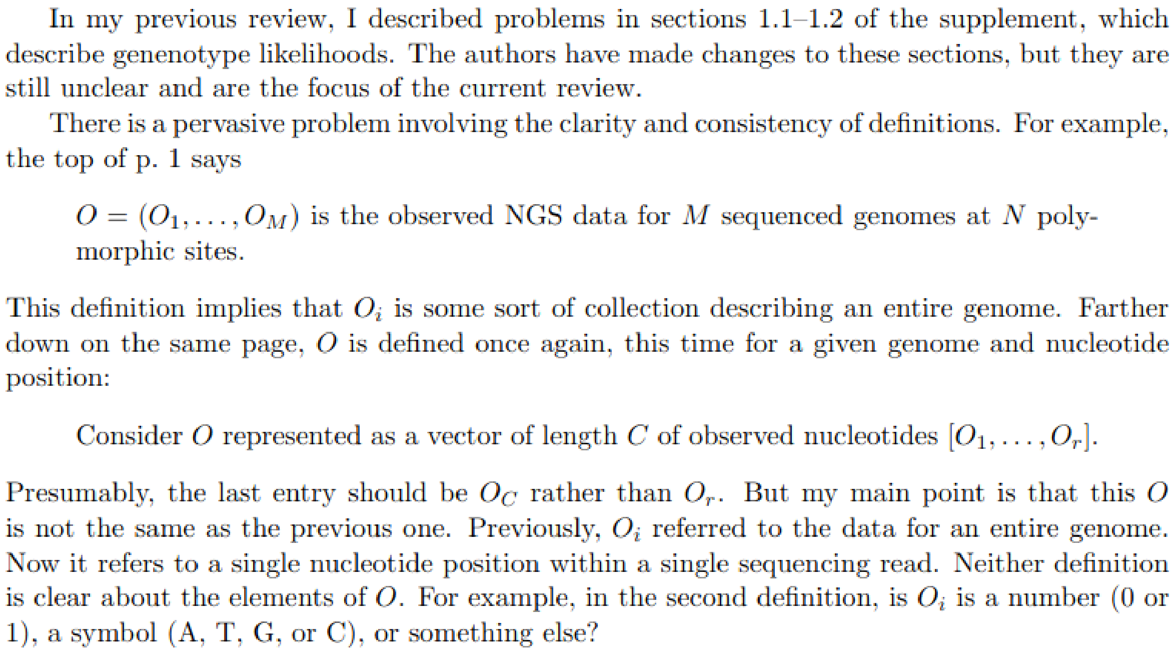
**Responses**



In my previous review, I described problems in sections 1.1–1.2 of the supplement, which describe genenotype likelihoods. The authors have made changes to these sections, but they are still unclear and are the focus of the current review.

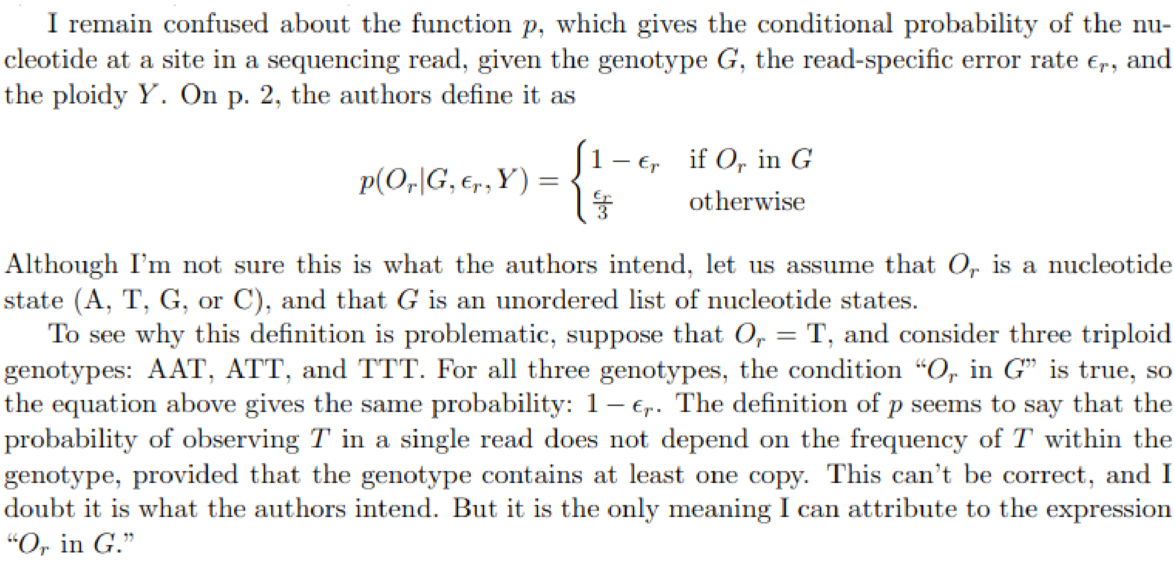
There is a pervasive problem involving the clarity and consistency of definitions. For example, the top of p. 1 says

