Review by Alan Rogers (recommender)

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.

Thank you for providing this amendment to the original decision letter. We rephrased our definition of genotype likelihoods, as described in our replies.

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.

HMMploidy has been primarily designed to detect ploidy variation among chromosomes and individuals. However, copy number variants and structural variations affect genomic patterns within chromosomes. Additionally, nuances of low-coverage short-read sequencing data may leave further local patterns in the data. Therefore, we implemented a HMM feature in our model for two main reasons.

Firstly, with a HMM, we are able to obtain a distribution of ploidy tracts along a chromosome and therefore provide further statistical support for each tested ploidy. In other words, we can infer how much of the sequencing data on said chromosome support the most likely ploidy.
Secondly, with a HMM, we are able to identify local regions where the predicted ploidy deviates from the whole-chromosome estimate. Said regions can be then further investigated, for instance as potential locations of CNVs. Generally, with a HMM, we are able to identify large segmental losses, polyplodisations, and rearrangements, as well as other genomic features (i.e. pseudo-automsomal regions in sex chromosomes) as a byproduct of HMMploidy's predictions.
Furthermore, It is worth mentioning that, unlike other methods, HMMploidy allows to differentiate between local changes in sequencing depth levels (i.e. due to the presence of CNVs) and genotype frequencies (i.e. due to selection or inbreeding) by symply disabling or enabling the use of genotype likelihoods.
Finally, the statistical framework in HMMploidy can be adopted to calculate ploidy likelihoods to obtain chromosome-wide estimates, as suggested by the recommender and reviewers.

We now explicitly mention these justifications in the introduction:
"HMMploidy infers ploidy variation in sliding windows among chromosomes and among individuals. While ploidy is not expected to vary within each chromosome, the distribution of inferred ploidy tracts provides further statistical support to whole-chromosome estimates.
Additionally, HMMploidy can identify local regions with aberrant predicted ploidy to be further investigated, for instance as potential locations of copy number variants or structural rearrangements.
Finally, any detected within-chromosome ploidy variation can serve as a diagnostic tool to investigate possible mapping or assembly errors.
Notably, by training separate HMMs, HMMploidy can effectively infer aneuploidy among chromosomes and samples."

To suggest further applications of of HMMploidy, we now add in the discussion:
"We predict that HMMploidy will have a broad applicability in studies of genome evolution beyond the scenarios illustrated in this study. For instance, the statistical framework in HMMploidy can be adopted to infer aneuploidy in cancerous cells (6), or partial changes of copy numbers in polyploid genomes due to deletions or duplications (52)."

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.)

We assume that this comment now refers only to the unclear exposition, which we addressed as detailed below.

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. The current wording initially led me to think that O_m referred to the set of raw reads for a given genome.

We now state that "We define a locus as a nucleotide site. We assume that sequencing reads are mapped and aligned so that bases can be assigned to a single nucleotide site."

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

We now specify that "..... Y is the set of ploidy levels included in the model and |Y| is the number of ploidy levels (i.e. cardinality of 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 $C^{(k)}_{m,n}$.

Thanks for noticing this. "K-th ploidy" is indeed wrong, as well as the "n" index for C^(k)_m,n and O^(k)_m,n, since observations and average depth are inside a window. The correct notation is indeed the one shown in Figure S1. We probably created this error when we had a workaround of the notation earlier during the writing of the manuscript. We now write:
".... each of the |Y| ploidies emits two observations. Those contain a dependency on which ploidy is assigned to that window. The observations consist of the sequenced reads O^{(k)}_m and the average sequencing depth C^{(k)}_m in the k-th window …."

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."

We agree and now state that "... HMMploidy's power increased with sample size up to 20 - the largest we considered - in all scenarios excluding the tetraploid case …", as suggested.

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.

As explained above, we edited and moved this paragraph in the introduction which now reads "HMMploidy infers ploidy variation in sliding windows among chromosomes and among individuals. While ploidy is not expected to vary within each chromosome, the distribution of inferred ploidy tracts provides further statistical support to whole-chromosome estimates.
Additionally, HMMploidy can identify local regions with aberrant predicted ploidy to be further investigated, for instance as potential locations of copy number variants or structural rearrangements.
Finally, any detected within-chromosome ploidy variation can serve as a diagnostic tool to investigate possible mapping or assembly errors."

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.

