Response to reviewers for "HMMploidy: inference of ploidy levels from short-read sequencing data"

Recommender

> 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.

Dear Recommender,

Many thanks for collecting useful comments on our manuscript. Many thanks for the patience while waiting for our responses. We addressed all points raised by the reviewers.

Specifically, in the revised submission, we now:

1) **stress the biological importance of our method and provide more details on the application to aneuploid genomes of C. neoformans;**

2) **reorganise the text to clarify the connection between parts. In particular, we moved all mathematical details in the supplementary material to avoid redundancies. We also modified the main text accordingly to ensure all (but only the essential) terms were introduced and explained. We believe that the main text is much easier to follow;**

3) **fixed any typographical mistakes and simplified the maths;**

4) **explain all algorithms as requested;**

5) **discuss any further changes in the modelling as suggested.**

Please find point-to-point responses to all comments below.

We hope that the new version is now suitable to be recommended at PCI Math Comp Biol and published at Peer Community Journal.

Kind Regards,

Samuele Soraggi and Matteo Fumagalli on behalf of all authors.

Reviewer

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?

We further stressed the biological motivation in the introduction and added more references. In this study, we focus on applications to aneuploid species, not on sex chromosomes or cancer cells (the latter would require a different modelling).

We also expanded the introduction on NGS data analysis.

The introduction now states that:

In recent years, advances in Next-generation sequencing (NGS) technologies allowed for the generation of large amount of genomic data (33; 24). Many statistical and computational methods, and accompanying software, to process NGS data for genotype and SNP calling have been proposed (26; 20; 3). Additionally, dedicated software have been developed to analyse low-coverage sequencing data (35; 16), a popular and cost-effective approach in population genomic studies (29). However, most of these efforts have been focused towards model species with known genomic information. In particular, there has been a lack of research into modelling sequencing data from non-diploid species or organisms with unknown ploidy.

Polyploidy is typically defined as the phenomenon whereby the chromosome set is multiplied, resulting the organism to have three or more sets of chromosomes (36). Polyploidy is common to many organisms at different genic and cellular levels, and it can be the consequence of hybridisation or whole genome duplication (14). For instance, polyploidy plays a significant role in the evolution and speciation of plants (41), as 34.5% of vascular plants (including leading commercial crop species) are shown to be polyploid (48).

Of particular interest is the case of aneuploidy, whereby chromosomal aberrations cause the number of chromosomal copies to vary within populations and individuals. Ploidy variation can be associated with a response or adaptation to environmental factors (9), and it is a phenomenon commonly detected in cancer cells (10) and several pathogenic fungi (i.e. *Cryptococcus neoformans*, *Candida albicans* and *Candida glabrata*) and monocellular parasites (43; 34; 12; 50; 49; 15).

Among aneuploid species, *Cryptococcus neoformans* is a fungal pathogen capable of causing meningitis in immunocompromised individuals, particularly HIV/AIDS patients. Ploidy variation, via aneuploidy and polyploidy, is an adaptive mechanism in *Cryptococcus neoformans* capable of generating variation within the host in response to a harsh environment and drug pressure (34). Aneuploidy-driven heteroresistance to the frontline antifungal drug fluconazole has been described (43), resulting in treatment failure in patients. Within fluconazole resistant colonies, aneuploidy was common, particularly disomy of chromosome 1 which harbours the gene encoding the main drug target of fluconazole, *ERG11* (43). For these reasons, 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.

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.

As pointed out, HMMploidy was not designed to identify duplicated regions which result in a local increase in coverage. The reason is that, unlike other methods, HMMploidy uses the information on genotype likelihoods too, and therefore it is less sensitive to local variation in sequencing depth.

This behaviour is illustrated when presenting results from the application on C. neoformans:
"Interestingly, samples CCTP27 and CCTP27 at day 121 (CCTP27-d121) are inferred to have the same ploidy, even though CCTP27-d121 triplicates its sequencing depth on chromosome 12 (Fig. 3). "

We argue that the observation of ploidy inference being inconsistent with the local variation in depth can be used to infer on gene duplications / CNVs, as:
"This finding suggests the presence of a recent copy number variation. In fact, as no sufficient genetic variation has built up on the recently duplicated triploid chromosome yet, the data is modeled as a single chromosome by the genotype likelihoods."

We acknowledge some confusion on the scope of inferring within-chromosome ploidy variation. The distribution of inferred ploidy tracts returned by HMMploidy can serve as (i) confidence on whole-chromosome ploidy assignment and (ii) detection of local

data aberrations. We removed misleading references to HMMploidy used to infer within-chromosome ploidy variation in the abstract and text.
We also added:
"While ploidy is not expected to vary within each chromosome, the distribution of local ploidy tracts as inferred by HMMploidy can provide statistical support to whole-chromosome estimates. Additionally, any detected within-chromosome ploidy variation can serve as a diagnostic tool to investigate possible mapping or assembly errors."

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.

