Responses

General comments

This revision is much better. Sections 1.1–1.2 of the supplement, which I criticized in previous reviews, are now beautifully clear.

I'm willing to recommend this manuscript without modification, but I want to raise one issue (mentioned first in my first review), in case the authors would prefer to address it. On p. 1 of the supplement, the authors write

$$p(O_{mnc}|a_y) = \begin{cases} 1 - \epsilon_{mnc} & \text{if } O_{mnc} = a_y \\ \frac{\epsilon_{mnc}}{3} & \text{otherwise} \end{cases}$$

I finally understand what this equation means, and I think it would be correct were it not for the fact that the method excludes nucleotide sites at which the sample contains more than two alleles. After conditioning on this fact, I think the "3" in the denominator will disappear. (If the observed allele is not the true one, there is only one alternative—not three.)

As I said above, I am happy to recommend this paper as it stands. The numerical results demonstrate that the method compares favorably with competitors, especially when coverage is low. This suggests that the issue just raised has no large effect. On the other hand, addressing it might improve the method.

The formula above calculates genotype likelihoods from sequenced nucleotides at each locus and individual. Whilst in downstream analyses we consider only diallelic variation in SNPs, we allow for multiple nucleotides to be present in the sequencing reads and to contribute to the calculation of genotype likelihoods, as per standard practice in the field (McKenna et al. Genome research 20(9), 1297-303 2010). This being said, similarly to other software for NGS data analysis, HMMploidy has a filtering option to remove sequenced nucleotides representing non-major alleles if their proportion is below a threshold chosen by the user. which are not among the two most common ones at each locus. This information is provided in the software manual on github, specifically with options:

- **-m or --min_non_major_freq: Set the minimum frequency of non major alleles for bases to be included in the calculations.**
- **-M2 or --max_minor2_freq: Set the maximum frequency of third most prolific alleles for bases to be included in the calculations.**
- **-M3 or --max_minor3_freq: Set the maximum frequency of fourth most prolific alleles for bases to be included in the calculations.**

Minor comments

- 100: Should state that F is the frequency **of the genotype**.
- page 1 of Supplement: In the definitions of a_i and O_{mnc}, the authors should state what "0" and "1" represent. Is "0" the reference allele and "1" the alternate?

- **F is the population allele frequency, which is estimated as in Equation 1 from the supplementary material. F is used to calculate genotype prior probabilities.**

- **We now add in the text: “Values {0,1} can be considered as the reference and alternate allele, respectively, with other polarisations being possible (e.g., ancestral and derived allele).”**