, \ldots, R_n$ for all individuals at the current genetic locus, we can use Bayes' theorem to related the probability of a specific combination of genotypes to both the quality of sequencing observations and *a priori* expectations about the distribution of alleles within a set of individuals sampled from the same population:

$$P(G_1, \ldots, G_n|R_1, \ldots, R_n) = \frac{P(G_1, \ldots, G_n)P(R_1, \ldots, R_n|G_1, \ldots, G_n)}{P(R_1, \ldots, R_n)} \tag{1}$$

$$P(G_1, \ldots, G_n|R_1, \ldots, R_n) = \frac{P(G_1, \ldots, G_n)\prod_{i=1}^{n} P(R_i|G_i)}{\sum_{\forall G_1,\ldots,G_n} P(G_1, \ldots, G_n)\prod_{i=1}^{n} P(R_i|G_i)} \tag{2}$$

In this formulation, $P(R_1, \ldots, R_n|G_1, \ldots, G_n) = \prod_{i=1}^{n} P(R_i|G_i)$ represents the likelihood that our observations match a given genotype combination (our data likelihood), and $P(G_1, \ldots, G_n)$ represents the prior likelihood of observing a specific genotype combination. We estimate the data likelihood as the joint probability that the observations for a specific individual support a given genotype. We use a neutral model of allele diffusion conditioned on an estimated population mutation rate to estimate the prior probability of sampling a given collection of genotypes.

Except for situations with small numbers of samples and potential alleles, we avoid the explicit evaluation of the posterior distribution as implied by (2), instead using a number of

optimizations to make the algorithm tractable to apply to very large datasets (see section 4.3).

3.3 Estimating the probability of sequencing observations given an underlying genotype, $P(R_i|G)$

Given a set of reads $R_i = r_{i_1}, \ldots, r_{i_{s_i}}$ from a sample at a given locus, we can extract a set of k_i observed alleles $B'_i = b'_1, \ldots, b'_{k_i}$ corresponding to underlying alleles $b_1, \ldots, b_i$ which encapsulate the potential set of represented variants at the locus in the given sample, including erroneous observations. Each of these observed alleles b'_i has a count o_f within the observations of the individual sample : $\sum_{j=1}^{k_i} o_j = s_i$ and corresponds to a true allele b_i.

The probability of obtaining a single observation b'_i provided a genotype in a single sample is:

$$P(b'_i|G) = \sum_{\forall(b_i \in G)} f_i P(b'_i|b_i) \tag{3}$$

Here f_i is the genotype allele frequency of b_i in G. We observe that the process generating reads from a given locus in a given sample is a multinomial process in which the sampling probabilities for each allele are governed by both the counts of alleles in the genotype and the error process that generates b'_i from underlying b_i. However, for the case that the base observation agrees with the underlying genotype, sampling probability dominates the probability that the observations are derived from a given genotype, and in the case when the observation does not agree with the genotype, the dominant process is the observation error. Following this observation we introduce the approximation that:

$$P(b'|b) = \begin{cases} 1 & \text{if } b' = b \\ P(error) & \text{if } b' \neq b \end{cases} \tag{4}$$

Here $P(error)$ is the probability that the base is erroneous as determined by the sequencing process used to generate the reads from the sample. Provided this approximation, we can estimate the probability of a given set of reads conditioned on an underlying genotype by using the multinomial sampling probability to estimate the probability of obtaining the observations that support the genotype scaled by the probability that the observations that disagree with the genotype are erroneous:

$$P(R_i|G) \approx \binom{s_i}{o_1, \ldots, o_{k_i}} \prod_{j=1}^{k_i} f_{i_j}^{o_j} \prod_{l=1}^{s_i} P(b'_l|b_l) \tag{5}$$

3.4 Genotype combination priors, $P(G_1, \ldots, G_n)$

3.4.1 Decomposition of prior probability of genotype combination

Let $G_1, \ldots, G_n$ denote the set of genotypes at the locus and $f_1, \ldots, f_k$ denote the set of allele frequencies which corresponds to these genotypes. We estimate the prior likelihood of

observing a specific combination of genotypes within a given locus by decomposition into resolvable terms:

$$P(G_1, \ldots, G_n) = P(G_1, \ldots, G_n \cap f_1, \ldots, f_k) \tag{6}$$

The probability of a given genotype combination is equivalent to the intersection of that probability and the probability of the corresponding set of allele frequencies. This identity follows from the fact that the allele frequencies are derived from the set of genotypes and we always will have the same $f_1, \ldots, f_k$ for any equivalent $G_1, \ldots, G_n$.