As suggested, we replaced q with epsilon. We also rewrote the equation to make it clear that the inner summation is over all possible genotypes given a certain ploidy.

We also acknowledge that the previous formulation was confusing (and possibly formally incorrect), while we assure that the implementation follows the correct extension of the GATK genotype likelihood model (which is referenced in the main text). The equation for genotype likelihoods is now written as:

$$\ln p(O_{m,n}|G_{m,n},Y_{m,n}) = \sum_{r=1}^{C_{m,n}} \ln\Big(\sum_{G_{m,n}\epsilon\{0,1,\ldots,Y_{m,n}\}} \frac{1}{\mathcal{Y}_r} p(O_{m,n,r}|G_{m,n},\epsilon_{m,n,r},Y_{m,n})\Big), \tag{2}$$

$$where \quad p(O_{m,n,r}|G_{m,n},\epsilon_{m,n,r},Y_{m,n}) = \begin{cases} 1-\epsilon_{m,n,r}, & if\ O_{m,n,r}\ \text{in}\ G_{m,n} \\ \frac{\epsilon_{m,n,r}}{3} & otherwise \end{cases}$$

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.

This is a typo, it should be "with". Thanks for spotting it. This has been fixed.

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?

We believe this comment refers to figure 1 in the revised manuscript. This figure is simply a cartoon, an illustration of the HMM for two ploidy levels only. There are no restrictions on which ploidy to start with or end in practice. For this illustration, we can assume that we have selected only ploidy 2 and 4 as it is easy to visualise the different contribution of genotype likelihoods and depth.

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.

We state that these parameters refer to the mean and dispersion of the 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.

Each Poisson-Gamma distribution depends on a ploidy level. This means that all windows assigned the same ploidy will refer to the same mean and dispersion parameters. We write the text above after defining the Poisson-Gamma distribution parameters in the manuscript to avoid confusion.

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.

We moved most of the mathematical details in the supplementary text and refrained from cross-referencing equations.

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 $O(Y^2K)$.

There are both typos and an error spotted in the two questions. The correct letter is K (not in callygraphic font) and the number of ploidy levels is squared. Also the "Big O" notation for the computational cost was missing, since we mean the algorithm is O(Yˆ2K). The expression "bounded by" is incorrect and we rewrite "with computational complexity O(Yˆ2K) (i.e. linear in the number of loci windows)."

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?

We meant that at 0.5X the best accuracy is achieved with higher sample size. We rewrote the paragraph to make it clearer.

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.

The grammar was checked and fixed by co-authors who are native English speakers.

Turning now to the supplementary materials...

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

This is in fact a binomial distribution. This has been fixed.

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.

This was in fact incorrect and now fixed. Please note that this sentence is not present in the text anymore as it was redundant once we moved contents into the supplementary material.

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.

We agreed and we moved most of the mathematical details in the supplementary material and presented equations only once.

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

Thanks. This has been fixed.

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.

Fixed.

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."

Thanks. This has been fixed.

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?

The other reviewer also addressed the difficulty in understanding the ECM and Q. In the text, we now write the following explanations after the 4 steps of the ECM algorithm:

“The ECM algorithm for a HMM with negative binomial observations thus consists of two EC-steps and two maximization steps. Specifically for the four steps above:

1. the first EC-step calculates the expected complete-data log-likelihood with the Markov chain parameters and the dispersion (beta) parameters unknown and to be estimated at the next M-step (maximization step), conditionally to the mean (alpha) parameters estimated at the previous iteration of ECM;

2. the first M-step maximizes the intermediate quantity calculated at the first step w.r.t. the unknown parameters;

3. the second EC-step replicates the first one inverting the roles of known and unknown parameters;

4. now the second intermediate quantity can be maximized w.r.t. the mean parameters.

The EC-step of the ECM algorithm is very similar to the classical forward-backward formulation in the E-step of the EM algorithm. The E-step expresses the expected complete-data log-likelihood with all HMM parameters unknown and to be estimated at the next M-step. The E-step works for observations distributed with one parameter, or multiple parameters whose maximization equations can be solved in normal form, i.e. by isolating the parameter of interest in each equation (e.g. Poisson and Gaussian distribution). The EC-step is a formulation of the E-step where only a portion of the HMM parameters can be estimated in one maximization step (the M-step). This is a characteristic of emission distributions whose parameters can be estimated only in function of each other in a system of equation (e.g. gamma and negative binomial distributions). Look at the following calculations and explanations to see concretely how means and dispersions express each other when writing the functions to maximize.

The calculation of A, delta at iteration l is solved by using the classical forward-backward algorithm, therefore we will only briefly mention the necessary elements of it, while we analyzed more in depth the estimation of means and dispersions.

The scope of each ECM iteration is to maximize the intermediate quantities to achieve the highest value of the complete data log-likelihood. In this way, at each iteration, new parameters can be used to rewrite Q and remaximize it until convergence. It is worth remembering again that the resulting parameters maximize a quantity different from the log-likelihood of the observed data - the ECM uses two forms of Q to make the maximization possible to implement practically, since expressing the log-likelihood directly is not concretely achievable. “

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

The logarithm is element-wise, now we specify it in the text after the equation.

