HMMploidy: inference of ploidy levels from short-read sequencing data

Samuele Soraggi[1,2][0000-0002-1159-5535], Johanna Rhodes[3][0000-0002-1338-7860], Isin Altinkaya[2,4,5][0000-0002-6364-3332], Oliver Tarrant[2], François Balloux[6][0000-0003-1978-7715], Matthew C. Fisher[3][0000-0002-1862-6402], and Matteo Fumagalli[2,7][0000-0002-4084-2953]

[1] Bioinformatics Research Center (BiRC), University of Aarhus, 8000 Aarhus, Denmark, samuele@birc.au.dk
[2] Department of Life Sciences Silwood Park, Imperial College London, Ascot, SL5 7PY, UK
[3] MRC Centre for Global Infectious Disease Analysis, Department of Infectious Disease Epidemiology, Imperial College London, London, W2 1PG, UK
[4] Department of Biology, Hacettepe University, 06800 Beytepe Campus, Ankara, Turkey
[5] GLOBE, Section for Geogenetics, Øster Voldgade 5-7, 1350, Copenhagen, Denmark
[6] UCL Genetics Institute, University College London, London, WC1E 6BT, UK
[7] School of Biological and Behavioural Sciences, Queen Mary University of London, London, E1 4NS, UK, m.fumagalli@qmul.ac.uk

Abstract. The inference of ploidy levels from genomic data is important to understand molecular mechanisms underpinning genome evolution. However, current methods based on allele frequency and sequencing depth variation do not have power to infer ploidy levels at low- and mid-depth sequencing data, as they do not account for data uncertainty. Here we introduce `HMMploidy`, a novel tool that leverages the information from multiple samples and combines the information from sequencing depth and genotype likelihoods. We demonstrate that `HMMploidy` outperforms existing methods in most tested scenarios, especially at low-depth with large sample size. We apply `HMMploidy` to sequencing data from the pathogenic fungus *Cryptococcus neoformans* and retrieve pervasive patterns of aneuploidy, even when artificially downsampling the sequencing data. We envisage that `HMMploidy` will have wide applicability to low-depth sequencing data from polyploid and aneuploid species.

Keywords: high-throughput DNA sequencing · ploidy · polyploidy · aneuploidy · hidden Markov model · genotype likelihood

Introduction

In recent years, advances in Next Generation Sequencing (NGS) technologies allowed for the generation of large amount of genomic data (38; 27). Many statistical and computational methods, and accompanying software, to process NGS data for genotype and variant calling have been proposed (30; 23; 3). Additionally, dedicated software have been developed to analyse low-coverage sequencing data (40; 19), a popular and cost-effective approach in population genomic studies (34). 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 (42). Polyploidy is common to many organisms at different genic and cellular levels, and it can be the consequence of hybridisation or whole genome duplication (17). For instance, polyploidy plays a significant role in the evolution and speciation of plants (47), as 34.5% of vascular plants (including leading commercial crop species) are shown to be polyploid (55).

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 (12), and it is a phenomenon commonly detected in cancer cells (13) and several pathogenic fungi (i.e. *Cryptococcus neoformans*, *Candida albicans* and *Candida glabrata*) and monocellular parasites (49; 39; 15; 57; 56; 18; 4).

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 (39). Aneuploidy-driven heteroresistance to the frontline antifungal drug fluconazole has been described (49), 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* (49). 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.

Available computational methods to infer ploidy levels from genomic data are based either on modelling the distribution of observed allele frequencies (`nQuire` (54)), comparing frequencies and coverage to a reference data set (`ploidyNGS` (2)), or using inferred genotypes and information on GC-content, although the latter is an approach specific for detecting aberrations in cancer genomes (e.g. `AbsCN-seq` (5), `sequenza` (16)). A popular approach is based on the simple eyeballing method, that is, on the visual inspection of variation of sequencing depth (compared to another ground-truth data set sequenced with the same

setup) and allele frequencies (2). However, methods based only on sequencing depth, allele frequencies and genotypes limit the inference on the multiplicity factor of different ploidy levels only (if present). Additionally, they often need a reference data with known ploidy to be compared to, and they generally lack power for low- or mid-depth sequencing data applications, which are typically affected by large data uncertainty. As low-coverage whole genome sequencing is a common strategy in population genetic studies of both model and non-model species (50), a tool that incorporates data uncertainty is in dire need.

To overcome these issues, we introduce a new method called `HMMploidy` to infer ploidy levels from low- and mid-depth sequencing data. `HMMploidy` comprises a Hidden Markov Model (HMM) (43) where the emissions are both sequencing depth levels and observed reads. The latter are translated into genotype likelihoods (40) and population frequencies to leverage the genotype uncertainty. The hidden states of the HMM represent the ploidy levels which are inferred in windows of polymorphisms. Notably, `HMMploidy` determines automatically its number of latent states through a heuristic procedure and reduction of the transition matrix. Moreover, our method can leverage the information from multiple samples in the same population by estimate of population frequencies, making it effective at very low depth.

`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 (CNVs) 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.

`HMMploidy` is written in `R/C++` and `python`. Source code is freely available at https://github.com/SamueleSoraggi/HMMploidy, integrated into `ngsTools` (22), and FAIR data sharing is available at the OSF repository https://osf.io/5f7ar/. We will first introduce the mathematical and inferential model underlying `HMMploidy`, then show its performance to detect ploidy levels compared to existing tools, and finally illustrate an application to sequencing data from the pathogenic fungus *Cryptococcus neoformans*.

Material and methods

This section describes the methods used in the implementation of the `HMMploidy` software. In what follows, data is assumed to be diallelic (i.e. we observe at most two states at a particular genotype regardless of the number of copies), without loss of generality. Allowing for more than two alleles would add a summation over all possible pairs of alleles in all calculations. In our notation, indices are

lower case and vary within an interval ranging from 1 to the index's upper case letter, e.g. $m = 1, \dots, M$.

Probability of sequenced data

Let $O = (O_1, \dots, O_M)$ be the observed NGS data for M sequenced genomes at N sites. Consider an m-th genome and n-th locus. 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. For ease of notation, we suppress the two indices, since they do not vary in the formula (1). For such genome and locus define Y, G and O as the ploidy, genotype and sequencing data, respectively. Given Y, the genotype G assumes values in $\{G_0, \dots, G_Y\}$, where each element is the collection of nucleotides of the genotype, and $|G_i|$ is the number of alternate (or derived) alleles of the i-th genotype.

The probability of the sequenced data, conditionally on the ploidy Y and the population frequency F at locus n, is expressed by

$$p(O|Y,F) = \sum_{G \in \{G_0,\dots,G_Y\}} p(O|G,Y)p(G|Y,F), \tag{1}$$

where the left-hand side of the equation has been marginalised over the genotypes, and the resulting probabilities have been rewritten as product of two terms using the tower property of the probability. The first factor of the product is the genotype likelihood (36). Note that the only varying parameter in it is the genotype; therefore it is also rewritten as $L(G|O,Y)$. The second factor is the probability of the genotype given the population frequency and the ploidy level, in other words the prior probability of the genotype. The marginalisation over all possible genotypes has therefore introduced a factor that takes into account the genotype uncertainty. The calculation of genotype likelihoods for an arbitrary ploidy number and the estimation of population allele frequencies are described in the Supplementary Material.

Throughout the analyses carried out in this paper, we assume Hardy-Weinberg equilibrium (HWE) and thus model the genotype probability with a binomial distribution (24; 53). 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 (10; 25; 26). However, we will leave the treatment of DHW for the inference of ploidy variation to future studies.

Hidden Markov Model for ploidy inference

Here, the HMM is defined, and the inferential process of ploidy levels from the HMM is illustrated. Further mathematical details, proofs and algorithms are available in the Supplementary Material.

Consider the N sites arranged in K adjacent and non-overlapping windows. For each individual m, `HMMploidy` defines a HMM with a Markov chain of length

K of latent states $Y_m^{(1)}, \ldots, Y_m^{(K)}$, as shown for a sequence of two ploidy levels (Fig. 1A) in the graphical model of dependencies of Fig. 1B. Each k-th latent state represents the ploidy level at a specific window of loci, and each window's ploidy level depends only on the previous one. Therefore, the sequence of states is described by a transition matrix $\boldsymbol{A}$ of size $|\mathcal{Y}| \times |\mathcal{Y}|$ and a $|\mathcal{Y}|$-long vector of starting probabilities $\boldsymbol{\delta}$, where $\mathcal{Y}$ is the set of ploidy levels included in the model and $|\mathcal{Y}|$ is the number of ploidy levels (i.e. cardinality of $\mathcal{Y}$) (Fig. 1C).

In the HMM structure, each of the $|\mathcal{Y}|$ ploidy levels emits two observations (Fig. S1). Those contain a dependency on which ploidy is assigned to that window. The observations consist of the sequenced reads $O_m^{(k)}$ and the average sequencing depth $C_m^{(k)}$ in the k-th window (Fig. 1B). The former is modelled by the probability in Equation 1; the latter by a Poisson-Gamma distribution (7; 9) (Fig. 1D). The Poisson-Gamma distribution consists of a Poisson distribution whose mean parameter is described by a Gamma random variable. This generates a so-called super-Poissonian distribution, for which the mean is lower than the variance. This allows us to model overdispersed counts, a common issue in NGS datasets (1).

For the m-th HMM, the Poisson-Gamma distribution in window k is modelled by the ploidy-dependent parameters $\alpha_{Y_m^{(k)}}, \beta_{Y_m^{(k)}} \in \mathbb{R}$, describing mean and dispersion, where $Y_m^{(k)}$ is the ploidy in the considered window. In each window, the estimated population frequencies serve as a proxy for the probability of sequenced reads. 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.

We propose a heuristic optimisation algorithm to automatically find the number of latent states of the HMM, and to assign them to the correct ploidy through the genotype likelihoods. Our implementation, described in the Supplementary Material, is a heuristic version of the well-known Expectation Conditional Maximisation (ECM) algorithm (8).

