This manuscript is promising but still needs work. There are three reviews (including my own), and they are in remarkable agreement. The main issues are as follows. First, the manuscript lacks biological motivation. It is still not clear to me what biological problems the method is designed to address. Second, the organization needs work. It should be possible to read the main text without the supplement, but that isn't possible in the current draft. In addition, much of the material in the main text is duplicated in the supplement, sometimes in inconsistent ways. Third, there are mistakes (possibly only typographical) in the math. Fourth, the algorithm on pp. 4–5 of the supplement is not explained clearly. Finally, Prof. Galtier (in section C of his review) has an interesting suggestion about the way coverage is modelled. I would be interested in your reaction to that suggestion.

None of these seem like insurmountable problems, so I anticipate a positive decision after a round of revision and re-review.

My own detailed comments

1. The manuscript needs more in the way of biological motivation, as noted by both reviewers. Are there organisms with extensive variation in ploidy within individual genomes? There is some variation of this sort in the human Y chromosome, which has a couple of diploid bits but is mostly haploid. Would this method be useful for mammalian sex chromosomes? For cancer cells? How does ploidy vary across the genomes of the pathenogenic yeast that you study?

2. In genome sequence data, duplicated regions are often assembled on top of each other and show up in the data as regions of high coverage. I don't think this is what you mean be variation in ploidy. It would be useful to say this and to explain without math how it is possible to distinguish variation in ploidy from this sort of error in genome assembly.

3. Eqn. 2 is unclear and probably incorrect. Let me begin with a few suggestions designed to improve readability.

   (a) The Phred-scaled error, $q_{m,n,r}$ is a distraction here. It doesn't matter that uncertainties are expressed in Phred scale in the data file. All that matters is the error probability, $\epsilon_{m,n,r}$. I would ditch q and retain ϵ.

   (b) For consistency with the rest of the notation, the inner index of the summation should be y rather than i.

   (c) This equation would be easier to read if you suppressed the m, n subscripts throughout.

   (d) It is also easier to read if you don't write it in log scale.

   With these cosmetic adjustments, the equation becomes

   $$p(O, G, Y) = \prod_{r=1}^{R} \frac{1}{Y} \sum_{y=1}^{Y} p(O_r | G, \epsilon_r, Y)$$

   I turn now to substantive concerns.

   (e) The sum is over y, but there is no y in the summand. What are we summing across? This formulation would make sense if G were a vector with Y entries, each representing one nucleotide within the genotype. In that interpretation G should be G_y. But the text says that G is an integer. If so, then I don't understand why we are averaging over Y values, all of which appear to be identical.

(f) The ambiguity about G also affects the second line of the equation, which defines $p(O_r|G, \epsilon_r, Y)$. That line includes the condition "if O_r in G," which would make sense if G were a vector, but doesn't work if G is an integer. If G is really an integer, then the averaging operation in the first line isn't needed, and the second line might be something like

$$p(O_r|G, \epsilon_r, Y) = \frac{G}{Y}(1 - \epsilon_r) + \epsilon_r/3$$

because G/Y is the frequency of the focal allele within the genotype. If G is really a vector, then perhaps "O_r in G" should be "$O_r = G_y$."

(g) However, neither of these options is quite right either, because they take no account of the fact that we have filtered out sites with more than two alleles. I suspect (but am not sure) that after conditioning on that filter, $\epsilon_r/3$ will become simply ϵ_r.

4. sec 2.3, line 4 up: Shouldn't "sampled without replacement" be "sampled with replacement?" Several reads may refer to a single haploid copy of the locus. If sampling were without replacement, there couldn't be more than two reads per site in a diploid genome. This is not a problem for Eqn. 3, which works fine under sampling with replacement.

5. Fig. 4. Why is "4" an absorbing state? Can't there be a tetraploid segment of chromosome surrounded by diploid segments? Why is the initial state necessarily diploid? Isn't it possible that the chromosome starts with a tetraploid segment?

6. Bottom of p. 5. You should define $\alpha_{Y_m^{(k)}}$ and $\beta_{Y_m^{(k)}}$. It would be sufficient to say that these are the shape and scale parameters of the underlying Gamma distribution.

7. Bottom of p. 5. As this is written, it appears that you must estimate two parameters ($\alpha_{Y_m^{(k)}}$ and $\beta_{Y_m^{(k)}}$) for each segment of the HMM. This parameter count seems excessive. Are you assuming that all segments with the same ploidy have the same values of these parameters? If so, please make this explicit.

8. p. 6, top. It should be possible to follow the text without consulting the supplementary materials. The reference here to Eqn. 6 makes that impossible. You should either say in words what this equation does, or move it from the supp to the text.

9. p. 6, ¶3: I'm lost at $\mathcal{YK}$. We already know about $\mathcal{Y}$, but I don't recall seeing $\mathcal{K}$. You do define K as the number of segments in the HMM. Is this the same as $\mathcal{K}$?

10. p. 6, ¶3: What do you mean by "bounded by $\mathcal{YK}$?" Is this an upper bound on the number of iterations? Or do you mean the algorithm is $O(\mathcal{YK})$? Apparently not: p. 8 says the algorithm is $O(Y^2K)$.

11. p. 8, ¶ 2: You say that power is maximal at coverage 0.5X and 20 individuals. Do you really mean that power is greater at 0.5X than at 30X? Why would it not increase with coverage? Why would it not increase with sample size? Or do you mean that you only considered 0.5X coverage, and power increased with sample size up to 20, the largest sample size you considered?

12. Grammar: At several points, there are constructions such as: "would allow to estimate." This should be "would allow us to estimate," or "would allow one to estimate." You need the noun.

Turning now to the supplementary materials. . .

13. Sect. 6.2. Why is this distribution negative binomial?

14. Sec. 6.3. I think this sentence is incorrect: "The genotype likelihood is the probability of observing a specific genotype given the observed sequencing data." The likelihood is ordinarily defined as the probability of the data interpreted as a function of the parameters.

15. Eqn. 5. I think this is meant to be the same as Eqn. 2, but it isn't. It has r in two places where Eqn. 2 had $O_{m,n,r}$. It's not a good idea to present the same equation both in the text and in the supplement.

16. Sec. 6.4. The problem with "sampling without replacement" occurs here too.

17. Just after Fig. S1. In this section, $\mathcal{Y}$ is the set of ploidies and $|\mathcal{Y}|$ is the number of ploidies. This is inconsistent with the text, in which $\mathcal{Y}$ was the number of ploidies. The notation should be consistent.

18. p. 4, ¶ 2. "Triplet" is the wrong word to use in describing $(\boldsymbol{A}, \boldsymbol{\delta}, \boldsymbol{\beta}, \boldsymbol{\alpha})$. Reading further, I also see "triplets" that consist of two elements. I think the word you want is "tuple."

19. pp. 4–5: I can't make sense of these equations. They need to be much more carefully explained. Can you provide some intuition about what the "intermediate quantity," Q, represents?

20. Eqn. 8. What is meant by $\ln \mathbf{A}$? Is this the element-wise logarithm of the matrix, or the matrix logarithm?