Reviewer 1 (Benjamin Peter)

In this paper, Soraggi et al. introduce a new model for inferring the ploidy of an organism from low-coverage sequence data using genotype likelihoods. This seems like a useful program; but the current manuscript requires substantial revising and editing to make it suitable for publication.

Thanks. We addressed all concerns.

Major points:

1. Introduction: I think the authors should define better what they mean with ploidy, particularly when we talk about ploidy at the sub-chromosome level, and how the authors expect it to differ from structural variation. I.e. I can think of cases like the pseudo-autosomal-region in humans and crazy systems like the platypus X-chromosome; but it would be nice to be explicit about this, I assume it has somehow to do with homologous recombination?

We now extended what we mean by ploidy and aneuploidy in the introduction, which now states that:

Polyploidy is typically defined as the phenomenon whereby the chromosome set is multiplied, resulting the organism to have three or more sets of chromosomes (36). Polyploidy is common to many organisms at different genic and cellular levels, and it can be the consequence of hybridisation or whole genome duplication (14). For instance, polyploidy plays a significant role in the evolution and speciation of plants (41), as 34.5% of vascular plants (including leading commercial crop species) are shown to be polyploid (48).

Of particular interest is the case of aneuploidy, whereby chromosomal aberrations cause the number of chromosomal copies to vary within populations and individuals. Ploidy variation can be associated with a response or adaptation to environmental factors (9), and it is a phenomenon commonly detected in cancer cells (10) and several pathogenic fungi (i.e. *Cryptococcus neoformans*, *Candida albicans* and *Candida glabrata*) and monocellular parasites (43; 34; 12; 50; 49; 15).

Among aneuploid species, *Cryptococcus neoformans* is a fungal pathogen capable of causing meningitis in immunocompromised individuals, particularly HIV/AIDS patients. Ploidy variation, via aneuploidy and polyploidy, is an adaptive mechanism in *Cryptococcus neoformans* capable of generating variation within the host in response to a harsh environment and drug pressure (34). Aneuploidy-driven heteroresistance to the frontline antifungal drug fluconazole has been described (43), resulting in treatment failure in patients. Within fluconazole resistant colonies, aneuploidy was common, particularly disomy of chromosome 1 which harbours the gene encoding the main drug target of fluconazole, *ERG11* (43). For these reasons, 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.

The focus of this method is on applications to aneuploid species, like the applications herein presented, and not on sex chromosomes or cancer cell lines. We removed references to ploidy variation within chromosomes and clarified what we meant by

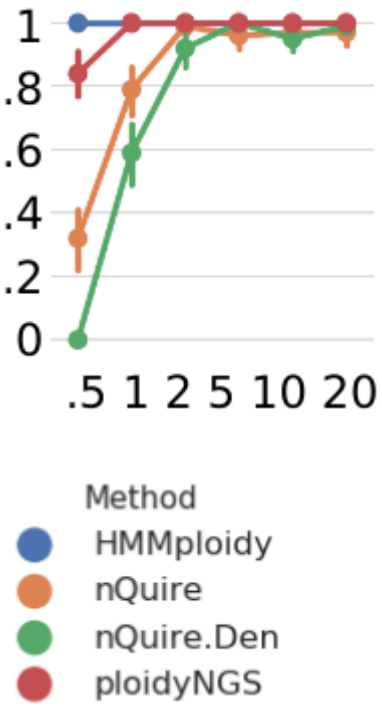
this. In fact, we use the distribution of within-chromosome ploidy levels as both confidence and as a diagnostic tool to highlight local regions with aberrant features. This is now clarified and added as *"While ploidy is not expected to vary within each chromosome, the distribution of local ploidy tracts as inferred by HMMploidy can provide statistical support to whole-chromosome estimates. Additionally, any detected within-chromosome ploidy variation can serve as a diagnostic tool to investigate possible mapping or assembly errors."*

2. Why are coverage-based methods not considered in comparisons? In ancient DNA, sexing is often done by comparing the ploidy of the X-chromosome. This works well at coverages < 0.01x, so I don't understand why these approaches wouldn't work on sufficiently large ploidy-regions. I would imagine at least aneuploidies would be easy to discover with those approaches as well. This needs to be better justified.

Traditional approaches based on allele frequencies (nQuire) and coverage variation (ploidyNGS) are indeed considered and compared against HMMploidy. Please note their description in the introduction.

It is trivial to test between diploid and haploid levels even for very low-coverage data, as in the case of ancient DNA. It is less trivial to test against multiple ploidy levels, as the main focus of this method. In fact, from Figure 2, traditional methods have remarkable less power to detect the correct ploidy

In this snippet from Figure 2:



We report the accuracy (on y-axis) of inferring ploidy 5 at different depths (on x-axis) with only 1 sample. At low-depth, HHMploidy, has the highest accuracy. These results are present in the text.