Following Bayes' Rule, this identity further decomposes to:

$$P(G_1, \ldots, G_n \cap f_1, \ldots, f_k) = P(G_1, \ldots, G_n | f_1, \ldots, f_k) P(f_1, \ldots, f_k) \tag{7}$$

We now can estimate the prior probability of $G_1, \ldots, G_n$ in terms of the genotype combination sampling probability, $P(G_1, \ldots, G_n | f_1, \ldots, f_k)$, and the probability of observing a given allele frequency in our population, $P(f_1, \ldots, f_k)$.

3.4.2 Genotype combination sampling probability $P(G_1, \ldots, G_n | f_1, \ldots, f_k)$

The multinomial coefficient $\binom{M}{c_1, \ldots, c_k}$ gives the number of ways which a set of alleles with frequencies $f_1, \ldots, f_k : f_i = c_i/M$ may be distributed among M copies of a locus. For phased genotypes $\hat{G}_i$ the probability of sampling a specific $\hat{G}_1, \ldots, \hat{G}_n$ given allele frequencies $f_1, \ldots, f_k$ is thus provided by the inverse of this term:

$$P(\hat{G}_1, \ldots, \hat{G}_n | f_1, \ldots, f_k) = \binom{M}{c_1, \ldots, c_k}^{-1} \tag{8}$$

However, our model is limited to unphased genotypes because our primary data only allows phasing within a limited context. Consequently, we must adjust (8) to reflect the number of phased genotypes which correspond to the unphased genotyping $G_1, \ldots, G_n$. Each unphased genotype corresponds to as many phased genotypes as there are permutations of the alleles in G_i. Thus, for a given unphased genotyping $G_1, \ldots, G_n$, there are $\prod_{i=1}^{n} \binom{m_i}{c_{i_1}, \ldots, c_{i_{k_i}}}$ phased genotypings.

In conjunction, these two terms provide the probability of sampling a particular unphased genotype combination given a set of allele frequencies:

$$P(G_1, \ldots, G_n | f_1, \ldots, f_k) = \binom{M}{c_1, \ldots, c_k}^{-1} \prod_{i=1}^{n} \binom{m_i}{c_{i_1}, \ldots, c_{i_{k_i}}} \tag{9}$$

In the case of a fully diploid population, the product of all possible multiset permutations of all genotypes reduces to 2^h, where h is the number of heterozygous genotypes, simplifying (9) to:

$$P(G_1, \ldots, G_n | f_1, \ldots, f_k) = 2^h \binom{M}{c_1, \ldots, c_k}^{-1} \tag{10}$$

3.4.3 Derivation of $P(f_1, \ldots, f_k)$ by Ewens' sampling formula

Provided our sample size n is small relative to the population which it samples, and the population is in equilibrium under mutation and genetic drift, the probability of observing a given set of allele frequencies at a locus is given by Ewens' sampling formula [Ewens, 1972]. Ewens' sampling formula is based on an infinite alleles coalescent model, and relates the probability of observing a given set of allele frequencies to the number of sampled chromosomes at the locus (M) and the population mutation rate θ.

The application of Ewens' formula to our context is straightforward. Let a_f be the number of alleles among $b_1, \ldots, b_k$ whose allele count within our set of samples is c. We can thus transform our set of frequencies $f_1, \ldots, f_k$ (equivalently, allele counts, $c_1, \ldots, c_k$) into a set of non-negative frequency counts $a_1, \ldots, a_M : \sum_{c=1}^{M} ca_c = M$. As many $c_1, \ldots, c_k$ can map to the same $a_1, \ldots, a_M$, this transformation is not invertible, but it is unique from $a_1, \ldots, a_M$ to $c_1, \ldots, c_k$.