Simulated data

The required memory, runtime and ploidy detection power of `HMMploidy` were compared to the ones obtained by other methods using simulated data. We simulated sequencing reads under a wide range of scenarios using a previously proposed approach (21; 20). Specifically, each locus is treated as an independent observation, without modelling the effect of linkage disequilibrium. The number of reads is modelled with a Poisson distribution with parameter given by the input depth multiplied by the ploidy level. At each locus, individual genotypes are randomly drawn according to a probability distribution defined by a set of population parameters (e.g., shape of the site frequency spectrum). Once genotypes are assigned, sequencing reads (i.e. nucleotidic bases) are sampled with replacement with a certain probability given by the base quality scores.

For comparing the performance of detecting ploidy between `HMMploidy` and existing tools, 100 simulations of M genomes are performed for every combina-

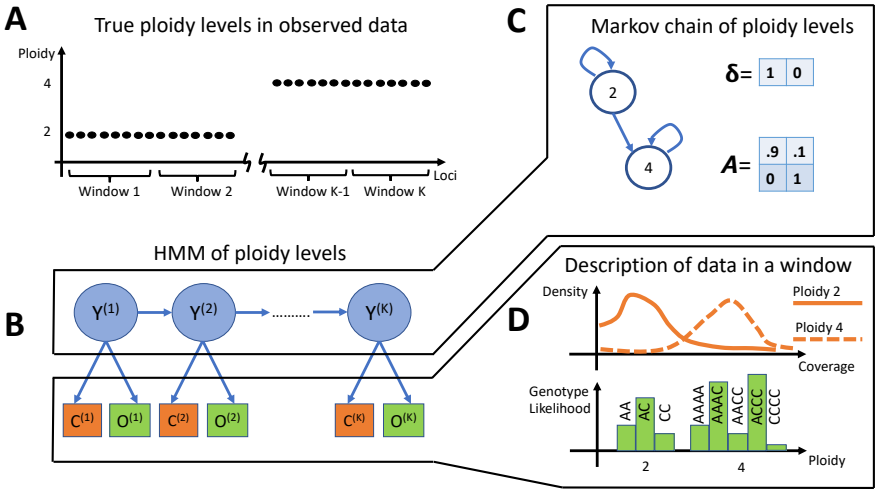

Fig. 1: **HMM for two ploidy levels.** (A) Consider a NGS dataset consisting of a sequence of two ploidy levels. (B) The HMM describing the data has a sequence of hidden states $Y^{(1)}, \ldots, Y^{(K)}$ - one for each window of loci - that can assume one of two values of the ploidies. Observations $C^{(1)}, \ldots, C^{(K)}$ and $O^{(1)}, \ldots, O^{(K)}$ describe the sequencing depth and observed reads in each window, respectively. The index related to the sample is omitted to simplify the notation. (C) The sequence of ploidy levels is described by a Markov chain with two states, governed by a starting vector $\boldsymbol{\delta}$ and a Markov matrix $\boldsymbol{A}$. (D) At each window, the observations are described by the distribution of depth. There are two distributions, each one dependant on the ploidy level. Similarly, genotype likelihoods describe the observed reads by modelling the genotypes at two distinct ploidy levels.

tion of 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). 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). Each simulated genome has a length of 5Kb with all loci being polymorphic in the population.

Simulated data for the analysis of runtime and memory usage consist of 100 diploid genomes of length 10kb, 100kb, 1Mb, 10Mb. Each simulated genome comprises an expected proportion of polymorphic sites equal to 1%. The simulation scripts and pipelines are included in the Github and OSF repositories. Performance analysis was performed on a cluster node with four reserved cores of an Intel Xeon Gold 6130 @1.00GHz with 24GB of RAM and the Ubuntu 18.04.3 OS.

Application to real data

To illustrate the use of `HMMploidy`, we apply it to sequencing data from 23 isolates of the pathogenic fungus *Cryptococcus neoformans* recovered from HIV-infected patients showing clinical evidence cryptococcal meningitis (45). Whole-genome sequencing data was performed on an Illumina machine following an established protocol for sample preparation (44) and data processing (45). Reads are mapped onto *C. neoformans* H99 reference genome (32), yielding an average depth of approximately 100 reads per site. We generated an additional data set by randomly sampling only 20% of reads for each sample. All sequencing raw reads were retrieved from the European Nucleotide Archive under the project accession PRJEB11842.

Results and discussion

Predictive performance

We assess the power of `HMMploidy` to infer ploidy levels on simulated genomes ranging from haploid to pentaploid. Samples sizes varied from 1 to 20 individuals haplotypes, and sequencing depths from 0.5X to 20X. `HMMploidy` is compared to the two state-of-the-art methods `ploidyNGS` (2) and `nQuire` (including a version with denoising option, `nQuire.Den`) (54). The former performs a Kolmogorov-Smirnov test between the minor allele frequencies of the observed data and of simulated data sets at different ploidy levels (simulated at 50X). The latter models the minor allele frequencies with a Gaussian mixture model. We exclude depth-based methods because they are hardly applicable to low sequencing depth (Fig. S2,S3) and work as empirical visual checks rather than algorithmic procedures. 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.

At low-depth (0.5X), `HMMploidy`'s power increases with sample size up to 20 - the largest we considered - in all scenarios excluding the tetraploid case (Fig. 2)).

This might be because it is difficult to distinguish diploid and tetraploid genotypes at such low depth. In the haploid and diploid case `ploidyNGS` has a remarkable 100% success at very low depths (Fig. 2). This is likely because having only few reads makes it easier to compare the data to a simulated genome with low ploidy level and a simpler distribution of observed alleles. However, this erratic behaviour disappears at higher ploidy levels, and `ploidyNGS` is generally outperformed by `nQuire.Den` and/or `HMMploidy`. `HMMploidy` is outperformed at low depth in the tetraploid scenario by both versions of `nQuire`. This might indicate that genotype likelihoods are not successful in modelling tetraploid genotypes as well as allele frequencies in this specific scenario.

Note also that none of the methods performs well with a single haploid sample. This happens because many loci show only one possible genotype, and even with the genotype likelihoods it is impossible to determine the multiplicity of the ploidy. With more samples it is possible to exploit loci with at least another allele to inform on the most likely genotype.

In all tested scenarios, `HMMploidy` greatly improves its accuracy with increasing sample size, with unique good performances at low depth (Fig. 2) not observed with other methods. Additionally, `HMMploidy` infers ploidy levels in sliding windows across the data (as in Fig. 3). Moreover, `HMMploidy` does not require a reference genome at a known ploidy, unlike `ploidyNGS`. `HMMploidy` can identify haploid genomes, unlike `nQuire`. Note that either deeper sequencing depth or larger sample size is likely to be required for `HMMploidy` to detect higher levels of ploidy, as the power of the method decreases with increasing ploidy (Fig. S4).

Computational performance

The benchmark of `HMMploidy` shows a rather constant CPU time across genome lengths by keeping the number of windows fixed at $K = 100$ (Fig. S5A). The shortest simulations are an exception, due to a very fast processing of the data to be used in the HMM. Occasionally, runtimes are elevated for cases where the inference algorithm is converging with difficulty. Fig. S5B shows the effect of increasing the number of windows on 10MB genomes. The growth of execution time follows linearly the increase of K, plus a probable extra overhead for preprocessing the data in many windows, showing that the forward-backward complexity $O(|\mathcal{Y}|^2 K)$ dominates the algorithm. In both the length- and windows-varying scenarios, memory usage was kept at an almost constant value of $350MB$. This is possible thanks to the implementation of file reading and frequency estimation in `C++`. Both `nQuire` and `ploidyNGS` are obviously extremely fast and run in less than one second because they only need to calculate and compare observed allele frequencies, with a cost approximately comparable to the number of loci in the data. Therefore, their performance is not reported in the benchmark figures. Analogous trends on execution times would follow for genomes longer than 10MB and we expect `HMMploidy` to run without issues on larger genomes.

Note that `HMMploidy` trains a separate HMM on each genome even for larger sample sizes. As shown above, each HMM might require considerable CPU time

if many windows are used, or if the heuristic ECM algorithm has a slow convergence. 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. Additionally, 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. The only case in which one can use the same HMM for all genomes is when they have all the same ploidy levels. However, this function is not implemented in `HMMploidy`. On the latter point, it would not be possible to detect sample-specific variation in ploidy levels when training the HMM on pooled genomic data. Therefore, training a separate HMM on each genome is an important feature in `HMMploidy`. However, a simple extension of `HMMploidy` would allow to estimate an HMM on the pooled data from multiple genomes, and to initiate HMM parameters and number of latent states to reduce the model estimation runtime. These options might be implemented in future versions of the software.

Application to real data

We used `HMMploidy` to infer ploidy variation in 23 isolates of *Cryptococcus neoformans* recovered from HIV-infected patients (45). By analysing variation in normalised sequencing coverage, Rhodes and coworkers identified extensive instances of aneuploidy, especially on chromosome 12, in several pairs of isolates (45), in line with previous findings using karyotypic analysis (41). We sought to replicate these inferences using `HMMploidy` and assessed its performance on a downsampled data set to mirror data uncertainty.

In accordance with the original study (45), we retrieve patterns of polyploidy and aneuploidy within each isolate. Most of the analysed samples are haploid (Fig. 3 and Fig. S6-S28). 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 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. Sample CCTP50 had on average a higher depth at day 409, but chromosome 1 changed from diploid (day 1) to haploid (day 409). Chromosome 12 was triploid at day 409 although the high variability of sequencing depth is not informative on the ploidy.