3. Why does the probability on the rhs of equation 2 not depend on i? Also, why does one not have to correct for the abundance of alleles? I.e. if we have a tetraploid and the genotype is AAAG, why would the probability of seeing As and Gs be equal? I think Equation 2 as stated is simply wrong, and if not, needs to be much better motivated.

Following additional comments from another reviewer, the equation for genotype likelihoods has been rewritten for ease of clarity, and it is now:

$$\ln p(O_{m,n}|G_{m,n},Y_{m,n}) = \sum_{r=1}^{C_{m,n}} \ln\Big(\sum_{G_{m,n}\epsilon\{0,1,\ldots,Y_{m,n}\}} \frac{1}{\mathcal{Y}_,} p(O_{m,n,r}|G_{m,n},\epsilon_{m,n,r},Y_{m,n})\Big), \tag{2}$$

$$where \quad p(O_{m,n,r}|G_{m,n},\epsilon_{m,n,r},Y_{m,n}) = \begin{cases} 1-\epsilon_{m,n,r}, & if\ O_{m,n,r}\ \text{in}\ G_{m,n} \\ \frac{\epsilon_{m,n,r}}{3} & otherwise \end{cases}$$

4. So is the only signal considered in G the heterozygosity? Could that be confounded with population structure?

HMMploidy jointly uses the information on genotype likelihoods and sequencing coverage. This can be evinced from Figure 1 and the introduction *"HMMploidy comprises a Hidden Markov Model (HMM) (31) where the emissions are both sequencing depth levels and observed reads. The latter are translated into genotype likelihoods (29) and population frequencies to leverage the genotype uncertainty.*" More details are given in the methods and supplementary text.

Population structure will affect the probability of genotypes given allele frequencies. In these examples, we only assumed HWE but HMMploidy can receive in input individual inbreeding coefficients to model genotype probabilities in case of deviation from HWE. In fact, we write *"Throughout the analyses carried out in this paper, we assume Hardy-Weinberg equilibrium (HWE) and thus model the genotype probability with a binomial distribution (20; 39). Other methods considering departure from HWE (DHW), can be considered and implemented by ad hoc substitutions of the formula coded in the software. Such functions can be useful in specific situations, such as pathology-, admixture- and selection-induced DHW scenarios (8; 21; 22). How ever, we will leave the treatment of DHW for the inference of ploidy variation in future studies."*

5. I do not think that essentially copying half the paper to the supplement is a good idea. It just makes the manuscript unnecessarily bloated. Why not reduce the supplement to p 4 and 5 which do the heavy lifting. That little care has been given to this arrangement is also apparent that the main text refers to superflues equations in the supplement

We moved most of the mathematical details and methods in the supplementary material. By doing so, we removed all redundancies.

6. Section 2.3: Is reference/sequencing bias an issue here?

We assume that all sequencing data uncertainty is captured in the calculation of genotype likelihoods. More complex genotype likelihoods, not considered herein but implementable in the software, allow for the inclusion of mapping errors, non-independence among reads, and other sources of errors. Additionally, there is no allele polarisation requirement in HMMploidy.

Minor:

Fig 1: I am a bit confused by panel A. What do the little dots represent? Is the unit a window or a SNP?

In Fig1A, dots represent units of data within each window.

p3. why would HWE lead to a negative binomial distribution?

This is in fact a binomial distribution. This has been fixed.

p5. (eq 4) would be good to label equation numbers in the supplement separately. Also, why can't one use the main text equations here?

We changed equation numbering as suggested.

p5. (m-th HMM) should that mean m-th hidden state?) otherwise I don't understand this section

m-th HMM is the HMM for the m-th genome.

p6. The difference between EM and ECM should be explained. Also, in the Baum-Welch-algorithm I am familiar with, the Forward-Backward Algorithm is the E-step of the EM; so what exactly is the EM for each forward-backwards run calculating expectations over?

The EC-step of the ECM algorithm is very similar to the classical forward-backward formulation in the E-step of the EM algorithm. The E-step expresses the expected complete-data log-likelihood with the HMM parameters unknown and to be estimated at the next EM iteration. The E-step works for emission distributions with one parameter, or multiple parameters whose maximization equations can be solved in

normal form, i.e. by isolating the parameter of interest in each equation (e.g. Poisson and Gaussian distribution). The EC-step is a formulation of the E-step where only a portion of the HMM parameters can be estimated in one maximization step (the M-step). This is a characteristic of emission distributions whose parameters can be estimated only in function of each other (e.g. gamma and negative binomial distributions). In our specific example, at the M-step we first estimate all parameters except the scales of the negative binomial distributions (on whose values we condition, from which the C of the EC-step), that are then calculated by conditioning on all parameters except the means. So the ECM procedure is composed of two expected conditional log-likelihoods (conditioned on a portion of the parameters) and two maximization steps.

We write the following text after the ECM steps to clarify the concepts.