Having transformed a set of frequencies over alleles to a set of frequency counts over frequencies, we can now use Ewens' sampling formula to approximate $P(f_1, \ldots, f_k)$ given θ:

$$P(f_1, \ldots, f_k) = P(a_1, \ldots, a_M) = \frac{M!}{\theta \prod_{z=1}^{M-1}(\theta + z)} \prod_{j=1}^{M} \frac{\theta^{a_j}}{j^{a_j} a_j!} \tag{11}$$

In the bi-allelic case in which our set of samples has two alleles with frequencies f_1 and f_2 such that $f_1 + f_2 = M$:

$$P(a_{f_1} = 1, a_{f_2} = 1) = \frac{M!}{\prod_{z=1}^{M-1}(\theta + z)} \frac{\theta}{f_1 f_2} \tag{12}$$

While in the monomorphic case, where only a single allele is represented at this locus in our population, this term reduces to:

$$P(a_M = 1) = \frac{(M-1)!}{\prod_{z=1}^{M-1}(\theta + z)} \tag{13}$$

In this case, $P(f_1, \ldots, f_k) = 1 - \theta$ when $M = 2$. This is sensible as θ represents the population mutation rate, which can be estimated from the pairwise heterozygosity rate of any two chromosomes in the population [Tajima, 1983, Watterson, 1975].

3.5 Expanding the model to incorporate the observability of the locus and alleles

The bayesian model described in section 3.2 can generate posterior estimates based on sequencing quality information and genotype distribution in a panel of samples. However, this estimate can incorporate only information captured in base quality information and read counts. This may fail to assess the ability of the sequencing and alignment methods to accurately characterize the locus and alleles that we genotype, which is an important consideration for downstream use of sequencing-derived genotype data.

Previous authors have addressed this limitation by adding post-processing steps to recalibrate the estimated quality of variants using training sets of known variants and known

artifacts. Once variant calls have been made we can annotate them with a variety of features and apply standard machine learning methods to "recalibrate" the quality estimates produced from genotype distribution, allele frequency, observation counts, and base quality. For instance, DePristo et al. [2011] apply a guassian mixture model (VQSR) to features extracted from putatively polymorphic loci to remove variants which are outliers in multiple feature dimensions.

Problematically, such an approach requires a training set, which may not be applicable in contexts with limited validation data, such as is commonly the case in non-model organisms. Furthermore, the training set may bias our results towards established patterns, decreasing sensitivity to novel variation that might have been previously uncharacterized due to technological limitations.

In contrast, we address the issue of loci sequencability in a general, *a priori* fashion by extending the traditional Bayesian variant detection model to incorporate an indicator, S, which describes the ability of our sequencing and alignment methods to characterize the locus we are considering. We define $S = true$ when we can sequence the locus and alleles and $S = false$ otherwise, and redefine our model (2) to estimate the posterior probability of a particular set of genotypes in our samples $(G_1, \ldots, G_n)$ and that the locus is sequenceable (S) given our aggregate read evidence $(R_1, \ldots, R_n)$:

$$P(G_1, \ldots, G_n, S|R_1, \ldots, R_n) = \frac{P(G_1, \ldots, G_n)P(S)\prod_{i=1}^{n} P(R_i|G_i)}{\sum_{\forall G_1,\ldots,G_n}(P(G_1, \ldots, G_n)P(S)\prod_{i=1}^{n} P(R_i|G_i))} \tag{14}$$

We will describe the development of $P(S)$ using aggregate statistics built from the read evidence overlapping the locus in section 3.6.