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 (Fig. S6-S28). Interestingly, `ploidyNGS`, `nQuire` and `nQuire.Den` infer

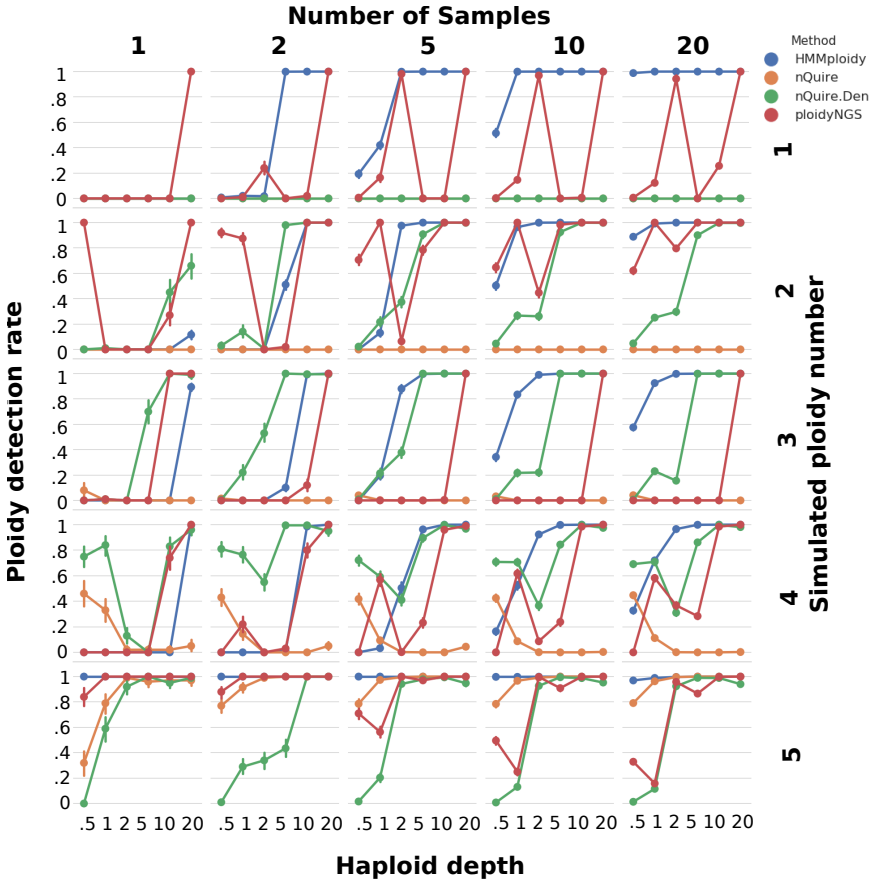

Fig. 2: **Comparison of ploidy detection rates for different methods at various experimental scenarios.** The rate of detecting the correct ploidy (y-axis) is shown against the haploid sequencing depth (x-axis) for different sample sizes (on columns) and ploidy levels (on rows). For every simulated ploidy level, at each value of the sequencing depth we generate M genomes 100 times, where M is the number of simulated samples. The ploidy detection rate is the proportion of correctly detected ploidy levels in the genomic windows with the HMM method, and the proportion of correctly detected ploidy levels along each whole genome with the other tested methods.

the highest tested ploidy in almost all windows of the 23 samples (Supplementary Table 1). This is likely because these methods fit the distribution of widely varying allele frequencies in each sample with the most complex ploidy model, as they do not consider the information of genotype likelihoods.

Cryptococcal meningitis, caused by the fungal yeasts *Cryptococcus neoformans* and *Cryptococcus gattii*, is a severe infection mostly affecting HIV/AIDS patients (35). Oral fluconazole antifungal therapies are widely used for treatment of Cryptococcal meningitis, although their efficacy is reported to be poor especially in Sub-Saharan Africa (33). Resistance to antifungal drugs is thought to be responsible for such poor outcomes and relapse episodes, but its molecular mechanisms are not yet understood (49). Resistance to oral fluconazole antifungal drugs in *Cryptococcus neoformans* was associated with aneuploidy (48). Recent genomic studies identified multiple occurrences of aneuploidy in resistant and relapse isolates (49). Our genomics inferences of aneuploidy in *Cryptococcus neoformans* from HIV-infected patients can serve as diagnostic and molecular surveillance tools to predict and monitor drug resistance isolates, whilst further providing novel insights into the pathogen's evolution (46) We envisage that `HMMploidy` can be deployed to large-scale genomics data of pathogenic species to characterise aneuploidy-mediated drug resistance.

Conclusions

Here we introduce `HMMploidy`, a method to infer ploidy levels suitable for low- and mid-depth sequencing data, as it jointly uses information from sequencing depth and genotype likelihoods. `HMMploidy` outperforms traditional methods based on observed allele frequencies, especially when combining multiple samples. 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).

Acknowledgements

We are grateful to Alan Rogers, Barbara Holland, Benjamin Peter, Nicolas Galtier, and several anonymous reviewers for improving the manuscript.

Funding

JR and MCF were supported by a grants from Natural Environmental Research Council (NERC; NE/P001165/1 and NE/P000916/1), the UK Medical Research Council (MRC; MR/R015600/1) and the Wellcome Trust (219551/Z/19/Z). MCF is a CIFAR Fellow in the 'Fungal Kingdom' programme. We acknowledge support from the Erasmus+ programme to IA.

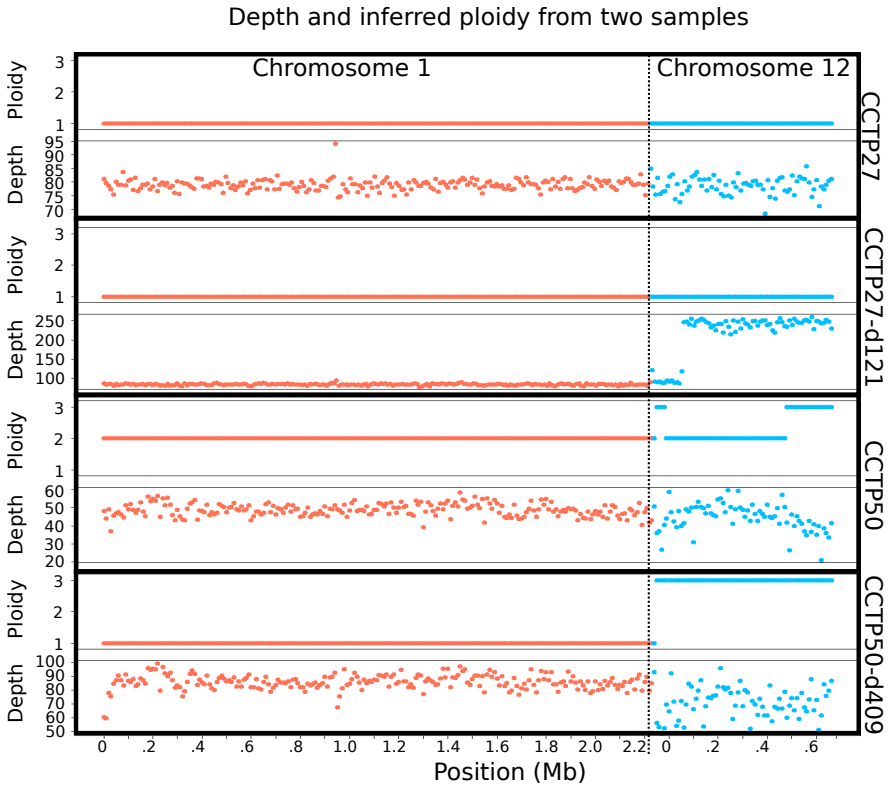

Fig. 3: **Inference of ploidy levels on two samples of *Cryptococcus neoformans* at different time points using `HMMploidy`.** Inferred ploidy and corresponding sequencing depth are shown in genomic windows for two samples at day 1 (CCTP27 and CCTP50), day 121 (CCTP27-d121) and 409 (CCTP50-d409) on chromosomes 1 and 12.

Conflicts of interest

Matteo Fumagalli is a recommender for Peer Community In Mathematical and Computational Biology.

Bibliography

[1] Anders, S., Huber, W.: Differential expression analysis for sequence count data. Genome biology **11**(10), R106–R106 (2010). https://doi.org/10.1186/gb-2010-11-10-r106, https://pubmed.ncbi.nlm.nih.gov/20979621, 20979621[pmid]

[2] Augusto Corrêa dos Santos, R., Goldman, G.H., Riaño-Pachón, D.M.: ploidyNGS: visually exploring ploidy with Next Generation Sequencing data. Bioinformatics **33**(16), 2575–2576 (aug 2017). https://doi.org/10.1093/bioinformatics/btx204, http://www.ncbi.nlm.nih.gov/pubmed/28383704

[3] Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K.V., Altshuler, D., Gabriel, S., DePristo, M.A.: From fastq data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. Current Protocols in Bioinformatics **43**(1), 11.10.1–11.10.33 (2013). https://doi.org/https://doi.org/10.1002/0471250953.bi1110s43, https://currentprotocols.onlinelibrary.wiley.com/doi/abs/10.1002/0471250953.bi1110s43

[4] Avramovska, O., Rego, E., Hickman, M.A.: Tetraploidy accelerates adaption under drug-selection in a fungal pathogen. bioRxiv (2021). https://doi.org/10.1101/2021.02.28.433243, https://www.biorxiv.org/content/early/2021/02/28/2021.02.28.433243

[5] Bao, L., Pu, M., Messer, K.: AbsCN-seq: a statistical method to estimate tumor purity, ploidy and absolute copy numbers from next-generation sequencing data. Bioinformatics **30**(8), 1056–1063 (apr 2014). https://doi.org/10.1093/bioinformatics/btt759, https://academic.oup.com/bioinformatics/ARTICLE-lookup/doi/10.1093/bioinformatics/btt759

[6] Ben-David, U., Amon, A.: Context is everything: aneuploidy in cancer. Nature Reviews Genetics **21**(1), 44–62 (Jan 2020). https://doi.org/10.1038/s41576-019-0171-x, https://doi.org/10.1038/s41576-019-0171-x

[7] Bishop, C.M.: Pattern recognition and machine learning. Springer (2006)