"The ECM algorithm for a HMM with negative binomial observations thus consists of two EC-steps and two maximization steps. Specifically for the four steps above:

1. the first EC-step calculates the expected complete-data log-likelihood with the Markov chain parameters and the dispersion (beta) parameters unknown and to be estimated at the next M-step (maximization step), conditionally to the mean (alpha) parameters estimated at the previous iteration of ECM;

2. the first M-step maximizes the intermediate quantity calculated at the first step w.r.t. the unknown parameters;

3. the second EC-step replicates the first one inverting the roles of known and unknown parameters;

4. now the second intermediate quantity can be maximized w.r.t. the mean parameters.

The EC-step of the ECM algorithm is very similar to the classical forward-backward formulation in the E-step of the EM algorithm. The E-step expresses the expected complete-data log-likelihood with all HMM parameters unknown and to be estimated at the next M-step. The E-step works for observations distributed with one parameter, or multiple parameters whose maximization equations can be solved in normal form, i.e. by isolating the parameter of interest in each equation (e.g. Poisson and Gaussian distribution). The EC-step is a formulation of the E-step where only a portion of the HMM parameters can be estimated in one maximization step (the M-step). This is a characteristic of emission distributions whose parameters can be estimated only in function of each other in a system of equation (e.g. gamma and negative binomial distributions). Look at the following calculations and explanations to see concretely how means and dispersions express each other when writing the functions to maximize.

The calculation of A, delta at iteration l is solved by using the classical forward-backward algorithm, therefore we will only briefly mention the necessary elements of it, while we analyzed more in depth the estimation of means and dispersions.

The scope of each ECM iteration is to maximize the intermediate quantities to achieve the highest value of the complete data log-likelihood. In this way, at each iteration, new parameters can be used to rewrite Q and remaximize it until convergence. It is worth remembering again that the resulting parameters maximize a quantity different from the log-likelihood of the observed data - the ECM uses two forms of Q to make the maximization possible to implement practically, since expressing the log-likelihood directly is not concretely achievable. "

----

p6. why is overfitting sets of ploidy levels a concern? How is the number of ploidy levels defined/constrained in the first place?

The distributions modelling the number of observed reads might overlap erroneously, recognizing two or more distinct hidden states matching similar mean and dispersion in the emission distributions. Moreover, noisy data and sequencing errors might create multiple hidden states having quite different emission distributions, when the underlying ploidy would be the same. To avoid this effect, we use the genotype likelihoods to penalize the existence of ploidies that do not fit the data according to genotype likelihoods. The BIC score of the HECM is used to remove the hidden states "in excess" from the model. The initial number of ploidy levels for the model is constrained by allowing the user to set a maximum value or a set of desired ploidy levels to consider in the algorithm.

Typos:

p2 incorporates

p3 (lower case) letters

across reads

In general, the English is quite poor and requires further editing. Also line numbers would greatly help pointing out typos and issues more specifically. This is compounded by the issue that the paper is at times jargon heavy (e.g. Tower property, Markov matrix) and worse, the jargon is not explained and used inconsistently (Markov matrix vs Transition matrix).

We fixed typos and co-authors who were native English speakers checked and fixed the grammar. We fixed the issue with jargon as suggested.

Reviewer 2 (Nicolas Galtier)

This manuscript introduces a method for inferring ploidy and its variation across genomes and loci based on next-generation sequencing data. The main novelty is the introduction of a hidden Markov Model (HMM) in which ploidy is assumed to vary across genomic windows. Ploidy is an important aspect of genome structure, and underlies key technical challenges of genome assembly and analysis, so this manuscript, in my opinion, addresses an important problem. I like much the idea of explicitly modelling ploidy variation and the resulting predictions on patterns of sequence coverage and base counts. I think that the HMMploidy approach has a great potential of significantly advancing the field. That said, I have a number of concerns regarding the manuscript, both content and form, which I detail below. Briefly, I do not think the approach is particularly well motivated or illustrated, I have technical issued with the maths and the way the method is presented, and a suggestion of improvement regarding sequence coverage modeling.

Thanks for the comments. We improved both the motivation and presentation of our method, as detailed in the responses below.

A. Awkward/insufficient justification of the method:

It is not totally intuitive why HMM would be appropriate to model ploidy, since ploidy is typically thought of as a constant, for a given species. In reality, the realized ploidy can vary across chromosomes or chromosomal regions and/or between individuals, making the HMM approach a promising one. The introduction very briefly mentions aneuploidy in cancer cells, and polyploidization in plants, as two possible instances of variable ploidy. The manuscript, however, does not develop on these examples, and rather presents (i) an analysis of data simulated in the absence of any variation in ploidy, and (ii) an analysis of a data set in Cryptococcus neoformans, introduced with very limited biological context. I did not find that the HMMploidy method performs particularly well in these two analyses. It was not obviously better than competing methods in the simulation benchmark, and failed to detect a conspicuous instance of triploidy in the real data analysis.