Thank you for these very detailed comments and corrections in the review. We agree that the mentioned parts needed a rewriting for better understanding and ease of reading. We reworked the whole genotype likelihood explanation, in which a mention of the logic behind varying Phred errors and the Li et al (2008) framework used in SAMtools are also included. We simplified the notation making the m,n indices implicit, so that formulae became shorter and readable. Also, we defined the genotypes as collections of nucleotides, so that it fits in the formulations we need. There was an error in the genotype likelihood formula we had before, and we corrected it. Analogous corrections to simplify the notation have been done in the first part of the main paper, where we illustrate briefly the probability of the data. Now the supplementary part reads as follows:

1.2 Genotype likelihood for arbitrary ploidy number

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. We will calculate genotype likelihoods using the base quality of each nucleotide. Bases across reads are assumed to be independent, so that each base quality can be treated as the probability of incorrectly sequenced nucleotide across reads (35).

For ease of visualisation, we will consider a diallelic locus n in a genome m but suppress the two indices, because the formula of the genotype likelihoods depends on r and the possible genotypes. Given m, n, consider the observed sequencing data O, the coverage C and the ploidy Y at such genome and locus. Consider O represented as a vector of length C of observed nucleotides $[O_1, \ldots, O_r]$. Let ϵ_r be the Phred probability calculated from the Phred quality score (12) for each observed nucleotide O_r.

If each ϵ_r was constant, then we would be able to observe the alternate alleles with probability

$$p(O) = \frac{|G|}{Y}\Big(1 - \epsilon_r\Big) + \Big(1 - \frac{|G|}{Y}\Big)\epsilon_r,$$

for a given genotype and ploidy, where $\frac{|G|}{Y}$ represents the observed frequency of the alternate alleles. The equation above would produce the following genotype likelihood:

$$L(G|O, Y) = \binom{C}{|G|} p(O)^{|G|}\big(1 - p(O)\big)^{C-|G|}.$$

However, ϵ_r varies at each nucleotide. This means that the likelihood $L(G|O, Y)$ is no longer binomial. The analytical procedure to calculate the likelihood implies calculating all possible error-dependent assignments of nucleotides to a genotype, for every genotype at each ploidy. This requires a large amount of

combinatorics and calculations at each locus, and therefore approximation is necessary to tackle this problem.

The approximation used in our software is an extension of the diploid GATK model (35) and mostly resemble the approach of (27), where the idea of the authors is to estimate genotypes at each nucleotide site without considering linked loci, and to set the error ϵ_r to an uniform value $\bar{\epsilon}$. Further considerations lead to an approximation of the genotype likelihood that ignores the Phred error. Such method is essentially what is implemented in `SAMtools` (29). In our case, the Phred error is still included in the model and varies across reads in a nucleotide locus. Our assumption leads to the following genotype likelihood:

$$L(G|O,Y) = \prod_{r=1}^{C} \frac{1}{Y+1} p(O_r|G,\epsilon_r,Y), \tag{1}$$

where $p(O_r|G,\epsilon_r,Y)$ is defined analogously as in the case of costant Phred error:

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

---

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

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

This has now been corrected in the text.

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

We corrected all instances of this typo.

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

We now clarify that "The average depths are modelled with a negative binomial distribution to take data overdispersion into account (Choudhary et al. 2022)."

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

We changed the text according to the reviewer's suggestion.

13. Eqns. 4–6. Minor typographic note: It's conventional in typesetting math to put multi-letter 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.

We changed all the occurrences of ln into \ln in the text.

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.

In this part of the text, by setting the partial derivative to zero and solving for the parameter of the derivative, we can find the parameter's optimum. We can see it is quite unclear from how the text is formulated. Now we rewrite this part of the supplementary as (equations and math text skipped with dots)
"By setting the partial derivative of $…..$ w.r.t. a certain $……..$ equal to zero as below, we will be able to calculate the optimum for the derivative's parameter
……
Solving for …….. leads to the optimum of the parameter:
……. "

Review by Nicolas Galtier

I found the manuscript to be substantially improved in many respects, and would like to thank the authors for the hard work and willingness to address all the reviewers' remarks. I still have a couple of questions.

1. From the authors' response and corrections, it is my understanding that the HMMploidy method is intended to be applied to segments across which ploidy does not vary. This is perceptible from the modified introduction, in which the emphasis is put on aneuploidy (i.e., single-ploidy chromosomes), and the simulation part, in which constant ploidy is assumed. This is a perfectly valid goal, but one might then ask, why taking an HMM approach? If ploidy is assumed to be constant then the likelihood can probably be calculated based on the provided equations without the HMM layer. The authors might like to clarify the choice of an HMM approach if ploidy is supposed not to change across the analyzed segments.