[8] Cappe, O., Moulines, E., Ryden, T.: Inference in Hidden Markov Models. Springer Science+Business Media, Inc (2005)

[9] Casella, G., Berger, R.L.: Statistical inference. Thomson Learning (2002)

[10] Chen, B., Cole, J.W., Grond-Ginsbach, C.: Departure from Hardy Weinberg Equilibrium and Genotyping Error. Front Genet. (8) (2017)

[11] Choudhary, S., Satija, R.: Comparison and evaluation of statistical error models for scrna-seq. Genome Biology **23**(1), 27 (Jan 2022). https://doi.org/10.1186/s13059-021-02584-9, https://doi.org/10.1186/s13059-021-02584-9

[12] Coward, J., Harding, A.: Size does matter: Why polyploid tumor cells are critical drug targets in the war on cancer. Frontiers in Oncology **4**, 123 (2014). https://doi.org/10.3389/fonc.2014.00123, https://www.frontiersin.org/article/10.3389/fonc.2014.00123

[13] Davoli, T., de Lange, T.: The causes and consequences of polyploidy in normal development and cancer. Annual Review of Cell and Developmental Biology **27**(1), 585–610 (2011). https://doi.org/10.1146/annurev-cellbio-092910-154234, https://doi.org/10.1146/annurev-cellbio-092910-154234, pMID: 21801013

[14] Ewing, B., Hillier, L., Wendl, M.C., Green, P.: Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome research **8**(3), 175–85 (mar 1998), http://www.ncbi.nlm.nih.gov/pubmed/9521921

[15] Farrer, R.A., Henk, D.A., Garner, T.W.J., Balloux, F., Woodhams, D.C., Fisher, M.C.: Chromosomal Copy Number Variation, Selection and Uneven Rates of Recombination Reveal Cryptic Genome Diversity Linked to Pathogenicity. PLoS Genetics **9**(8), e1003703 (aug 2013). https://doi.org/10.1371/journal.pgen.1003703, http://dx.plos.org/10.1371/journal.pgen.1003703

[16] Favero, F., Joshi, T., Marquard, A.M., Birkbak, N.J., Krzystanek, M., Li, Q., Szallasi, Z., Eklund, A.C.: Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. Annals of Oncology **26**(1), 64–70 (jan 2015). https://doi.org/10.1093/annonc/mdu479, http://www.ncbi.nlm.nih.gov/pubmed/25319062

[17] Fox, D.T., Soltis, D.E., Soltis, P.S., Ashman, T.L., Van de Peer, Y.: Polyploidy: A biological force from cells to ecosystems. Trends in Cell Biology **30**(9), 688–694 (Sep 2020). https://doi.org/10.1016/j.tcb.2020.06.006, https://doi.org/10.1016/j.tcb.2020.06.006

[18] Fu, C., Davy, A., Holmes, S., Sun, S., Yadav, V., Gusa, A., Coelho, M.A., Heitman, J.: Dynamic genome plasticity during unisexual reproduction in the human fungal pathogen cryptococcus deneoformans. PLOS Genetics **17**(11), 1–31 (11 2021). https://doi.org/10.1371/journal.pgen.1009935, https://doi.org/10.1371/journal.pgen.1009935

[19] Fumagalli, M., Vieira, F.G., Linderoth, T., Nielsen, R.: ngsTools: methods for population genetics analyses from next-generation sequencing data. Bioinformatics **30**(10), 1486–1487 (May 2014)

[20] Fumagalli, M.: Assessing the effect of sequencing depth and sample size in population genetics inferences. PLOS ONE **8**(11), 1–11 (11 2013). https://doi.org/10.1371/journal.pone.0079667, https://doi.org/10.1371/journal.pone.0079667

[21] Fumagalli, M., Vieira, F.G., Korneliussen, T.S., Linderoth, T., Huerta-Sánchez, E., Albrechtsen, A., Nielsen, R.: Quantifying population genetic differentiation from next-generation sequencing data. Genetics **195**(3), 979–992 (2013). https://doi.org/10.1534/genetics.113.154740, https://www.genetics.org/content/195/3/979

[22] Fumagalli, M., Vieira, F.G., Linderoth, T., Nielsen, R.: ngsTools: methods for population genetics analyses from next-generation

sequencing data. Bioinformatics **30**(10), 1486–1487 (01 2014). https://doi.org/10.1093/bioinformatics/btu041, https://doi.org/10.1093/bioinformatics/btu041

[23] Garrison, E., Marth, G.: Haplotype-based variant detection from short-read sequencing (2012)

[24] Hardy, G.H.: Mendelian Proportions in a Mixed Population. Science, New Series **28**(706), 49–50 (1908)

[25] Jacqueline, K.W., Anna, P., Nancy, J.C.: Rational Inferences about Departures from Hardy-Weinberg Equilibrium. The American Journal of Human Genetics (6), 967–986 (2005)

[26] Lachance, J.: Detecting selection-induced departures from hardy-weinberg proportions. Genetics Selection Evolution (1), 15 (2009)

[27] Levy, S.E., Myers, R.M.: Advancements in next-generation sequencing. Annual Review of Genomics and Human Genetics **17**(1), 95–115 (2016). https://doi.org/10.1146/annurev-genom-083115-022413, https://doi.org/10.1146/annurev-genom-083115-022413, pMID: 27362342

[28] Li, C., Biswas, G.: Temporal Pattern Generation Using Hidden Markov Model Based Unsupervised Classification. In: IDA 1999: Advances in Intelligent Data Analysis, pp. 245–256. Springer, Berlin, Heidelberg (1999). https://doi.org/10.1007/3-540-48412-$4_2$1

[29] Li, H., J, R., Durbin, R.: Mapping short dna sequencing reads and calling variants using mapping quality scores. Genome research **11**(18), 1851–1858 (2008). https://doi.org/10.1101/gr.078212.108

[30] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., Subgroup, .G.P.D.P.: The Sequence Alignment/Map format and SAMtools. Bioinformatics **25**(16), 2078–2079 (06 2009). https://doi.org/10.1093/bioinformatics/btp352, https://doi.org/10.1093/bioinformatics/btp352

[31] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., Subgroup, .G.P.D.P.: The Sequence Alignment/Map format and SAMtools. Bioinformatics **25**(16), 2078–2079 (06 2009). https://doi.org/10.1093/bioinformatics/btp352, https://doi.org/10.1093/bioinformatics/btp352

[32] Loftus, B.J., Fung, E., Roncaglia, P., Rowley, D., Amedeo, P., Bruno, D., Vamathevan, J., Miranda, M., Anderson, I.J., Fraser, J.A., Allen, J.E., Bosdet, I.E., Brent, M.R., Chiu, R., Doering, T.L., Donlin, M.J., D'Souza, C.A., Fox, D.S., Grinberg, V., Fu, J., Fukushima, M., Haas, B.J., Huang, J.C., Janbon, G., Jones, S.J.M., Koo, H.L., Krzywinski, M.I., Kwon-Chung, J.K., Lengeler, K.B., Maiti, R., Marra, M.A., Marra, R.E., Mathewson, C.A., Mitchell, T.G., Pertea, M., Riggs, F.R., Salzberg, S.L., Schein, J.E., Shvartsbeyn, A., Shin, H., Shumway, M., Specht, C.A., Suh, B.B., Tenney, A., Utterback, T.R., Wickes, B.L., Wortman, J.R., Wye, N.H., Kronstad, J.W., Lodge, J.K., Heitman, J., Davis, R.W., Fraser, C.M., Hyman, R.W.: The genome of the basidiomycetous yeast and human pathogen ¡i¿cryptococcus neoformans¡/i¿. Science **307**(5713), 1321–1324 (2005). https://doi.org/10.1126/science.1103773

[33] Longley, N., Muzoora, C., Taseera, K., Mwesigye, J., Rwebembera, J., Chakera, A., Wall, E., Andia, I., Jaffar, S., Harrison, T.S.: Dose Response Effect of High-Dose Fluconazole for HIV-Associated Cryptococcal Meningitis in Southwestern Uganda. Clinical Infectious Diseases **47**(12), 1556–1561 (12 2008). https://doi.org/10.1086/593194, https://doi.org/10.1086/593194

[34] Lou, R.N., Jacobs, A., Wilder, A., Therkildsen, N.O.: A beginner's guide to low-coverage whole genome sequencing for population genomics. Molecular Ecology **n/a**(n/a) (2021). https://doi.org/https://doi.org/10.1111/mec.16077, https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.16077

[35] May, R.C., Stone, N.R., Wiesner, D.L., Bicanic, T., Nielsen, K.: Cryptococcus: from environmental saprophyte to global pathogen. Nature Reviews Microbiology **14**(2), 106–117 (Feb 2016). https://doi.org/10.1038/nrmicro.2015.6, https://doi.org/10.1038/nrmicro.2015.6

[36] McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., DePristo, M.A.: The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research **20**(9), 1297–303 (sep 2010). https://doi.org/10.1101/gr.107524.110, http://www.ncbi.nlm.nih.gov/pubmed/20644199

[37] McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., DePristo, M.A.: The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research **20**(9), 1297–303 (sep 2010). https://doi.org/10.1101/gr.107524.110, http://www.ncbi.nlm.nih.gov/pubmed/20644199

[38] Metzker, M.L.: Sequencing technologies — the next generation. Nature Reviews Genetics **11**(1), 31–46 (jan 2010). https://doi.org/10.1038/nrg2626, http://www.ncbi.nlm.nih.gov/pubmed/19997069

[39] Morrow, C.A., Fraser, J.A.: Ploidy variation as an adaptive mechanism in human pathogenic fungi. Seminars in Cell and Developmental Biology **24**(4), 339–346 (apr 2013)

[40] Nielsen, R., Paul, J., Albrechtsen, A., Song, Y.: Genotype and snp calling from next-generation sequencing data. Nature Reviews. Genetics **12**(6), 443–451 (2011). https://doi.org/10.1038/nrg2986