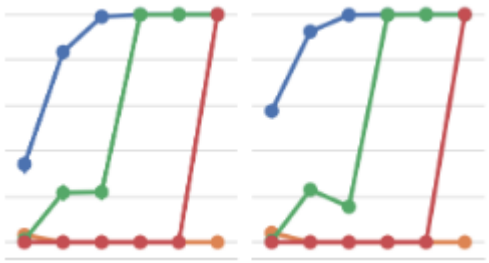
The focus of this method is on applications to aneuploid species, and not on sex chromosomes or cancer cell lines. Our main application is on *Cryptococcus neoformans*. We now provide more biological context in the introduction:

"Cryptococcus neoformans is a fungal pathogen capable of causing meningitis in immunocompromised individuals, particularly HIV/AIDS patients. Ploidy variation, via aneuploidy and polyploidy, is an adaptive mechanism in C. neoformans (and other pathogenic fungi, such as Candida albicans and Candida glabrata) capable of generating variation within the host in response to a harsh environment and drug pressure (Morrow and Fraser, 2013). Aneuploidy-driven heteroresistance to the frontline antifungal drug fluconazole has been described (Stone et al. 2019), resulting in treatment failure in patients. Within fluconazole resistant colonies, aneuploidy was common, particularly disomy of chromosome 1 which harbours the gene encoding the main drug target of fluconazole, ERG11 (Stone et al. 2019, Sionov et al. 2010)."

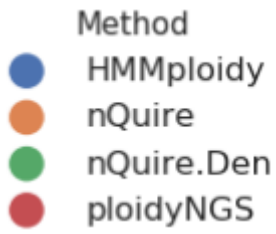
We removed references to ploidy variation within chromosomes and clarified what we meant by this. In fact, we now add in the discussion: *"While ploidy is not expected to vary within each chromosome, the distribution of local ploidy tracts as inferred by HMMploidy can provide statistical support to whole-chromosome estimates. Additionally, any detected within-chromosome ploidy variation can serve as a diagnostic tool to investigate possible mapping or assembly errors."*

Please also note that we extended the introduction with more information on ploidy, ploidy variation and next-generation sequencing data.

From Figure 2, HMMploidy has a remarkably better performance at inferring ploidy with increasing sample size. This snippet



Shows the accuracy (on y-axis from 0 to 1) to infer triploids with 10 and 20 samples (left and right panels), with depth on x-axis (0.5, 1, 2, 5, 10, 20). The labels of different methods are:



We also acknowledge that HMMploidy is not always the most performant method at lower sample sizes. However, we believe our results on the performance of various methods are useful for experimental design and provide general knowledge of the power to infer ploidy at various conditions. Importantly, please also note that *"While nQuire and ploidyNGS sweep the whole simulated genomes, HMMploidy analyses windows of 250bp, so the detection rate is calculated as the windows' average, making the comparison deliberately more unfair to our method."*

Regarding the point that HMMploidy "failed to detect a conspicuous instance of triploidy in the real data analysis", we actually argue the opposite. In fact, HMMploidy is not drastically sensitive to local changes in depth (unlike traditional methods) and was able to infer the correct ploidy in face of aberration increase in ploidy (possibly

due to gene duplications). In fact, where we write *"Interestingly, samples CCTP27 and CCTP27 at day 121 (CCTP27-d121) are inferred to have the same ploidy, even though CCTP27-d121 triplicates its sequencing depth on chromosome 12 (Fig. 3). "*. We argue that the observation that ploidy inference is inconsistent with the local variation in depth can be used to infer on gene duplications / CNVs. Thus, we also write *"This finding suggests the presence of a recent copy number variation. In fact, as no sufficient genetic variation has built up on the recently duplicated triploid chromosome yet, the data is modeled as a single chromosome by the genotype likelihoods."* The ability of HMMploidy to prevent being biased by local changes in depth is given by the integrated use of genotype likelihoods (a feature that no other methods has).

There are a number of reasons why ploidy is expected to vary among/across assembled genomes that are not mentioned or considered in the manuscript. The realized ploidy can be locally increased due to large-scale duplications, when several distinct regions of a genome are so similar that they are assembled as a single piece. Counting gene copy number is indeed a difficult problem (eg see papers by Schrider and Hahn). Another typical artefact with genome assembly is allele splitting, when heterozygosity is so high that assembling algorithms separate homologus alleles as if they were distinct loci (eg have a look at papers on the Ciona savignyi and Adineta vaga genomes, or the recent liteature on haplotig detection and cleaning). The HMMploidy approach seems to be a promising way to identify, annotate and possibly filter out such anomalous genomic regions. Another example of varying ploidy that comes to my mind are sex chromosomes, which are haploid in the heterogametic sex (male in XY systems, female in ZW systems) and diploid in the homogametic sex (see for instance papers by Muyle, Kafer and Marais on how to annotate sex-chromosome-associated contigs). Please note that in many systems (eg mammals) the Y/W chromosome is actually a mosaic of ploidy, with so-called pseudo-autosomal regions being diploid while the sex-specific region is haploid. Each of the topics I'm mentioning in this paragraph is the subject of a voluminous literature.