> $O = (O_1, \ldots, O_M)$ is the observed NGS data for $M$ sequenced genomes at $N$ polymorphic sites.

This definition implies that $O_i$ is some sort of collection describing an entire genome. Farther down on the same page, $O$ is defined once again, this time for a given genome and nucleotide position:

> Consider $O$ represented as a vector of length $C$ of observed nucleotides $[O_1, \ldots, O_r]$.

Presumably, the last entry should be $O_C$ rather than $O_r$. But my main point is that this $O$ is not the same as the previous one. Previously, $O_i$ referred to the data for an entire genome. Now it refers to a single nucleotide position within a single sequencing read. Neither definition is clear about the elements of $O$. For example, in the second definition, is $O_i$ is a number (0 or 1), a symbol (A, T, G, or C), or something else?



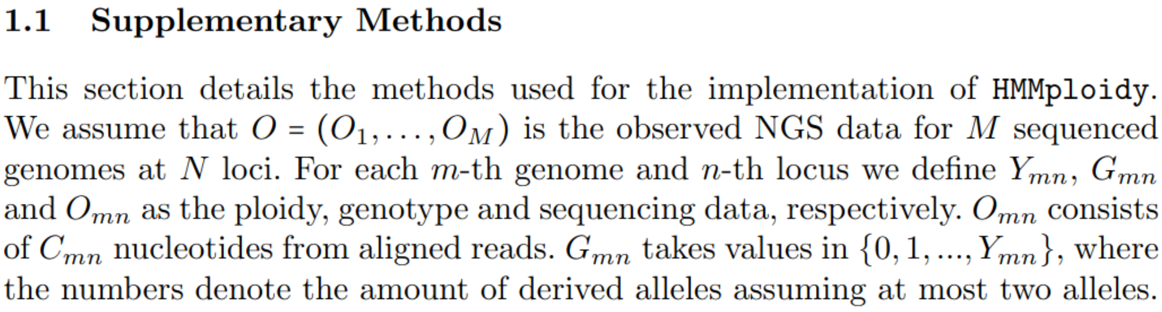
I remain confused about the function $p$, which gives the conditional probability of the nucleotide at a site in a sequencing read, given the genotype $G$, the read-specific error rate $\epsilon_r$, and the ploidy $Y$. On p. 2, the authors define it as

$$p(O_r|G, \epsilon_r, Y) = \begin{cases} 1 - \epsilon_r & \text{if } O_r \text{ in } G \\ \frac{\epsilon_r}{3} & \text{otherwise} \end{cases}$$

Although I'm not sure this is what the authors intend, let us assume that $O_r$ is a nucleotide state (A, T, G, or C), and that $G$ is an unordered list of nucleotide states.