As explained to the recommender, HMMploidy has been primarily designed to detect ploidy variation among chromosomes and individuals. However, copy number variants and structural variations affect genomic patterns within chromosomes. Additionally, nuances of low-coverage short-read sequencing data may leave further local patterns in the data. Therefore, we implemented a HMM feature in our model for two main reasons.
Firstly, with a HMM, we are able to obtain a distribution of ploidy tracts along a chromosome and therefore provide further statistical support for each tested ploidy. In other words, we can infer how much of the sequencing data on said chromosome support the most likely ploidy.
Secondly, with a HMM, we are able to identify local regions where the predicted ploidy deviates from the whole-chromosome estimate. Said regions can be then further investigated, for instance as potential locations of CNVs. Generally, with a HMM, we are able to identify large segmental losses, polyplodisations, and rearrangements, as well as other genomic features (i.e. pseudo-automsomal regions in sex chromosomes) as a byproduct of HMMploidy's predictions.
Furthermore, It is worth mentioning that, unlike other methods, HMMploidy allows to differentiate between local changes in sequencing depth levels (i.e. due to the presence of CNVs) and genotype frequencies (i.e. due to selection or inbreeding) by symply disabling or enabling the use of genotype likelihoods.
Finally, the statistical framework in HMMploidy can be adopted to calculate ploidy likelihoods to obtain chromosome-wide estimates, as suggested by the recommender and reviewers.

We now explicitly mention these justifications in the introduction:

“HMMploidy infers ploidy variation in sliding windows among chromosomes and among individuals. While ploidy is not expected to vary within each chromosome, the distribution of inferred ploidy tracts provides further statistical support to whole-chromosome estimates.
Additionally, HMMploidy can identify local regions with aberrant predicted ploidy to be further investigated, for instance as potential locations of copy number variants or structural rearrangements.
Finally, any detected within-chromosome ploidy variation can serve as a diagnostic tool to investigate possible mapping or assembly errors.
Notably, by training separate HMMs, HMMploidy can effectively infer aneuploidy among chromosomes and samples.“

To suggest further applications of of HMMploidy, we now add in the discussion:
“We predict that HMMploidy will have a broad applicability in studies of genome evolution beyond the scenarios illustrated in this study. For instance, the statistical framework in HMMploidy can be adopted to infer aneuploidy in cancerous cells (6), or partial changes of copy numbers in polyploid genomes due to deletions or duplications (52).”

2. The section on empirical analysis is still a bit unclear to me. In particular:
- do we have external knowledge on the real level of (aneu)ploidy in these samples?

The empirical analysis presented in this study aims at recapitulating previous findings on aneuploidy in C. neoformans. As described in the introduction, ploidy variation is an adaptive mechanism in Cryptococcus neoformans in response to a harsh environment and drug pressure. Aneuploidy-driven heteroresistance to antifungal drug fluconazole is emerging and, therefore, the inference of ploidy is this system is important in both evolutionary and clinical studies.
The data herein used was generated in a previous study (Rhodes et al. 2017 https://doi.org/10.1534/g3.116.037499), and authors inferred aneuploidy (see Figure 2) by investigating the variation of normalized coverage. Specifically, authors write “*In order to determine aneuploidy, whole-genome coverage data were normalized and regions displaying normalized coverage equal to two were deemed diploid events (likewise, normalized coverage equal to three was deemed a triploid event, and so on), whereas normalized coverage equal to zero was deemed a deletion event.*”
In Figure 2 of Rhodes et al. 2017, authors show the patterns of inferred ploidy variation, and concluded that aneuploidy events were observed in at least seven genome pairs of isolates, especially on chromosome 12, in line with previous reports (Omerod et al. 2013 https://pubmed.ncbi.nlm.nih.gov/23550133/). With this approach, Authors also

identified several instances of copy number variation (CNV) in genes known to be involved in drug resistance and virulence.

Our aim was to recapitulate their findings on extensive aneuploidy in C. neoformans using HMMploidy. Importantly, we sought to infer ploidy on both the original high-coverage and a downsampled data set, and assessed any difference.

We agree that these previous findings and our goal were not clear. In the Results section we now write:
"We used HMMploidy to infer ploidy variation in 23 isolates of Cryptococcus neoformans recovered from HIV-infected patients (44). In the original study (44), by analysing variation in normalised sequencing coverage, Rhodes and coworkers identified extensive instances of aneuploidy, especially on chromosome 12, in accordance with previous findings using karyotypic analysis (40). We sought to replicate these inferences using HMMploidy and assessed its performance on a downsampled data set to mirror data uncertainty."