I would suggest (i) strenghtening the introduction by discussing in more detail why among-loci variation in ploidy is actually relevant, thus justifying the HMM approach, and (ii) identifying a couple of real data sets with clear expectations regarding ploidy variation, and demonstrate the applicability and added value of the newly introduced method.

The focus of this method is on applications to aneuploid species, like the applications herein presented, and not on sex chromosomes or cancer cell lines. We removed references to ploidy variation within chromosomes and clarified what we meant by this. In fact, we use the distribution of within-chromosome ploidy levels as both confidence and as a diagnostic tool to highlight local regions with aberrant features. This is now clarified and added as *"While ploidy is not expected to vary within each chromosome, the distribution of local ploidy tracts as inferred by HMMploidy can provide statistical support to whole-chromosome estimates. Additionally, any detected within-chromosome ploidy variation can serve as a diagnostic tool to investigate possible mapping or assembly errors."*

B. Awkward/inaccurate presentation of the method:

I have several concerns with the way the method is presented, which I think mostly result from insufficient clarity. At any rate at the moment I can't say I totally understand what the method exactly does, and the manuscript apparently contains incorrect equations.

We fixed the equation on genotype likelihoods and moved most of the mathematical details in the supplementary material.

- 2.1 first sentence: "N polymorphic sites"; how do we know a site is polymorphic or not prior to the analysis? Should one perform SNP calling beforehand? Maybe remove "polymorphic"?

We removed “polymorphic” as suggested.

- 2.1: a genotype is described as the number of "alternate (or derived) alleles", suggesting that SNPs are assumed to be polarized. I do not think that the method presented here requires SNP polarization (which is good), so I would suggest clarifying.

The method does not assume SNP polarisation. We added “Any of the two alleles can be considered to define G_{m,n}.”

- 2.1: "We assume Hardy-Weinberg equilibrium (HWE) and thus model the genotype probability with a negative binomial distribution" -> I would rather think a binomial distribution?

This was a typo and it has been fixed as suggested.

- 2.2: Equation 2 appears awkward. The summation variable i does not appear in the term right to the Sigma symbol, which is suggestive of a problem. Also a genotype G_mn was defined above as an integer taking value in {0, ..., Y_mn}, but here appears the idea that O_mnr (some observed nucleotide) can be "in G_mn" (second part of equation 2), which is inconsistent.

I guess one could re-define a genotype as a vector of nucleotide instead of an integer, then replace in equation 2

p(O_mnr|G_mn,Q_mnr,Y_mn)

with

p(O_mnr|G_mni,Q_mnr,Y_mn)

and replace in second line of equation 2

"if O_mnr in G_mn"

with

"if O_mnr = G_mni"

Alternatively one could keep the text definition of genotype, call A_n and a_n the two alleles at locus n (say), and replace in equation 2:

sum_i p(Omnr|G_mn,Q_mnr,Y_mn)/Y_mn

with

((1-G_mn) p(O_mnr|A_n,Q_mnr,Y_mn) + G_mn p(O_mnr|a_n,Q_mnr,Y_mn))/Y_mn

and adjust second line of equation 2.

The above two options, which I think are equivalent (but different from the text), are what makes sense to me. In the rest of this review I'm assuming that the calculation that was actually made corresponds to the above modified equations.

We rewrote the equation of genotype likelihoods as it was confusing (and possibly formally incorrect). It now reads:

$$\ln p(O_{m,n}|G_{m,n},Y_{m,n}) = \sum_{r=1}^{C_{m,n}} \ln\Big(\sum_{G_{m,n}\epsilon\{0,1,\ldots,Y_{m,n}\}} \frac{1}{\mathcal{Y}}, p(O_{m,n,r}|G_{m,n},\epsilon_{m,n,r},Y_{m,n})\Big), \quad (2)$$

$$where \quad p(O_{m,n,r}|G_{m,n},\epsilon_{m,n,r},Y_{m,n}) = \begin{cases} 1-\epsilon_{m,n,r}, & if\ O_{m,n,r}\ \text{in}\ G_{m,n} \\ \frac{\epsilon_{m,n,r}}{3} & otherwise \end{cases}$$

Please note that the implementation follows the correct equation, as the extension of GATK genotype likelihood model to polyploids.

- 2.3: equation 3 is a rather complex way of saying that the estimated alternate allele frequency is the observed alternate allele frequency across all reads from the pooled genome sample. Indeed Fhat_mn in equation 3 can be written as f_mn/C_mn, where f_mn is the observed number of alternate alleles in genome m, so C_mn cancels out and we get Fhat_n=sum(f_mn)/C_n.

We changed the equation as suggested.

[now switching to Supplementary Material]

- 6.5: I am not sure what alpha and beta are. I guess these correspond to the shape and scale parameter of the Poisson-Gamma distribution of mean coverage across windows - this should be specified. Secondly, I do not understand why these parameters appear with a _k index, suggesting there is one alpha and one beta per window. The text and figure S1 instead suggest that there is one value of coverage per window, C_m(k), drawn from a

unique Poisson-Gamma distribution, the parameters of which should be shared across windows?