[41] Ormerod, K.L., Morrow, C.A., Chow, E.W.L., Lee, I.R., Arras, S.D.M., Schirra, H.J., Cox, G.M., Fries, B.C., Fraser, J.A.: Comparative genomics of serial isolates of cryptococcus neoformans reveals gene associated with carbon utilization and virulence. G3 (Bethesda, Md.) **3**(4), 675–686 (Apr 2013). https://doi.org/10.1534/g3.113.005660, https://pubmed.ncbi.nlm.nih.gov/23550133, 23550133[pmid]

[42] Van de Peer, Y., Mizrachi, E., Marchal, K.: The evolutionary significance of polyploidy. Nature Reviews Genetics **18**(7), 411–424 (Jul 2017). https://doi.org/10.1038/nrg.2017.26, https://doi.org/10.1038/nrg.2017.26

[43] Rabiner, L.: A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE **77**(2), 257–286 (1989). https://doi.org/10.1109/5.18626, http://ieeexplore.ieee.org/document/18626/

[44] Rhodes, J., Beale, M.A., Fisher, M.C.: Illuminating choices for library prep: A comparison of library preparation methods for whole genome sequencing of cryptococcus neoformans using illumina hiseq. PLOS ONE **9**(11), 1–9 (11 2014). https://doi.org/10.1371/journal.pone.0113501, https://doi.org/10.1371/journal.pone.0113501

[45] Rhodes, J., Beale, M.A., Vanhove, M., Jarvis, J.N., Kannambath, S., Simpson, J.A., Ryan, A., Meintjes, G., Harrison, T.S., Fisher, M.C., Bicanic, T.: A Population Genomics Approach to Assessing the Genetic Basis of Within-Host Microevolution Underlying Recurrent Cryptococcal Meningitis Infection. G3 Genes—Genomes—Genetics (2017). https://doi.org/10.1534/g3.116.037499, https://doi.org/10.1534/g3.116.037499

[46] Rhodes, J., Desjardins, C.A., Sykes, S.M., Beale, M.A., Vanhove, M., Sakthikumar, S., Chen, Y., Gujja, S., Saif, S., Chowdhary, A., Lawson, D.J., Ponzio, V., Colombo, A.L., Meyer, W., Engelthaler, D.M., Hagen, F., Illnait-Zaragozi, M.T., Alanio, A., Vreulink, J.M., Heitman, J., Perfect, J.R., Litvintseva, A.P., Bicanic, T., Harrison, T.S., Fisher, M.C., Cuomo, C.A.: Tracing Genetic Exchange and Biogeography of Cryptococcus neoformans var. grubii at the Global Population Level. Genetics **207**(1), 327–346 (07 2017). https://doi.org/10.1534/genetics.117.203836, https://doi.org/10.1534/genetics.117.203836

[47] Sattler, M.C., Carvalho, C.R., Clarindo, W.R.: The polyploidy and its key role in plant breeding. Planta (243), 281–296 (2016)

[48] Sionov, E., Chang, Y.C., Kwon-Chung, K.J.: Azole heteroresistance in cryptococcus neoformans: Emergence of resistant clones with chromosomal disomy in the mouse brain during fluconazole treatment. Antimicrobial Agents and Chemotherapy **57**(10), 5127–5130 (2013). https://doi.org/10.1128/AAC.00694-13, https://journals.asm.org/doi/abs/10.1128/AAC.00694-13

[49] Stone, N.R., Rhodes, J., Fisher, M.C., Mfinanga, S., Kivuyo, S., Rugemalila, J., Segal, E.S., Needleman, L., Molloy, S.F., Kwon-Chung, J., Harrison, T.S., Hope, W., Berman, J., Bicanic, T.: Dynamic ploidy changes drive fluconazole resistance in human cryptococcal meningitis. Journal of Clinical Investigation **129**(3), 999–1014 (mar 2019)

[50] Therkildsen, N.O., Palumbi, S.R.: Practical low-coverage genomewide sequencing of hundreds of individually barcoded samples for population and evolutionary genomics in nonmodel species. Molecular Ecology Resources **17**(2), 194–208 (2017). https://doi.org/https://doi.org/10.1111/1755-0998.12593, https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.12593

[51] Viterbi, A., A.: Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Transactions on Information The-

ory **13**(2), 260–269 (apr 1967). https://doi.org/10.1109/TIT.1967.1054010, http://ieeexplore.ieee.org/document/1054010/

[52] Vu, G.T.H., Cao, H.X., Reiss, B., Schubert, I.: Deletion-bias in dna double-strand break repair differentially contributes to plant genome shrinkage. New Phytologist **214**(4), 1712–1721 (2017). https://doi.org/https://doi.org/10.1111/nph.14490, https://nph.onlinelibrary.wiley.com/doi/abs/10.1111/nph.14490

[53] Weinberg, W.: Über den Nachweis der Vererbung beim Menschen. Jahresh. Ver. Vaterl. Naturkd. Württemb. **64**, 369–382 (1908)

[54] Weiß, C.L., Pais, M., Cano, L.M., Kamoun, S., Burbano, H.A.: nQuire: a statistical framework for ploidy estimation using next generation sequencing (2018). https://doi.org/10.1186/s12859-018-2128-z, https://doi.org/10.1186/s12859-018-2128-z

[55] Wood, T.E., Takebayashi, N., Barker, M.S., Mayrose, I., Greenspoon, P.B., H, R.L.: The frequency of polyploid speciation in vascular plants. Proc Natl Acad Sci USA (106), 13875–13879 (2009)

[56] Yang, F., Gritsenko, V., Lu, H., Zhen, C., Gao, L., Berman, J., ying Jiang, Y., Alanio, A.: Adaptation to fluconazole via aneuploidy enables cross-adaptation to amphotericin b and flucytosine in cryptococcus neoformans. Microbiology Spectrum **9**(2), e00723–21 (2021). https://doi.org/10.1128/Spectrum.00723-21, https://journals.asm.org/doi/abs/10.1128/Spectrum.00723-21

[57] Zhu, J., Tsai, H.J., Gordon, M.R., Li, R.: Cellular Stress Associated with Aneuploidy. Developmental cell **44**(4), 420–431 (feb 2018). https://doi.org/10.1016/j.devcel.2018.02.002, http://www.ncbi.nlm.nih.gov/pubmed/29486194

1 Supplementary Material

1.1 Supplementary Methods

This section describes supplementary information on methods used for the implementation of `HMMploidy`. $O = (O_1, \ldots, O_M)$ is the observed NGS data for M sequenced genomes at N polymorphic sites. For each m-th genome and n-th locus we define $Y_{m,n}$, $G_{m,n}$ and $O_{m,n}$ as the ploidy, genotype and sequencing data, respectively. $G_{m,n}$ assumes values in $\{G_0, G_1, ..., G_{Y_{m,n}}\}$, where each G_i represents the i-th genotype as a collection $\{G_i^1, \ldots, G_i^{Y_{m,n}+1}\}$ of its nucleotides, with $|G_i|$ defined as the number of alternate (or derived) alleles of the genotype.

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 (37).

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 (14) 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 (37) and mostly resemble the approach of (29), 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 a 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` (31). 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 constant Phred error:

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

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

1.4 Hidden Markov Model for ploidy inference

Here, the HMM is defined, and the inferential process of ploidy levels from the HMM is illustrated. Mathematical details, proofs and algorithm analysis of the HMM for ploidy inference (Fig. S1) are presented here.

The Markov chain of ploidy levels is characterised by a $|\mathcal{Y}| \times |\mathcal{Y}|$ transition matrix $\boldsymbol{A}$, and a $|\mathcal{Y}|$-long vector $\boldsymbol{\delta}$ of starting probabilities for the first latent state. Here, $\mathcal{Y}$ is the set of ploidy levels included in the model, and $|\mathcal{Y}|$ is its cardinality. The average depth for genome m in window k is characterised by the ploidy-dependent parameters $\alpha_{Y_m^{(k)}}, \beta_{Y_m^{(k)}} \in \mathbb{R}$, describing mean and dispersion of the data, for each $Y_m^{(k)} \in \mathcal{Y}$. For brevity we write the parameters of the depth distribution in vector form, i.e. $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$. The allele frequencies calculated through Equation (2) in the k-th window of loci serve as a proxy for the probability of sequenced reads.

Heuristic Expectation Conditional Maximization (HECM) The ECM algorithm is used to infer the parameters $\boldsymbol{A}$, $\boldsymbol{\delta}$ modelling the sequence of ploidy levels. This is done in two iterative steps by exploiting the ploidy-dependent distributions of the observed data (sequenced reads and coverage in each window). The first step is the well-known forward-backward algorithm (43; 8), that computes in each window the probability of a ploidy given all the observed data. This is done in an efficient way through dynamic programming and exploitation of Markov properties with computational complexity $\mathcal{O}(|\mathcal{Y}|^2 K)$ (i.e. linear in the number of loci windows) by implementing two calculation sweeps, starting respectively at the end and at the beginning of the observation sequence.

The forward-backward algorithm thus creates the mathematical link between ploidy levels and observed data, and allows us to update the parameters governing the Markov chain of ploidy levels with the ECM algorithm in a subsequent step. The ECM algorithm maximizes a value (called intermediate quantity) strictly related to the likelihood of the model, where the free variables of the maximization are the matrix $\boldsymbol{A}$, the vector $\boldsymbol{\delta}$ and the parameters of the distribution of observed data. This procedure continues iteratively by recalculating the forward-backward posteriors and the update parameters with the ECM, until the intermediate quantity cannot be further improved.