3.6 Estimation of the probability that the locus is sequencable $P(S)$

For accurate variant detection via resequencing, we require that the locus in question is sequencable. That is, we require that the reference is accurate, that we have an accurate model of copy number at the locus, that we have genomic coverage, and that reads can be aligned to the alleles of interest in the region. In a case where these conditions are met, we assume $S = true$. Where it is not, $S = false$.

The sequenceability of a locus and its alleles is assumed under previous Bayesian variant detection models [Li, 2011, Li et al., 2009, Marth et al., 1999]. Uncertainty about the genomic model characterization has been incorporated into data likelihoods or detection thresholds using read mapping quality [Li et al., 2008, Wang et al., 2013]. In practice, the incorporation of confidence in the characterizability of the locus and alleles requires the reclassification of variant calls on the basis of aggregate metrics describing the calls, such as the ratio of observations for an alternate allele to those for a reference allele among apparent heterozygotes, or the average observation quality (base quality) of alleles. In practice, variant detectors have been modified to supply annotations to downstream classifiers that "recalibrate" the quality estimates, but no existing method has incorporated estimates of sequenceability into the Bayesian inference model.

A quality score recalibrator utilizes training data, particularly as sets of known variants or validated errors, to describe the distribution of true events and errors across the space of

possible annotations in the data set to be recalibrated. The variant calling error function as described by these aggregate metrics can then be approximated using a variety of machine learning methods, such as support vector machines [O'Fallon et al., 2013] or a gaussian mixture model as implemented in the GATK's Variant Quality Score Recalibrator (VQSR).

We observe that S is proportional to a number of variables which can be estimated directly from the observations covering a genomic locus. For instance, if the locus and alleles are observable without bias, we expect the count of observations of a sample supporting a particular alternate allele $R_i \equiv b$ to approximate its frequency in the correct genotype G_i for the sample, $|R_i \equiv b|/|R_i| \approx |b \in G_i|/m_i$. Deviation from this expectation which is observed across many samples may indicate problems mapping reads containing the alternate against the reference, or hidden copy-number variations or paralogs that might frustrate our observation of the locus. Similarly, if we use whole-genome shotgun techniques, we have a number of other expectations about behavior of the reads in aggregate with respect to a particular allele and locus. We will express these in terms of a bias terms B_* that equal 0 when there is no bias for a particular metric.

In an unbiased context, we expect half of our reads to place to either side of the locus (placement bias B_p):

$$P(B_p = 0) \propto \binom{|R_{left}|}{|R|} 0.5^{|R_{left}|} \tag{15}$$

We expect half to contain the allele in the first half of their length (cycle bias B_c):

$$P(B_c = 0) \propto \binom{|R_{start}|}{|R|} 0.5^{|R_{start}|} \tag{16}$$

Half should be derived from each strand of DNA (strand bias B_s):

$$P(B_s = 0) \propto \binom{|R_{forward}|}{|R|} 0.5^{|R_{forward}|} \tag{17}$$

And, the aggregate fraction of reads supporting a particular allele in samples with a particular genotype should approximate the frequency of the allele in that particular genotype (allele balance, B_a). Recall that the distinct alleles in a particular set of genotypes are $b_1, \ldots, b_K$, the corresponding allele frequencies in the set are $f_1, \ldots, f_k$, and the observation counts are represented by $o_1, \ldots, o_K$:

$$P(B_a = 0) \propto \prod_{\forall g \in \{G\}} \binom{|R|}{o_1, \ldots, o_K} \prod_{j=1}^{K} f_j^{o_j} \tag{18}$$

We use these relationships to determine relationships in $P(S)$ under various configurations of alleles and genotypes in the samples:

$$P(S) \propto P(B_p = 0)P(B_c = 0)P(B_s = 0)P(B_a = 0) \tag{19}$$

4 Direct detection of phase from short-read sequencing

By modeling multiallelic loci, this Bayesian statistical framework provides the foundation for the direct detection of longer, multi-base alleles from sequence alignments. In this section we describe our implementation of a haplotype-based variant detection method based on this model.

