Correction

After my initial submission of this decision letter, I realized I had made a mistake in items 7 and 8 below. I correct this mistake in this opening section and then continue with the decision letter as originally submitted, mistakes and all. The previous version claimed that the genotype likelihood is binomial. But because error (ϵ) varies among sequencing reads, the probability (p) of observing the reference allele also varies, and the likelihood is not binomial. To calculate it without approximations, one would need to sum across all ways of partitioning the C reads among the two alleles of each heterozygous genotype. Avoiding this sum requires approximations, even in the diploid case.

For example, consider the model of Li et al. [2008, sec 1 and Eqns. 9–11 of Supplementary Materials]. Their approach is similar to that of the current manuscript in that it estimates each genotype from sequencing reads at an individual nucleotide site, rather than from several linked sites. It differs in that it deals only with diploids. To avoid summing across partitions, those authors approximate the likelihood of heterozygote genotypes using a binomial formula that ignores sequencing error altogether.

In the manuscript of Sorragi et al, the central problem is a lack of clarity in section 1.2 of Supplementary Materials, both in the text and in the equations. In addition to the points I make below, I would add that we need some discussion of the approximations used to avoid the sum over partitions.

Original decision letter

This manuscript is improved, but I'm not yet prepared to recommend it. The computational results convince me that the math is at least approximately correct and that the method is an improvement over its competitors. I am therefore optimistic about the manuscript. There are however still problems.

In response to the first set of reviews, the authors now emphasize that their method is not designed to detect variation in ploidy along a chromosome. Instead they are interested in variation among chromosomes and among individuals. This raises the question: why use a HMM at all? All three reviewers (including me) asked this question.

The authors do provide a rationale for this decision, but it is buried on lines 214–219 (see below). This rationale should be in the introduction, and it should be given more emphasis, as it justifies the entire approach taken in this paper. In my view, this rationale is still a bit thin. I find it strange that a HMM would be used to model something that doesn't vary along the chromosome.

Second, and more seriously, there are real problems in sections 1.2 and 1.3 of the supplement. The exposition is unclear in these sections, and the math seems to be incorrect. (See below for details.)

My own detailed comments

1. Line 80: I assume that loci are nucleotide sites here. I would make that explicit. Second, the definitions imply that the reads have already been assembled, so that we can associate bases that refer to a single nucleotide site. I would make that explicit too.

2. Line 115: This line defines $|\mathcal{Y}|$, but we don't yet have a definition of $\mathcal{Y}$.

3. Line 120: The notation here seems confused. $O^{(k)}_{m,n}$ represents the sequence reads emitted by the HMM in the kth window. However, the text also says (line 118) that this is a value emitted by the HMM for the kth ploidy. How can k refer both to ploidy and to the index of current window? If this value refers to a window (which includes several loci), then why the subscript n, which refers to a single locus? The same comment applies to $C^{(k)}_{m,n}$.

4. Line 194: The sentence is a bit misleading, because it seems to imply that power would be lower with a sample of 25 than with one of 20. I would say instead that "power increased with sample size up to sample size 20, the largest we considered."

5. Lines 214–219: This passage hints at a rationale for using an HMM, which models variation in ploidy along a chromosome, even where ploidy is constant on each chromosome. This argument belongs in the introduction, and it should be given emphasis. It seems to represent the (otherwise missing) rationale for using a HMM in spite of the fact that the authors don't anticipate that ploidy will vary along the chromosome.

Turning now to the supplementary materials...

6. Section 1.2: This section is of central importance but is difficult to follow. The exposition is unclear, and I think the math is incorrect. One difficulty is that the symbol "p" is used for two different purposes. On the left side of the equations 1 and 2, p is the probability of the entire set of sequencing data at a particular locus (nucleotide site). On the right, it is the probability of the observation at a single sequencing read. I would suggest using "L" on the left side of these equations, since this is the likelihood function.

7. The definition of $p(O_{m,n,r}|G_{m,n}, \epsilon_{m,n,r}, Y_{m,n})$ seems to be incorrect. First, it doesn't make sense to say "if $O_{m,n,r}$ in $G_{m,n}$," because both of these quantities are integers. "In" would make sense only if $G_{m,n}$ were a collection of some sort. (I made this same comment in my previous review.) Apart from that, the definition seems to depend only on sequencing error and not on genotype, which can't be right. I would suggest the following rewrite:

   > Consider a locus (nucleotide site) with ploidy Y, which is covered by C sequencing reads. Let $G \in \{0, 1, \ldots, Y\}$ represent the number of copies of the reference allele in this genotype. At a single read, and in the absence of error, we would observe the reference allele with probability G/Y, assuming that the nucleotide sequenced is chosen at random from among the Y available. Suppose however that errors arise with probability ϵ, and that when an error occurs, the observed nucleotide is equally likely to be any of the other 3 nucleotide states. With this model of error, we observe the reference allele in a single read with probability
   >
   > $$p = (G/Y)(1 - \epsilon) + (1 - G/Y)\epsilon/3$$
   >
   > Here, the first term accounts for the possibility that the true nucleotide is the reference allele and is sequenced without error. The second term accounts for the possibility that the true nucleotide is the alternate allele but is erroneously read as the reference allele.

8. I am also unable to make sense of Eqns. 1 and 2. The observed data consist of C independent observations, each of which is either the reference allele (probability p) or the

alternate allele (probability $1 - p$). This implies that O, the number of copies of the reference allele in the sequencing data, is binomial with parameters C and p. The likelihood is therefore

$$L = \binom{C}{O} p^O (1-p)^{C-O}$$

which is not equivalent to the authors' Eqns. 1 and 2.

9. Eqn. 3, which aims to estimate the population allele frequency, also appears to be incorrect. It sums M allele frequencies, but does not divide this sum by M. Instead, it divides by C_n, which is larger than M. Consequently, the quantity calculated will be much too small. I suspect the authors meant to write

$$\hat{F}_n = \frac{1}{C_n} \sum_{m=1}^{M} C_{m,n} F_{m,n}$$

10. Page 3, paragraph 2: "allows to update" → "allows us to update." Also p. 5, line 1.

11. P. 3, line 13 up: Should "negative binomial" should be "binomial?"

12. P. 4, middle of page: I suggest "feasible" instead of "possible to be implemented."

13. Eqns. 4–6. Minor typographic note: It's conventional in typesetting math to put multiletter functions such as "ln" in Roman type, so that they don't look like "l times n." In LaTeX, the command `\ln` will do this for you.

14. P. 5, just after 1st displayed eqn.: The phrase "Let us equal the partial derivative" is unclear. I'm not sure what was intended here.

References

Heng Li, Jue Ruan, and Richard Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. 18(11):1851–1858, 2008.