We start by illustrating the steps of the ECM algorithm, and subsequently adding the heuristic procedure. For ease of notation, λ is the tuple of parameters $(\boldsymbol{A}, \boldsymbol{\delta}, \boldsymbol{\beta}, \boldsymbol{\alpha}) \in \Lambda$, considered as two separate tuples: $\lambda = (\lambda_1, \lambda_2) = ((\boldsymbol{A}, \boldsymbol{\delta}, \boldsymbol{\beta}), (\boldsymbol{\alpha})) \in \Lambda_1 \times \Lambda_2$. Given the parameters $\lambda^{\ell-1}$ calculated at the $(\ell-1)$-th step of the ECM, the ℓ-th iteration to calculate λ^{ℓ} follows the steps below:

1. calculate the intermediate quantity
$$Q(\lambda_1^{\ell} | \lambda^{\ell-1}) = \mathbb{E}\Big[\ln p\big(O_m^{(1:K)}, Y_m^{(1:K)} | (\lambda_1^{\ell}, \lambda_2^{\ell-1})\big) \,\Big|\, O_m^{(1:K)}, \lambda^{\ell-1} \Big];$$
2. calculate $\lambda_1^{\ell} = \arg_{\lambda_1^{\ell} \in \Lambda_1} \max Q(\lambda_1^{\ell} | \lambda^{\ell-1}))$;
3. calculate the intermediate quantity $Q(\lambda_2^{\ell} | (\lambda_1^{\ell}, \lambda_2^{\ell-1}))$ analogously to step 1;
4. calculate $\lambda_2^{\ell} = \arg_{\lambda_2^{\ell} \in \Lambda_2} \max Q((\lambda_2^{\ell}) | (\lambda_1^{\ell}, \lambda_2^{\ell}))$.

Here, we used $O_m^{(1:K)}$, $Y_m^{(1:K)}$ to denote $O_m^1, \ldots, O_m^K$ and $Y_m^1, \ldots, Y_m^K$, respectively.

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 ($\boldsymbol{\beta}$) parameters unknown and to be estimated at the next M-step (maximization step), conditionally to the mean ($\boldsymbol{\alpha}$) 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. 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 (7; 43). The E-step expresses the expected complete-data log-likelihood with the 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). The average depths are modelled with a negative binomial distribution to take data overdispersion into account (11).

The calculation of $\boldsymbol{A}, \boldsymbol{\delta}$ at iteration ℓ is solved by using the classical forward-backward algorithm (43; 8), 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 feasible, since expressing the log-likelihood directly is not concretely achievable.

The intermediate quantity can be explicitly written as the sum of three terms involving separately the matrix $\boldsymbol{A}$, the vector $\boldsymbol{\delta}$ and the vectors $\boldsymbol{\alpha}, \boldsymbol{\beta}$:

$$Q(\lambda_1^\ell|\lambda^{\ell-1}) = \sum_{Y_m^{(1:K)} \epsilon \mathcal{Y}} \ln(\delta^\ell_{Y_m^{(1)}}) p(Y_m^{(1:K)}|O_m^{(1:K)}, C_m^{(1:k)}, \lambda^{\ell-1}) \tag{3}$$

$$+ \sum_{Y_m^{(1:K)} \epsilon \mathcal{Y}} \sum_{k=2}^{K} \ln(\boldsymbol{A}^\ell_{Y_m^{(k-1)} Y_m^{(k)}}) p(Y_m^{(1:K)}|O_m^{(1:K)}, C_m^{(1:k)}, \lambda^{\ell-1}) \tag{4}$$

$$+ \sum_{Y_m^{(1:K)} \epsilon \mathcal{Y}} \sum_{k=1}^{K} \left(\ln\left(p(O_m^{(k)}|Y_m^{(k)}, F^{(k)})\right) + \ln\left(p(C_m^{(k)}|Y_m^{(k)}, \alpha^{\ell-1}_{Y_m^{(k)}}, \beta^\ell_{Y_m^{(k)}})\right)\right)$$

$$p(Y_m^{(1:K)}|O_m^{(1:K)}, C_m^{(1:k)}, \lambda^{\ell-1}) \tag{5}$$

where the logarithm of a matrix is intended element-wise.

Consider the (m,k)-th forward variable defined by

$$f(y_m^{(k)}) = p(O_m^{(1:k)}, C_m^{(1:k)}, Y_m^{(k)} = y_m^{(k)}|\lambda),$$

that is, the probability of the first k observations and k-th ploidy $y_m^{(k)}$ given the parameters λ. Define the (m,k)-th backward variable as

$$b(y_m^{(k)}) = p(O_m^{(k+1:K)}, C_m^{(k+1:K)}|Y_m^{(k)} = y_m^{(k)}, \lambda),$$

that is, the probability of the latest $(K-k)$ observations, given the k-th ploidy $y_m^{(k)}$ and the parameters λ. The forward and backward variables can be computed with an iterative procedure (43, eq. 19,20,24,25) and allow us to calculate efficiently the likelihood of the data as

$$p(O_m^{(1:K)}, C_m^{(1:k)}|\lambda) = \sum_{y_m^{(k)} \in \mathcal{Y}} f(y_m^{(k)}) b(y_m^{(k)}) \quad \text{for any } k = 1, \ldots, K.$$

The two terms in the equation lines (3) and (4) include only the parameters $\boldsymbol{\delta}$ and $\boldsymbol{A}$, respectively. This simplifies finding optimisation formulae for those parameters by considering separately each term of lines (3) and (4). Such optimisation equations for $\boldsymbol{\delta}$ and $\boldsymbol{A}$ are easily derived through Lagrange multipliers (43, eq. 40a,40b). This does not solve the second step of the ECM algorithm, because the optimum for $\boldsymbol{\beta}$ is still not calculated.

It is easy to see that both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ concur in defining line (5). This is what originates the conditional nature of the ECM algorithm, i.e. $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ cannot be optimised independently. Therefore we first optimise $\boldsymbol{\beta}$ considering the values of $\boldsymbol{\alpha}$ calculated at the $(\ell-1)$-th iteration of the ECM algorithm. Using the forward and backward variables, and excluding terms independent from the Poisson-Gamma parameters, Equation (5) can be written as follows:

$$\begin{aligned}
&\sum_{Y_m^{(1:K)} \in \mathcal{Y}} \sum_{k=1}^{K} u(m,k) \ln\left(p(C_m^{(k)}|Y_m^{(k)}, \alpha^{\ell-1}_{Y_m^{(k)}}, \beta^{\ell}_{Y_m^{(k)}})\right) \\
&= \sum_{Y_m^{(1:K)} \in \mathcal{Y}} \sum_{k=1}^{K} u(m,k) \ln\left(\frac{\Gamma(\alpha^{\ell-1}_{Y_m^{(k)}} + C_m^{(k)})}{\Gamma(C_m^{(k)}+1)\Gamma(\alpha^{\ell-1}_{Y_m^{(k)}})}\right) \\
&+ \sum_{Y_m^{(1:K)} \in \mathcal{Y}} \sum_{k=1}^{K} u(m,k)\left(C_m^{(k)} \ln\left(\frac{1}{\beta^{\ell}_{Y_m^{(k)}}+1}\right) + \alpha^{\ell-1}_{Y_m^{(k)}} \ln\left(\frac{\beta^{\ell}_{Y_m^{(k)}}}{\beta^{\ell}_{Y_m^{(k)}}+1}\right)\right)
\end{aligned}$$

where $u^{\ell}(m,k) = f(y_m^{(k)})b(y_m^{(k)})/p(O_m^{(1:K)}, C_m^{(1:k)}|\lambda^{\ell})$. By setting the partial derivative of $Q(\lambda_1^{\ell}|\lambda^{\ell-1})$ w.r.t. a certain $\beta^{\ell}_{y_m^{(k)}}$, $y_m^{(k)} \in \mathcal{Y}$ equal to zero as below, we will be able to calculate the optimum for the derivative's parameter:

$$\frac{\partial Q(\lambda_1^{\ell}|\lambda^{\ell-1})}{\partial \beta^{\ell}_{y_m^{(k)}}} = \sum_{k=1}^{K} -u(m,k)\frac{C_m^{(k)}}{\beta^{\ell}_{y_m^{(k)}}+1} + \sum_{k=1}^{K} u(m,k)\frac{\alpha^{\ell-1}_{Y_m^{(k)}}}{\beta^{\ell}_{y_m^{(k)}}\left(\beta^{\ell}_{y_m^{(k)}}+1\right)} = 0.$$

Solving for $\beta^{\ell}_{y_m^{(k)}}$ leads to the optimum of the parameter:

$$\beta^{\ell}_{y_m^{(k)}} = \frac{\alpha^{\ell-1}_{Y_m^{(k)}} \sum_{k=1}^{K} u(m,k)}{\sum_{k=1}^{K} u(m,k) C_m^{(k)}}.$$

This completes the step 2 of the ECM. In our implementation of `HMMploidy`, we want to leverage the information contained in the genotype likelihoods, whose

partial derivative goes to zero and in principle are not integrated in the optimisation. In `HMMploidy`, we add the genotype likelihoods to the depth distribution prior to optimisation, so that forward and backward variables contain information on both depth and genotypes, and allow the identification of different states with distinct ploidy levels.

The value of $Q(\lambda_2^\ell|(\lambda_1^\ell, \lambda_2^{\ell-1}))$ can be easily calculated as in step 1, and by setting the partial derivative of $Q(\lambda_2^\ell|(\lambda_1^\ell, \lambda_2^{\ell-1}))$ w.r.t. $\alpha^\ell_{y_m^{(k)}}$, $y_m^{(k)} \in \mathcal{Y}$, to zero, we obtain:

$$\sum_{k=1}^{K} u(m,k)\left(\ln\left(\frac{\beta^\ell_{y_m^{(k)}}}{\beta^\ell_{y_m^{(k)}}+1}\right) + \psi_0\left(\alpha^\ell_{y_m^{(k)}} + C_m^{(k)}\right) - \psi_0\left(\alpha^\ell_{y_m^{(k)}}\right)\right) = 0.$$

Solving for $\alpha^\ell_{y_m^{(k)}}$ is done through the Newton-Rapson method (7), completing step 4 of the ECM.