Our method assembles haplotype observations over minimal, dynamically-determined, reference-relative windows which contain multiple segregating alleles. To be used in the analysis, haplotype observations must be derived from aligned reads which are anchored by reference-matching sequence at both ends of the detection window. These haplotype observations have derived quality estimations which allow their incorporation into the general statistical model described in section 3. We then employ a gradient ascent method to determine the maximum *a posteriori* estimate of a mutual genotyping over all samples under analysis and establish an estimate of the probability that the loci is polymorphic.

4.1 Parsing haplotype observations from sequencing data

In order to establish a range of sequence in which multiple polymorphisms segregate in the population under analysis, it is necessary to first determine potentially polymorphic windows in order to bound the analysis. This determination is complicated by the fact that a strict windowing can inappropriately break clusters of alleles into multiple variant calls. We employ a dynamic windowing approach that is driven by the observation of multiple proximal reference-relative variations (SNPs and indels) in input alignments.

Where reference-relative variations are separated by less than a configurable number of non-polymorphic bases in an aligned sequence trace, our method combines them into a single haplotype allele observation, H_i. The observational quality of these haplotype alleles is given as $\min(q_l \,\forall\, b'_i \in H_i,\, Q_i)$, or the minimum of the supporting read's mapping quality and the minimum base quality of the haplotype's component variant allele observations.

4.2 Determining a window over which to assemble haplotype observations

At each position in the reference, we collect allele observations derived from alignments as described in 4.1. To improve performance, we apply a set of input filters to exclude alleles from the analysis which are highly unlikely to be true. These filters require a minimum number of alternate observations and a minimum sum of base qualities in a single sample in order to incorporate a putative allele and its observations into the analysis.

We then determine a haplotype length over which to genotype samples by a bounded iterative process. We first determine the allele passing the input filters which is longest relative to the reference. For instance, a longer allele could be a multi-base indel or a composite haplotype observation flanked by SNPs. Then, we parse haplotype observations from all the alignments which fully overlap this window, finding the rightmost end of the longest haplotype allele which begins within the window. This rightmost position is used to update the haplotype window length, and a new set of haplotype observations are assembled from the reads fully overlapping the new window. This process repeats until the rightmost end of the window is not partially overlapped by any haplotype observations which pass the

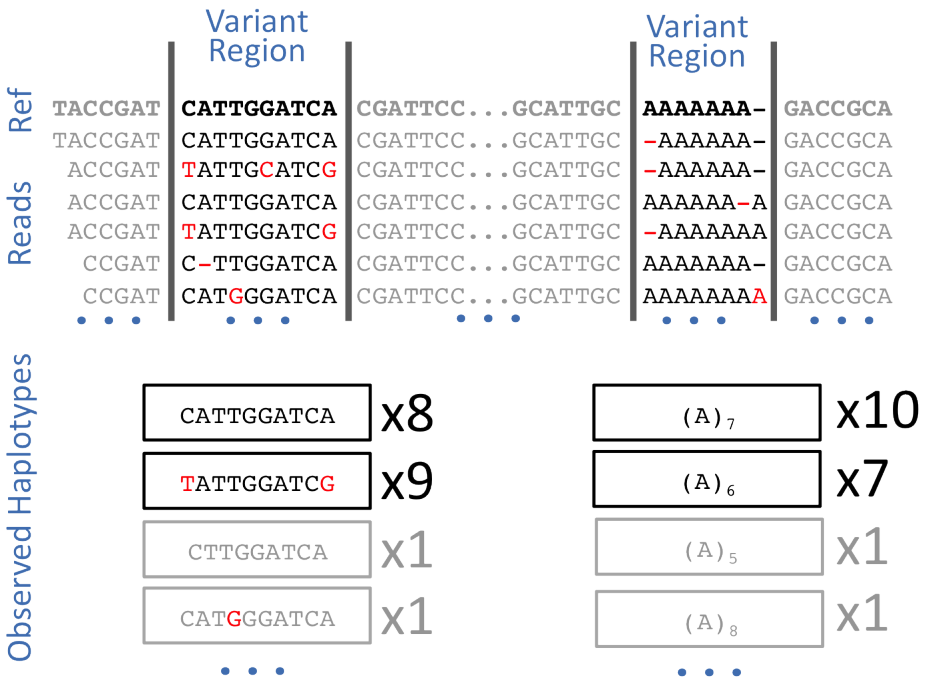

Figure 6: The direct detection of phase from short-read sequencing traces and counting of haplotypes across dynamically-determined windows.

input filters. This method will converge given reads have finite length and the only reads which fully overlap the detection window are used in the analysis.