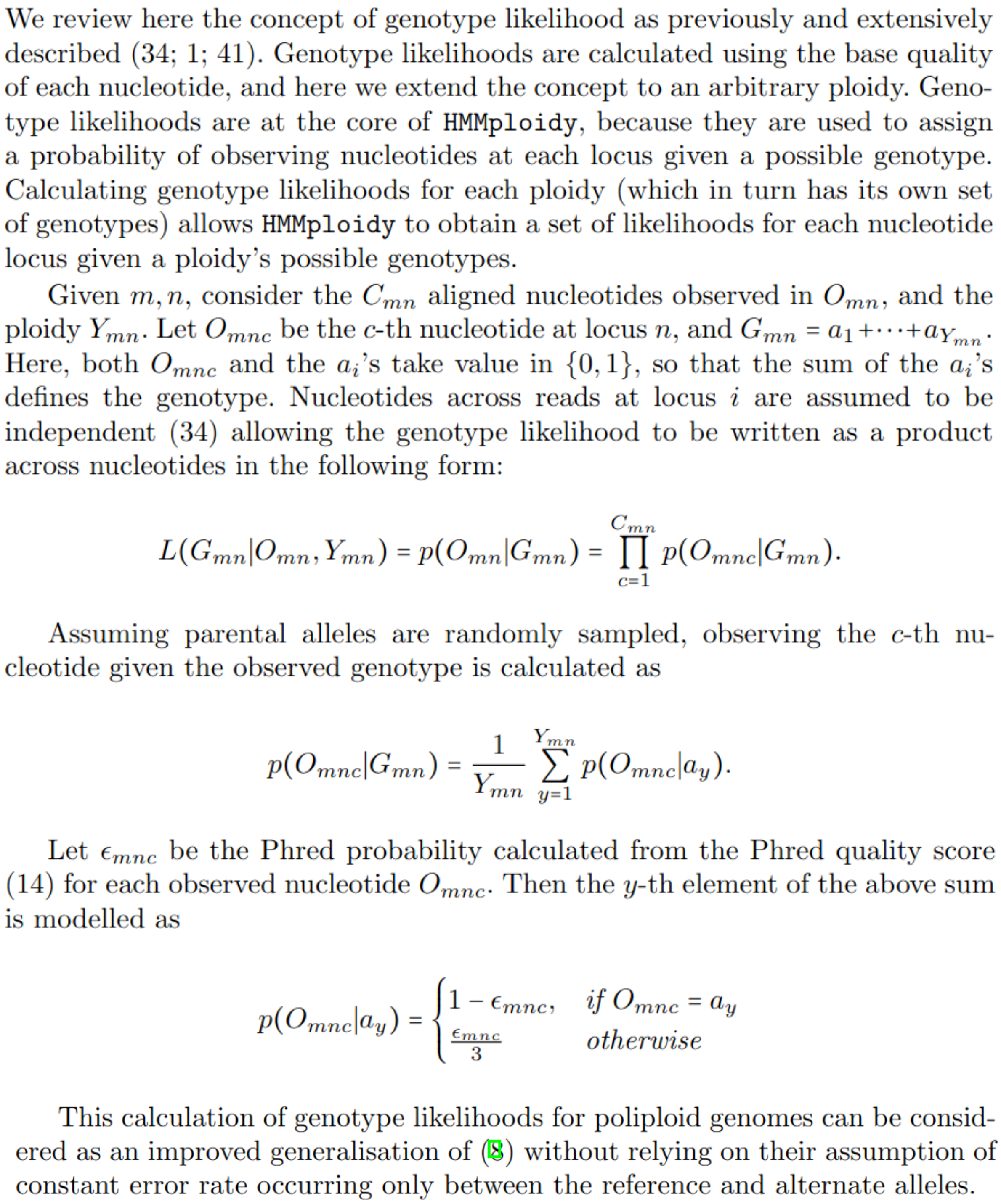
To see why this definition is problematic, suppose that $O_r = \text{T}$, and consider three triploid genotypes: AAT, ATT, and TTT. For all three genotypes, the condition "$O_r$ in $G$" is true, so the equation above gives the same probability: $1 - \epsilon_r$. The definition of $p$ seems to say that the probability of observing $T$ in a single read does not depend on the frequency of $T$ within the genotype, provided that the genotype contains at least one copy. This can't be correct, and I doubt it is what the authors intend. But it is the only meaning I can attribute to the expression "$O_r$ in $G$."

**We fixed the notation and revised sections 1.1 and 1.2 of supplementary material which now read as:**



## 1.1 Supplementary Methods

This section details the methods used for the implementation of `HMMploidy`. We assume that $O = (O_1, \ldots, O_M)$ is the observed NGS data for $M$ sequenced genomes at $N$ loci. For each $m$-th genome and $n$-th locus we define $Y_{mn}$, $G_{mn}$ and $O_{mn}$ as the ploidy, genotype and sequencing data, respectively. $O_{mn}$ consists of $C_{mn}$ nucleotides from aligned reads. $G_{mn}$ takes values in $\{0, 1, \ldots, Y_{mn}\}$, where the numbers denote the amount of derived alleles assuming at most two alleles.