Heuristic step and ploidy inference The ECM algorithm is repeated as an iterative sequence of EC and M steps, until the expected conditional log-likelihood of the model satisfies a convergence criteria. When convergence is achieved, `HMMploidy` performs the heuristic step, by running few iterations of the ECM over the HMM, where the set of ploidy levels is reduced by one, and the parameters for initialisation are the final ones from the ECM. We assume that, if the HMM has an overfitting set of ploidy levels, observation parameters are overlapping (28) for two or more ploidy levels. Therefore, removing one unnecessary ploidy requires only few extra iterations for the EM to converge again. The Bayesian Information Criterion (BIC) (7; 8) is used to compare the HMM with the reduced HMMs. If there is a reduced HMM with a better BIC score, then the ECM runs again on such HMM, otherwise it stops. Such method is an adaptation of the suggestion in (28).

After the HMM is reduced through the BIC comparison, we reduce the transition matrix between ploidy levels, i.e. we remove ploidy values for which there is almost zero probability of lasting a reasonable number of adjacent windows. In other words, we remove ploidy levels that will last for the length of one or few more windows of loci. Once the HMM parameters are determined through the heuristic sweep, the standard Viterbi algorithm (51) is applied to infer the most likely sequence of ploidy levels from the parameters of the HMM. The Viterbi algorithm is another example of dynamic programming, allowing to bypass the calculation of all possible $|\mathcal{Y}|^K$ sequences of ploidy levels to determine which one is the best.

Reduction of the transition matrix An important element of a HMM is the transition matrix between states and the meaning of each state. Thanks to the heuristic ECM, `HMMploidy` is able to assign a ploidy to each state of the Markov chain in an unsupervised mode without overfitting the data. However, one needs to check whether transitions between states follow a biological meaning. For

example, it is unlikely that a ploidy occurs only in a small window of loci, and then shifts again to the previous value, i.e. such event is likely due to noise or other biological artefact altering the quality and behaviour of the data (e.g. the presence of a centromere).

Once the HECM algorithm has converged to a set of parameters $\lambda \in \Lambda$, it is possible to perform an optional filtering on the transition matrix $\boldsymbol{A}$ of the HMM. Given the matrix $\boldsymbol{A}$ of size $|\mathcal{Y}| \times |\mathcal{Y}|$, the time of permanence in a state $y \in \mathcal{Y}$ has geometric distribution with parameter $\boldsymbol{A}_{y,y}$ (9). If the user expects that a ploidy level has to remain uninterrupted for at least a certain number of windows N, then a corresponding minimum value for the parameter of the geometric distribution can be estimated. In fact, the probability of permanence in ploidy y for at least $N > 0$ windows is given by the cumulative distribution function of the geometric distributions, that is, $1 - (1 - y)^N$.

Given N, `HMMploidy` calculates the minimum value of y that has to be on the diagonal of $\boldsymbol{A}$. Rows and columns corresponding to diagonal entries lower than y are cancelled and $\boldsymbol{A}$ is rendered stochastic again. Corresponding values of δ, α, β are also removed. Afterwards, the HMM optimisation is performed again on the new subset of ploidy levels for adjustment of the remaining parameters.

Application in presence of sparse polymorphic sites Given an individual m, consider each k-th window of its genome. In presence of very few polymorphic sites in each window, the genotype likelihoods might not be enough to determine the ploidy, especially when the data is at low-depth and in presence of error, as it is often the case with high-throughput data.

To consider this case, the option `useGeno` is added to the software. When `useGeno='yes'`, the HMM infers the ploidy numbers as explained in the main text. Otherwise, if When `useGeno='no'`, only the sequencing depth data is used to infer the hidden states of a binomial HMM initially. This allows to consider the largest possible windows of loci. Each latent state is then assigned a ploidy by maximising Equation (1) over all the windows with same hidden state.

1.5 Results from the analysis of *Cryptococcus Neoformans*

Here we present all the inferred ploidy levels from the 23 isolates of *Cryptococcus Neoformans* from the original study (45). Each figure contains:

- In the first line, inferred ploidy levels from chromosome 1 and 12 using the full data,
- In the second line, inferred ploidy levels from chromosome 1 and 12 using 20% of the original sequencing data.

Most of the results from the downsampled data coincide with the inference from the whole data. Higher ploidy levels can be hard to detect in some cases, and are occasionally detected as a constant lower ploidy or as a highly varying sequence of adjacent ploidy levels. However, downsampling seems to recover a constant haploid chromosome 12 in sample cctp50 (Fig. S12B-D) according to what the

sequencing depth indicates. This means that downsampling might reduce the effect of noisy data points that could alter the detected ploidy. In fact, the triploid sections of chromosome 12 are at the extremities of the chromosome, where the data is more affected by noise and in general by a lower sequencing quality.

All the other samples recover successfully the original ploidy levels in downsampled data. However, note that there are few changes in ploidy probably due to noise or the presence of reads close to the centromere (Fig. S21, S16, S15, S13, S14). Ploidy levels for *Cryptococcus Neoformans* samples inferred by competing methods `nQuire`, `nQuire.Den`, and `ploidyNGS` are presented in Supplementary Table 1.

1.6 Supplementary Figures

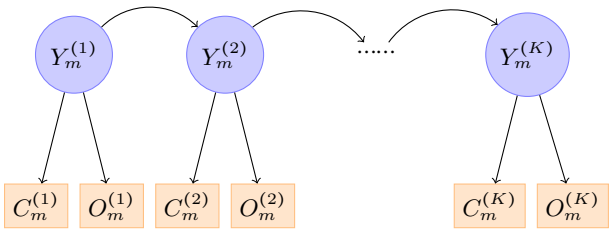

Figure S 1**: Hidden Markov Model for ploidy inference.** Graphical representation of the HMM to infer the ploidy levels of the m-th genome. $Y_m^{(k)}$ is the ploidy of the k-th window of genome m. The ploidy-dependent emissions consist of the average sequencing depth $C_m^{(k)}$ and the sequenced data $O_m^{(k)}$, whose distributions are respectively described by a Poisson-gamma distribution and by Equation (1).

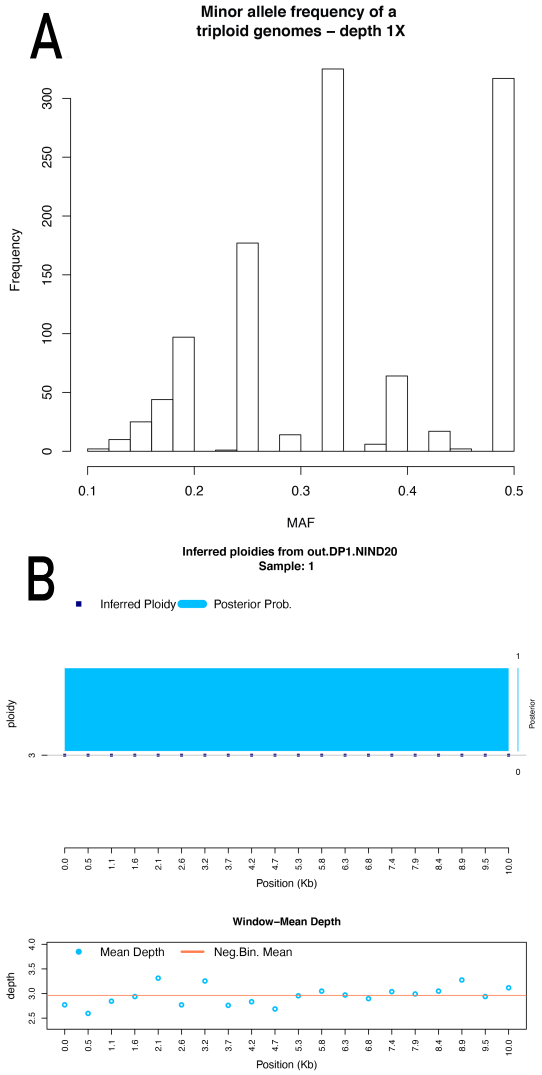

Figure S 2**: Histograms of minor allele frequencies and inferred ploidy with `HMMploidy` at low depth.** (A) Distribution of the minor allele frequencies of one simulated triploid genomes (out of a sample of 20 individuals) of 10kbp at depth $1X$ for the haploid state. It is not trivial to determine the ploidy by visual inspection of this graph. (B) Inferred ploidy with `HMMploidy` from the same individual on windows of 500 bases. Using the information contained in all the other individuals, it is possible to infer the correct ploidy.

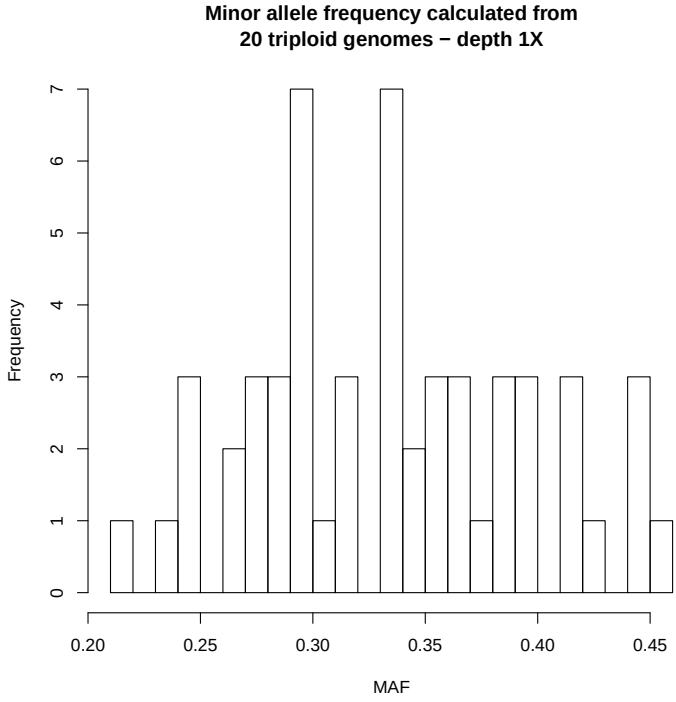

Figure S 3**: Histograms of minor allele frequencies for many samples at low depth.** Histogram of the estimated minor allele frequency for 20 simulated triploid individuals in a window of 500 bases. The distribution is closer to the one expected for a triploid individual, but it is still not possible to infer the ploidy by a simple visual inspection of the graph. The use of genotype likelihoods in `HMMploidy` supplies additional information to infer the correct ploidy.