4.3 Detection and genotyping of local haplotypes

Once a window for analysis has been determined, we parse all fully-overlapping reads into haplotype observations which are anchored at the boundaries of the window. Given these sets of sequencing observations $r_{i_1}, \ldots, r_{i_{s_i}} = R_i$ and data likelihoods $P(R_i|G_i)$ for each sample and possible genotype derived from the putative alleles, we then determine the probability of polymorphism at the locus given the Bayesian model described in section 3.

To establish a maximum *a posteriori* estimate of the genotype for each sample, we employ a convergent gradient ascent approach to the posterior probability distribution over the mutual genotyping across all samples under our Bayesian model. This process begins at the genotyping across all samples $G_1, \ldots, G_n$ where each sample's genotype is the maximum-likelihood genotype given the data likelihood $P(R_i|G_i)$:

$$G_1, \ldots, G_n = \underset{G_i}{\operatorname{argmax}}\, P(R_i|G_i) \tag{20}$$

The posterior search then attempts to find a genotyping $G_1, \ldots, G_n$ in the local space of genotypings which has higher posterior probability under the model than this initial genotyping. In practice, this step is done by searching through all genotypings in which a single sample has up to the Nth best genotype when ranked by $P(R_i|G_i)$, and N is a

small number (e.g. 2). This search starts with some set of genotypes $G_1, \ldots, G_n = \{G\}$ and attempts to find a genotyping $\{G\}'$ such that:

$$P(\{G\}'|R_1, \ldots, R_n) > P(\{G\}|R_1, \ldots, R_n) \tag{21}$$

$\{G\}'$ is then used as a basis for the next update step. This search iterates until convergence, but in practice must be bounded at a fixed number of steps in order to ensure optimal performance. As the quality of input data increases in coverage and confidence, this search will converge more quickly because the maximum-likelihood estimate will lie closer to the maximum *a posteriori* estimate under the model.

This method incorporates a basic form of genotype imputation into the detection method, which in practice improves the quality of raw genotypes produced in primary allele detection and genotyping relative to methods which only utilize a maximum-likelihood method to determine genotypes. Furthermore, this method allows for the determination of marginal genotype likelihoods via the marginalization of assigned genotypes for each sample over the posterior probability distribution.

4.4 Probability of polymorphism

Provided a maximum *a posteriori* estimate of the genotyping of all the individuals in our sample, we might like establish an estimate of the quality of the genotyping. For this, we can use the probability that the locus is polymorphic, which means that the number of distinct alleles at the locus, K, is greater than 1. While in practice the space of possible genotypings is too great to integrate over, it is possible to derive the probability that the loci is polymorphic in our samples by summing across the monomorphic cases:

$$P(K > 1|R_1, \ldots, R_n) = 1 - P(K = 1|R_1, \ldots, R_n) \tag{22}$$

Equation (22) thus provides the probability of polymorphism at the site, which is provided as a quality estimate for each evaluated locus in the output of FreeBayes.

4.5 Marginal likelihoods of individual genotypes

Similarly, we can establish a quality estimate for a single genotype by summing over the marginal probability of that specific genotype and sample combination under the model. The marginal probability of a given genotype is thus:

$$P(G_j|R_i, \ldots, R_n) = \sum_{\forall(\{G\}:G_j \in \{G\})} P(\{G\}|R_i, \ldots, R_n) \tag{23}$$

In implementation, the estimation of this term requires us to must sample enough genotypings from the posterior in order to obtain well-normalized marginal likelihoods. In practice, we marginalize from the local space of genotypings in which each individual genotype is no more than a small number of steps in one sample from the maximum *a posteriori* estimate of $G_i, \ldots, G_n$. This space is similar to that used during the posterior search described in section 4.3. We apply (23) to it to estimate marginal genotype likelihoods for the most likely

individual genotypes, which are provided for each sample at each site in the output of our implementation.