- I don't quite understand the interpretation of the CCTP27 vs CCTP27-d121 discrepancy. In this genome the sequencing depth of chromosome 12 was tripled at day 121, compared to the reference at day 0, suggesting some major biological event. HMMploidy infers the same ploidy (of 1) for chromosome 12 in the two samples, thus missing this biological event as far as I understand it. Still, this is interpreted as a success of the method.

Regarding Figure 3, the average depth for CCTP27-d121 almost tripled in (most of) chromosome 12 compared to chromosome 1. Despite this, HMMploidy (which also uses information on genotypes) infers no change in haploidy. We interpret this result as a large CNV on chromosome 12 for CCTP27-d121 rather than a recombining triploid chromosome. We see this as a further advantage (rather than a "success") of using a method that is not solely based on coverage variation.
We now rewrite the related paragraph as "We interpret this pattern as one CNV instance spanning most of chromosome 12 for CCTP27-d121. In fact, despite the increase in depth, the data is modelled as a haploid chromosome by the genotype likelihoods. This further illustrates the importance of jointly using information on genotypes and depth variation to characterise aneuploidy and CNV events."

The authors might like to clarify their specific goals with this analysis, and what kind of biological pattern or structure they are targeting. If the idea is to identify polyploid segments having accumulated a certain amount of sequence variation, as seems

implicit in the empirical analysis section, then this should probably be stated more explicitly and discussed.

We agree that the rationale for this analysis was not very clear. We believe that with the changes made in the responses above, both our aim and the biological reasons for aneuploidy in C. neoformans are evident now. Please also note that *"Notably, we were able to retrieve the same patterns of predicted ploidy variation when artificially down-sampling the sequencing data to 20% of the original data set"*, unlike competing tested methods, suggesting that our inferences are robust to lower quality data.

Review by Barbara Holland

This paper is an interesting application of a Hidden Markov Model to both inferring ploidy level and detecting changes in ploidy level. The authors make a convincing case for why this is an interesting problem with potential applications in both agriculture and medicine. The new method appears to be more accurate at inferring ploidy levels than existing alternatives particularly at low sequencing coverage.

The issue of whether the method is good at detecting changes in ploidy level does not appear to be explored. My understanding of the model is that the HMM part of the model is used to model the changes in ploidy level. Perhaps I am missing something obvious, but the authors don't seem to exploit this feature in their simulations (i.e. they all have constant ploidy level).

As explained to the recommender and the previous reviewer, HMMploidy has been primarily designed to detect ploidy variation among chromosomes and individuals. However, copy number variants and structural variations affect genomic patterns within chromosomes. Additionally, nuances of low-coverage short-read sequencing data may leave local patterns in the data. Therefore, we implemented a HMM feature in our model for two main reasons.
Firstly, with a HMM, we are able to obtain a distribution of ploidy tracts along a chromosome and therefore provide further statistical support for each tested ploidy. In other words, we can infer how much of the sequencing data on said chromosome support the most likely ploidy.
Secondly, with a HMM, we are able to identify local regions where the predicted ploidy deviates from the whole-chromosome estimate. Said regions can be then further investigated, for instance as potential locations of CNVs. Generally, with a HMM, we are able to identify large segmental losses, polyplodisations, and rearrangements, as well as other genomic features (i.e. pseudo-automsomal regions in sex chromosomes) as a byproduct of HMMploidy's predictions.
Furthermore, It is worth mentioning that, unlike other methods, HMMploidy allows us to differentiate between local changes in sequencing depth levels (i.e. due to the presence of CNVs) and genotype frequencies (i.e. due to selection or inbreeding) by symply disabling or enabling the use of genotype likelihoods.
Finally, the statistical framework in HMMploidy can be adopted to calculate ploidy likelihoods to obtain chromosome-wide estimates, as suggested by the recommender and reviewers.

We now explicitly mention these justifications in the introduction:
"HMMploidy infers ploidy variation in sliding windows among chromosomes and among individuals. While ploidy is not expected to vary within each

chromosome, the distribution of inferred ploidy tracts provides further statistical support to whole-chromosome estimates.
Additionally, HMMploidy can identify local regions with aberrant predicted ploidy to be further investigated, for instance as potential locations of copy number variants or structural rearrangements.
Finally, any detected within-chromosome ploidy variation can serve as a diagnostic tool to investigate possible mapping or assembly errors.“

So my guess is that the superior performance of the method compared to the existing approaches is mostly coming from having a better model for the genotype likelihood and the error process rather than it being an HMM. With this in mind it would be good to move the discussion of this part of the model into the main text rather than the Supplementary Material.