We review here the concept of genotype likelihood as previously and extensively described (34; 1; 41). Genotype likelihoods are calculated using the base quality of each nucleotide, and here we extend the concept to an arbitrary ploidy. Genotype likelihoods are at the core of `HMMploidy`, because they are used to assign a probability of observing nucleotides at each locus given a possible genotype. Calculating genotype likelihoods for each ploidy (which in turn has its own set of genotypes) allows `HMMploidy` to obtain a set of likelihoods for each nucleotide locus given a ploidy's possible genotypes.

Given $m, n$, consider the $C_{mn}$ aligned nucleotides observed in $O_{mn}$, and the ploidy $Y_{mn}$. Let $O_{mnc}$ be the $c$-th nucleotide at locus $n$, and $G_{mn} = a_1 + \cdots + a_{Y_{mn}}$. Here, both $O_{mnc}$ and the $a_i$'s take value in $\{0, 1\}$, so that the sum of the $a_i$'s defines the genotype. Nucleotides across reads at locus $i$ are assumed to be independent (34) allowing the genotype likelihood to be written as a product across nucleotides in the following form:

$$L(G_{mn}|O_{mn}, Y_{mn}) = p(O_{mn}|G_{mn}) = \prod_{c=1}^{C_{mn}} p(O_{mnc}|G_{mn}).$$

Assuming parental alleles are randomly sampled, observing the $c$-th nucleotide given the observed genotype is calculated as

$$p(O_{mnc}|G_{mn}) = \frac{1}{Y_{mn}} \sum_{y=1}^{Y_{mn}} p(O_{mnc}|a_y).$$

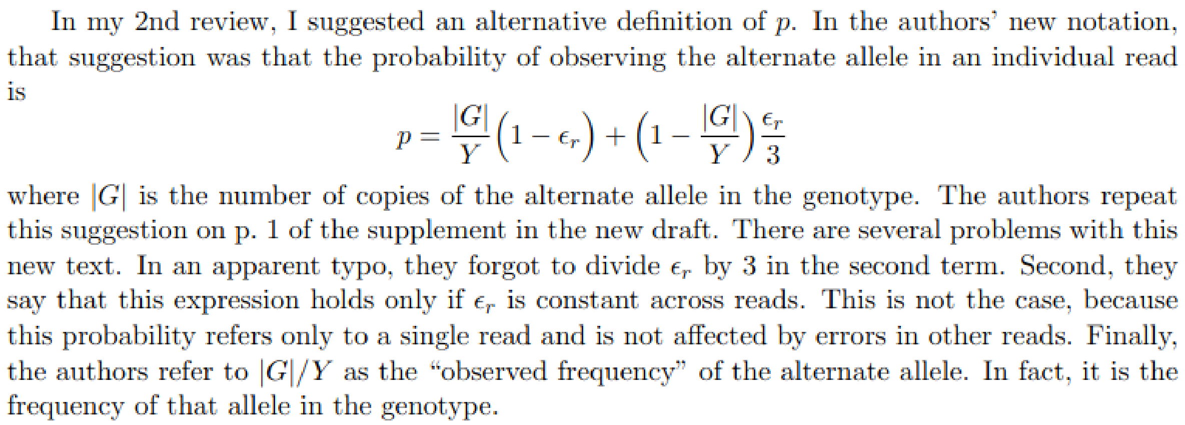
Let $\epsilon_{mnc}$ be the Phred probability calculated from the Phred quality score (14) for each observed nucleotide $O_{mnc}$. Then the $y$-th element of the above sum is modelled as

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

This calculation of genotype likelihoods for poliploid genomes can be considered as an improved generalisation of (8) without relying on their assumption of constant error rate occurring only between the reference and alternate alleles.