References

1000 Genomes Project Participants T. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**: 56–65.

Albers CA, Lunter G, MacArthur DG, McVean G, Ouwehand WH, and Durbin R. 2011. Dindel: accurate indel calls from short-read data. *Genome Res.* **21**: 961–973.

Branton D, Deamer DW, Marziali A, Bayley H, Benner SA, Butler T, Di Ventra M, Garaj S, Hibbs A, Huang X, et al.. 2008. The potential and challenges of nanopore sequencing. *Nat. Biotechnol.* **26**: 1146–1153.

Browning SR and Browning BL. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**: 1084–1097.

Clarke J, Wu HC, Jayasinghe L, Patel A, Reid S, and Bayley H. 2009. Continuous base identification for single-molecule nanopore DNA sequencing. *Nat Nanotechnol* **4**: 265–270.

Cleary SP, Zhang W, Di Nicola N, Aronson M, Aube J, Steinman A, Haddad R, Redston M, Gallinger S, Narod SA, et al.. 2003. Heterozygosity for the BLM(Ash) mutation and cancer risk. *Cancer Res.* **63**: 1769–1771.

Delaneau O, Marchini J, and Zagury JF. 2012. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**: 179–181.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al.. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**: 491–498.

Ewens WJ. 1972. The sampling theory of selectively neutral alleles. *Theor Popul Biol* **3**: 87–112.

Garrison E. 2012a. FreeBayes source repository. `https://github.com/ekg/freebayes`.

Garrison E. 2012b. mutatrix population genome simulator. `https://github.com/ekg/mutatrix`.

Garrison E. 2012c. vcflib: variant call file processing and manipulation utilities. `https://github.com/ekg/vcflib`.

Holtgrewe M. 2010. Mason a read simulator for second generation sequencing data. Technical Report TR-B-10-06, Institut fr Mathematik und Informatik, Freie Universitt Berlin.

Howie B, Marchini J, and Stephens M. 2011. Genotype imputation with thousands of genomes. *G3 (Bethesda)* **1**: 457–470.

Huang FW, Hodis E, Xu MJ, Kryukov GV, Chin L, and Garraway LA. 2013. Highly recurrent TERT promoter mutations in human melanoma. *Science* **339**: 957–959.

Iqbal Z, Caccamo M, Turner I, Flicek P, and McVean G. 2012. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nat. Genet.* **44**: 226–232.

Lee WP and Strömberg M. 2012. MOSAIK reference-guided aligner for next-generation sequencing technologies. `https://github.com/wanpinglee/MOSAIK`.

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

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

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

Li Y, Willer CJ, Ding J, Scheet P, and Abecasis GR. 2010. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* **34**: 816–834.

Marth GT, Korf I, Yandell MD, Yeh RT, Gu Z, Zakeri H, Stitziel NO, Hillier L, Kwok PY, and Gish WR. 1999. A general approach to single-nucleotide polymorphism discovery. *Nat. Genet.* **23**: 452–456.

O'Fallon BD, Wooderchak-Donahue W, and Crockett DK. 2013. A support vector machine for identification of single-nucleotide polymorphisms from next-generation sequencing data. *Bioinformatics* **29**: 1361–1366.

Opitz DW and Maclin R. 1999. Popular ensemble methods: An empirical study. *J. Artif. Intell. Res. (JAIR)* **11**: 169–198.

Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**: 437–460.

Wang Y, Lu J, Yu J, Gibbs RA, and Yu F. 2013. An integrative variant analysis pipeline for accurate genotype/haplotype inference in population NGS data. *Genome Res.* **23**: 833–842.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. *Theor Popul Biol* **7**: 256–276.