There is an alpha and a beta for each ploidy level/hidden state of the HMM. The index k refers to the ploidy level of the k-th window, i.e. Y_(m)^k, so that for the same ploidy in different windows there will be the same pair of parameters describing the observed data.

This same issue was raised by the other reviewer, denoting that the definition was not entirely clear. We try to help the understanding of the text by stating the following after the definition of parameters:

Note that the Poisson-Gamma distributions depend each on a ploidy level. This means that all windows assigned the same ploidy will refer to the same mean and dispersion parameters.

C. Modeling scheme:

The way sequencing coverage is modeled lacks clarity and justification. Irrespective of ploidy, there might be differences in coverage among loci (e.g. GC-rich vs GC-poor regions) and among genomes (due to experimental setting or the experimental noise). It would appear natural to me to model the among-loci variation in coverage as suggested in the ms, to also model among-genomes variation in coverage (i.e., introduce genome specific coverage parameters), and to define C_m(k) as the product of these two terms - thus assuming that the locus-effect and the genome-effect are independent. If one thinks this is too strong an assumption, maybe some (de)correlation parameter could be introduced. My understanding of the current method is that the across_loci distribution of coverage is assumed to be independent across genomes, i.e., the fact that one locus is highly covered in one genome says nothing about coverage at the same locus in another genome. This sounds like an highly, maybe overly, versatile model, which I think might induce some loss of signal. For instance, the analysis of chromosome 12 in the Cryptococcus CCTP27-d121 sample did not detect any change in ploidy even though coverage is consistently tripled across a large portion of the chromosome (fig 2). I suggest that if coverage was modeled in a more constrained way - i.e. as the product of a genome-specific and a locus-specific term - this abnormality could be interpreted by the method as a triplication. A clarification of how coverage is modeled across loci and genomes, a discussion of this question, and an attempt to adopt a less versatile scheme, would appear required.

This is an interesting extension of modelling coverage across the genome which will have important applications, for instance, for the analysis of cancer genomes. We mention in the introduction that some methods use GC-content information. We state "Available computational methods to infer ploidy levels from genomic data are [...] based [...] on using inferred genotypes and information on GC content - although this is an approach specific for detecting aberrations in cancer genomes (e.g. AbsCN-seq (4), sequenza (13))". Adding dependency to GC-content for the depth-associated component of HMMploidy is beyond the scope of this paper, but an interesting extension to follow up in future studies.

Additionally, We wish to reiterate that HMMploidy successfully prevented a wrong inference of Cryptococcus CCTP27-d121 being triploid despite an aberrant increase in coverage, as explained earlier. This is thanks to the joint use of depth and genotype likelihoods.

D. Minor

- section 3: "averaged by the polyploid genome size" -> "divided by genome size" ?

We now write "The sequencing depth is defined as the average number of sequenced bases at one site for each chromosomal copy (i.e. divided by the ploidy level)."

- Simulations: section 3 says that ploidy 1 to 20 have been simulated, but the result section and figure 2 only consider ploidy 1 to 5.

Simulations were performed with this combination of parameters: ploidy (from 1 to 5, constant along each genome), sample size (1, 2, 5, 10, 20), and sequencing depth (0.5X, 1X, 2X, 5X, 10X, 20X).

- Discussion: "On the former point, rescaling sequencing depth across genomes is not possible since HMMploidy models a distribution of read counts." -> I do not understand this sentence.

We agree that the sentence is not clear enough.

We want to underline that one could limit itself to define only one HMM for all genomes. To do this, sequencing depth could be for example standardized in each genome. Firstly, this would make the estimation of the distributions of standardized counts very difficult, especially in samples with a lot of noise, errors, and/or low-depth sequencing. Secondly, two genomes could easily have two different ploidy levels matching the same distribution parameters. For example a diploid-tetraploid sample where the two ploidy levels have observations' mean parameters -1 and 1 could match haploid-diploid levels in another genome having the same mean parameters for the ploidy-related observations. The only case in which one can use the same HMM for all genomes is when they have all the same ploidy levels, but we did not implement this function because it is a very unusual and unlikely practical case.

We clarify this by substituting the sentence with:

On the former point, assume one could limit itself to define only one HMM for all genomes. To do this, sequencing depth could be for example standardized in each genome. Firstly, this would make the estimation of the distributions of standardized counts very difficult, especially in samples with a lot of noise, errors, and/or low-depth sequencing. Secondly, two genomes could easily have two different ploidy levels matching the same distribution parameters. For example a diploid-tetraploid sample where the two ploidy levels have observations' mean parameters -1 and 1

could match haploid-diploid levels in another genome having the same mean parameters for the ploidy-related observations. The only case in which one can use the same HMM for all genomes is when they have all the same ploidy levels, but we did not implement this function because it is a very unusual and unlikely practical case.