**We'd like to stress that this formulation of genotype likelihoods has been previously introduced by several studies, as cited accordingly. Here we extend it to the case of arbitrary ploidy. We believe that the notation is now correct and is equivalent to previous formulations. We also add a reference to a previous study (8) which proposed an alternative (albeit with more assumptions) formulation of genotype likelihoods in polyploid genomes.**

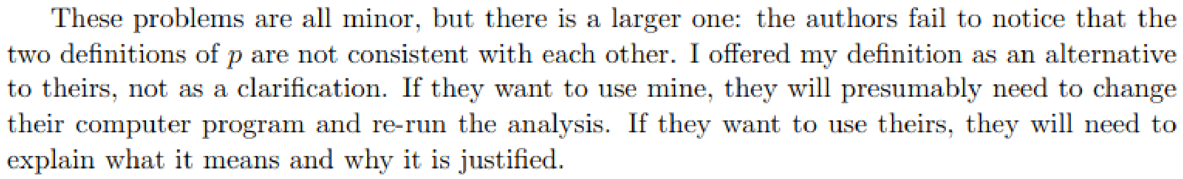
—-------------------------------------------------------------------------------------



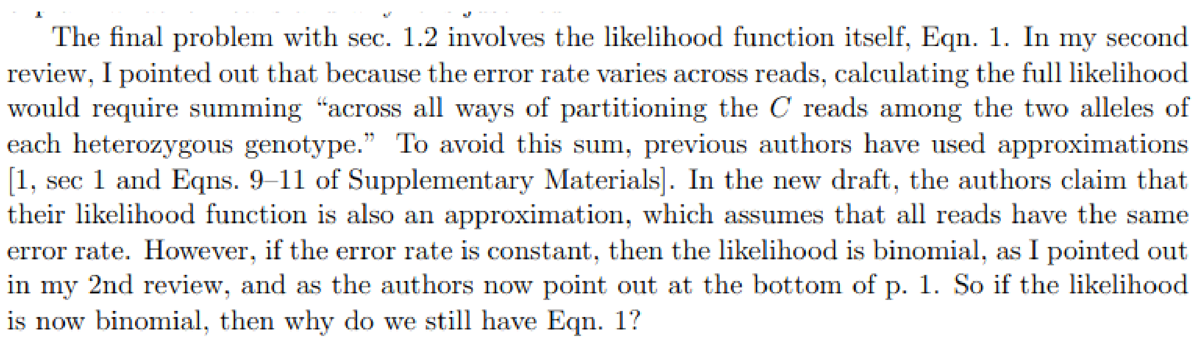
In my 2nd review, I suggested an alternative definition of $p$. In the authors' new notation, that suggestion was that the probability of observing the alternate allele in an individual read is

$$p = \frac{|G|}{Y}\left(1 - \epsilon_r\right) + \left(1 - \frac{|G|}{Y}\right)\frac{\epsilon_r}{3}$$

where $|G|$ is the number of copies of the alternate allele in the genotype. The authors repeat this suggestion on p. 1 of the supplement in the new draft. There are several problems with this new text. In an apparent typo, they forgot to divide $\epsilon_r$ by 3 in the second term. Second, they say that this expression holds only if $\epsilon_r$ is constant across reads. This is not the case, because this probability refers only to a single read and is not affected by errors in other reads. Finally, the authors refer to $|G|/Y$ as the "observed frequency" of the alternate allele. In fact, it is the frequency of that allele in the genotype.



These problems are all minor, but there is a larger one: the authors fail to notice that the two definitions of $p$ are not consistent with each other. I offered my definition as an alternative to theirs, not as a clarification. If they want to use mine, they will presumably need to change their computer program and re-run the analysis. If they want to use theirs, they will need to explain what it means and why it is justified.



The final problem with sec. 1.2 involves the likelihood function itself, Eqn. 1. In my second review, I pointed out that because the error rate varies across reads, calculating the full likelihood would require summing "across all ways of partitioning the $C$ reads among the two alleles of each heterozygous genotype." To avoid this sum, previous authors have used approximations [1, sec 1 and Eqns. 9–11 of Supplementary Materials]. In the new draft, the authors claim that their likelihood function is also an approximation, which assumes that all reads have the same error rate. However, if the error rate is constant, then the likelihood is binomial, as I pointed out in my 2nd review, and as the authors now point out at the bottom of p. 1. So if the likelihood is now binomial, then why do we still have Eqn. 1?

**We did not use this alternative definition in our method. Since it is not relevant to our study, we removed this section from the new version. To clarify, we do not assume the error to be constant in the calculation of genotype likelihoods as described in the previous reply.**