A monoallelic representation of genomic alignment data

Erik Garrison

November 4, 2014

A standard representation of genomic observation data involves the compression of sequence reads into a locus-based model where a set of alleles $A = \{a_1, \ldots, a_n\}$ and a particular copy number p are used to generate a set of genotypes $G = \{g_1, \ldots, g_{\binom{p}{n}}\}$ and corresponding likelihoods L of read evidence R given genotype $L = \{P(R|g_1), \ldots, P(R|g_{\binom{p}{n}})\}$ at a single genomic position. Note that in this model genotypes are multisets of alleles $g = \{a_1, \ldots, a_p\}$, and the number of genotypes is given by the multiset coefficient $\binom{p}{n}$. These likelihoods can then be used as the basis for further computation of data and statistics of interest, for instance to infer the phased haplotypes of a given sample using the aggregate evidence from many hundreds of thousands of other individuals for which we have likelihood information. Conveying likelihoods rather than hard calls for particular genotypes allows the preservation of information about uncertainty, and any similar compression scheme should preserve this quality. This approach has been standardized by groups like the 1000 Genomes Project and is encapsulated by the GL and PL fields provided in Variant Call Format (VCF).

This site-based (or *multiallelic*) model is conceptually straightforward to work with when our alleles are very short and do not overlap. As we consider longer alleles, such as deletions, we might want to consider all of the overlapping alleles within the same context in order to yield a likelihood of the read data for each potential genotype. Problematically, this feature of the model necessitates the redefinition of the context in which we describe our genotype likelihoods as we incorporate data from different genomes or collections of genomes. The multiallelic compression of read evidence into genotype likelihoods thus requires us to return to the original read evidence to regenerate likelihoods whenever changes in the set of alleles at a given site would alter the locus over which we define genotype likelihoods. An alternative approach would be to collapse read evidence on an allele-specific basis (a *monoallelic* model). Doing so would simplify the process of combining data from disparate sources, as a multiallelic model can be reconstructed out of the monoallelic one, and monoallelic descriptions of the data can be readily combined based on a sequence-based interpretation of the relationship between the alleles. To clarify the relationship between the monoallelic and multiallelic models, I will first describe a standard genotype

likelihood definition, and then show how it can be readily reconstructed from per-allele annotations.

To define our likelihoods, we require a few quantities of interest describing the sample and evidence at a specific genomic locus. We have s reads $R = \{r_1, \ldots, r_s\}$, observation counts for the particular alleles $o_i = |r \in R : r \equiv a_i|$, genotype allele fractions $f_i = |a \in g : a = a_i|/p$, and the number of unique alleles in a particular genotype k.

We introduce an approximation to simplify calculation of $P(r|g)$. In 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. Each read r is assigned a "quality" q that represents the probability that the read was derived from an underlying sequence that is not the same as that which it represents at a given position. As such, we describe the probability of observing a particular read given an underlying allele:

$$P(r|g) = \begin{cases} 1 & \text{if } r \equiv a \in g \\ q & \text{if } r \not\equiv a \in g \end{cases} \tag{1}$$

In other words, the probability of a single observation r given a particular genotype g is approximately the error estimate for that observation when the allele it supports is not part of the genotype in question.

Using these definitions, we can define a genotype likelihood function that combines the multinomial sampling probability of obtaining a given set of observations from an underlying genotype and the quality information encoded by the sequencing data:

$$P(R|g) \approx \binom{s}{o_1, \ldots, o_k} \prod_{j=1}^{k} f_j^{o_j} \prod_{i=1}^{s} P(r_i|g) \tag{2}$$

This model is very similar to that used in most genotyping algorithms based on short-read alignment data from contemporary DNA sequencing platforms.

We would like to modify our data representation to decouple the site-based representation implied by the genotype likelihood format from the compression of sequencing information. Although this format would not be identical to a site-based compression of the evidence, it can be equivalent. Here I show that likelihoods can be calculated "on-the-fly" from an intermediate data representation based on only the read counts $O = \{o_1, \ldots, o_k\}$ and partial products of $P(r|a)$ over the alleles implied by the sequencing data.

First, we can partition the likelihood calculation to operate over the distinct k alleles at the locus without changing our likelihood function:

$$P(R|g) \approx \binom{s}{o_1, \ldots, o_k} \prod_{j=1}^{k} f_j^{o_j} \prod_{m=1}^{k} \prod_{i=1}^{s} P(r_i|a_m) \tag{3}$$

Crucially, this representation will generalize to any new allele that was not observed in our read evidence because $\prod_{i=1}^{s} P(r_i|a) = \prod_{i=1}^{s} q_i$ for any allele a not represented in the read set for the sample. Trivially, our observation count o_i for such novel alleles is also = 0. As such, an allele-centric representation of the sequencing data in which each allele has only an observation count o_j and a product of read qualities $\prod_{i=1}^{o_j} q_i$ can be used to compress our sequencing information without preventing us from later constructing genotype likelihoods that include alleles which we did not refer to in our first pass through the data.