The better performance of HMMploidy is indeed given by the inclusion of genotype likelihoods, information that is typically not considered by other methods.

Minor points
Keywords: check spelling of poliplody

Fixed.

Line 37 ‘the evolution’ -> ‘evolution’

The sentence is “...inferring the ploidy of a sample from genomic data, like in the case of Cryptococcus neoformans, is essential to shed light onto the evolution and adaptation across the domains of life.” We believe that the use of “the” is correct but we are fine removing it if the Recommender wishes so.

The paragraph at the top of page 3 has a few typos/grammatical issues. E.g ‘reference data at known ploidy set...’,
incorporate -> incorporates

Fixed.

Line 74, by diallelic do you mean that you only see at most two states at a particular site across the sample of genomes under consideration (e.g A/G or C/T) regardless of how many copies of the site there are? Or is diallelic with respect to a sequencing read, i.e. there are only two variants of a read?

By diallelic we mean that we observed at most two states at a particular genotype regardless of the number of copies. For instance, with alleles A and

G, triploid genotypes are AAA AAG AGG GGG. Therefore, we do not consider multiallelic variation. We now write *"In what follows, data is assumed to be diallelic (i.e. we observed at most two states at a particular genotype regardless of the number of copies), without loss of generality."* to make it clearer.

Line 83. Is O_{m,n} a sequencing read or just a single site?

We mean a single nucleotide site n for a genome m. We added this term when defining the variable, so that it is clearer.

line 92. Is the population frequency F_n assumed to be known or is it also something that needs to be estimated? How is this done?

F_n is estimated as in section 1.3 of the supplementary material, as shown below.

1.3 Estimation of population frequencies

Population allele frequencies are calculated prior to the HMM optimisation to decrease the computational time. Specifically, the population frequency F_n at the n-th locus is estimated under the assumption of ploidy level being arbitrarily very high to let frequencies represent any possible genotype.

Let $\hat{F}_{m,n}$ be the observed minor allele frequency for sample m at locus n. The population frequency estimator for F_n, say $\hat{F}_n$, is defined as

$$\hat{F}_n = \frac{1}{C_n} \sum_{m=1}^{M} C_{m,n} \hat{F}_{m,n}, \tag{2}$$

where $C_n = \sum_{m=1}^{M} C_{m,n}$.

Line 92 It seems odd to have the genotype likelihood relegated to the supplementary material when it is a very important component of the model. It isn't a long section, so I'd suggest moving it to the main text.

We agree that the genotype likelihoods are an important part of the method. However, during previous rounds of review, we were asked to move this part to the supplementary material (alongside other mathematical details). We feel that the current structure is a good balance between narrative and technical details. Furthermore, the genotype likelihood model presented herein is a simple extension of the GATK model to polyploidy. Therefore, we don't feel the

need to move this section to the main text but we are happy to change it depending on the recommender's opinion.

line 128 how are the alpha and beta parameters for the Poisson Gamma distribution selected/estimated?

Estimation of the parameters is done using the maximization part of the heuristic EM algorithm. We explain those steps in the supplementary material, where we show the equations for finding an optimal alpha, that in turn will be used to calculate an optimal value for beta. It is not much different from optimizing other distribution's parameters, e.g. gaussian, with the difference that the maximization equations are more complex to solve and require some form of approximation and a few controls to avoid numerical problems.

Line 155 I find the description of the simulation a bit confusing, you say that ploidy chosen from 1 to 5 is constant along the genome. I thought the point of the HMM was to be able to detect changes in ploidy level?

In order to compare the performance HMMploidy with existing methods, we simulated genomes with constant ploidy. How HMMploidy can infer ploidy variation can be appreciated with the application on the real data set, and specifically with the robust estimation after artificially downsampling the data.

Line 245 "allows to overcome" -> "allows the method to overcome"

Fixed.

Line 248, I don't understand the sentence starting "On the former point..."

The paragraph now reads as:
"However, training a separate HMM on each genome allows the method to overcome two main issues: samples sequenced at different coverage, and ploidy varying among samples.
When samples are sequenced at different coverage, it is common practice to standardise the sequencing depth across all genomes. However, this would make the estimation of the distributions of standardised counts difficult, especially in samples with noise, errors, and limited coverage."

Line 256 missing full stop

Fixed.

Line 264 tuntimes -> runtimes
Fixed.