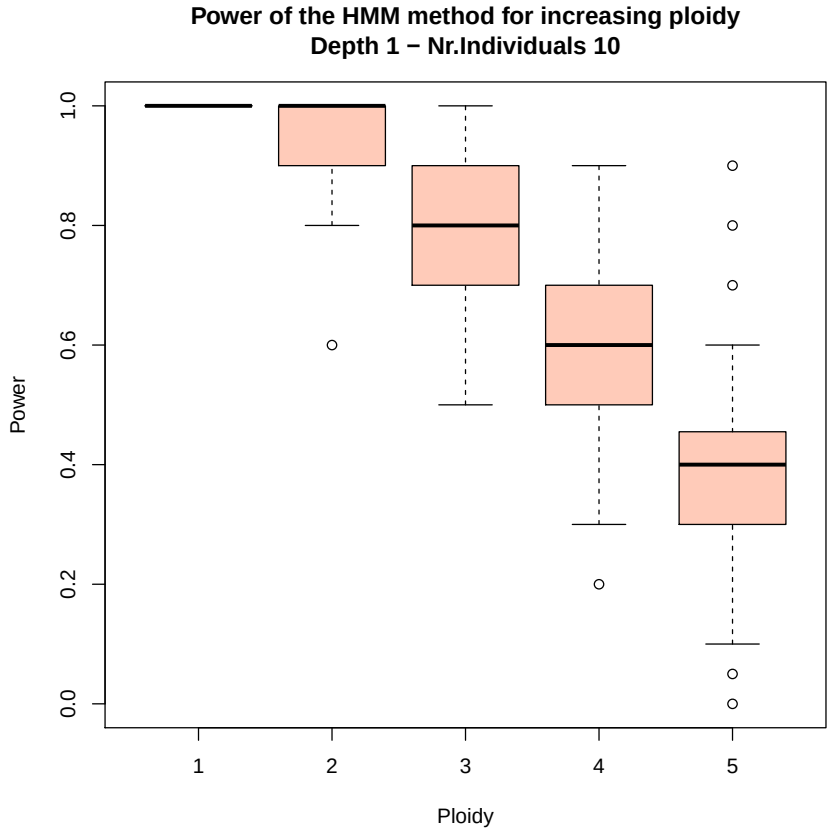

Figure S 4**: Relationship between ploidy levels and detection rate.** Power of `HMMploidy` to detect the correct ploidy level (on y-axis) on simulated genomes with increasing ploidy (on x-axis) from one to six at depth $1X$. The power decreases with higher ploidy numbers because genotype likelihoods lack information to characterise correct genotypes.

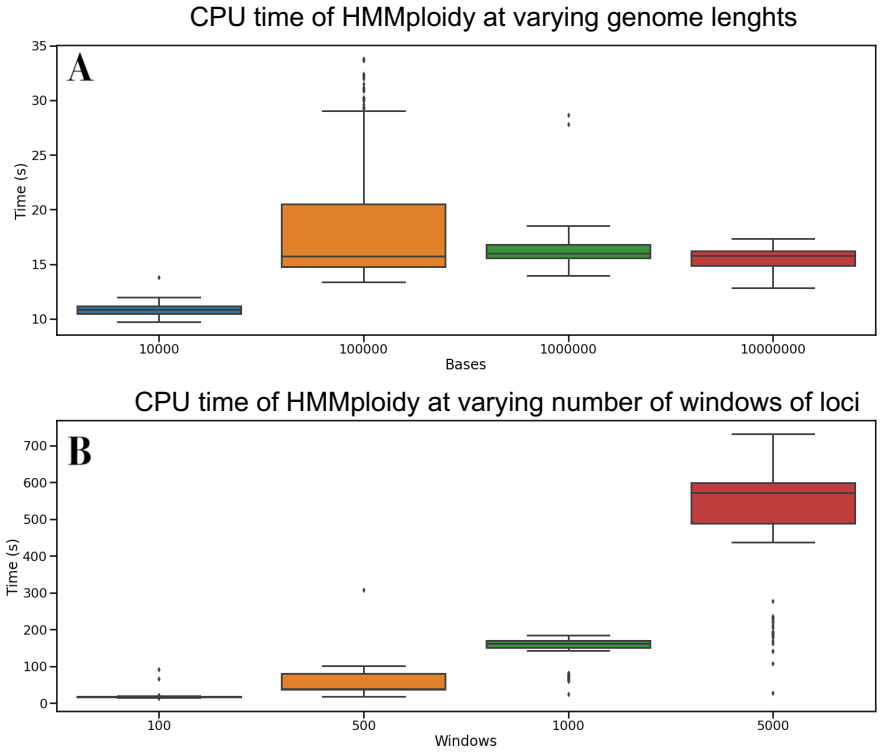

Figure S 5**: CPU running time for** `HMMploidy`**.** (A) CPU running time of `HMMploidy` by simulating genomes of various lengths and keeping the windows number to 100. The time is quite constant, meaning that the loading and processing of the data is very fast, and most of the time is taken by the HMM inference. (B) CPU running time of `HMMploidy` by increasing the number of windows on a $10MB$ genome. The time grows accordingly with K in an almost linear fashion (due to a probable overhead for preprocessing the data in many windows), as predicted by the computational cost of the forward-backward algorithm.

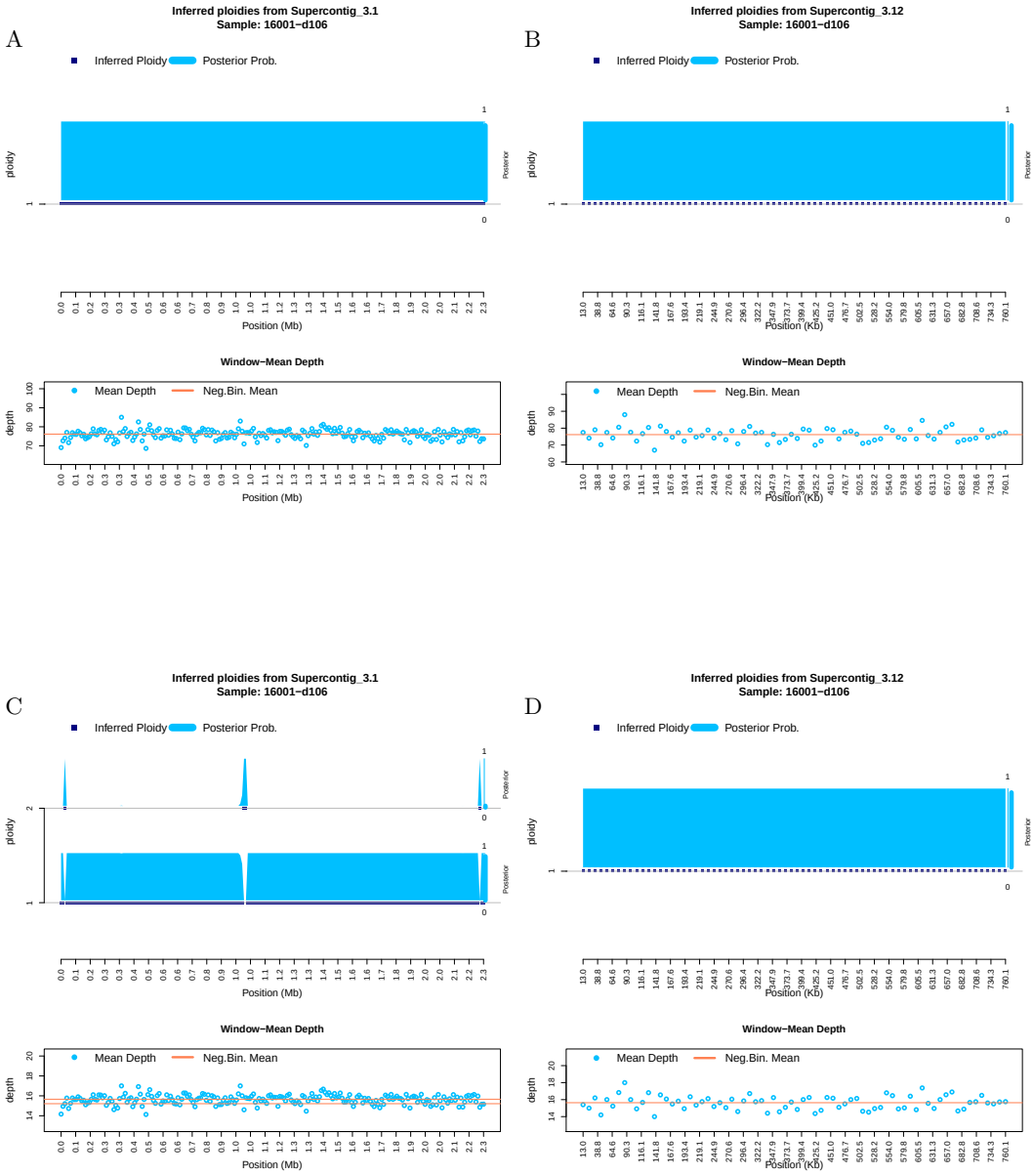

Figure S 6**: Ploidy inference on full and downsampled sequencing data.** Inferred ploidy levels from `HMMploidy` for chromosome 1 and 12 of isolate 16001-d106. **(A-B)** Results using the whole data on chromosomes 1 (A) and 12 (B). **(C-D)** Results using the data downsampled to 20% of its original depth on chromosomes 1 (C) and 12 (D).

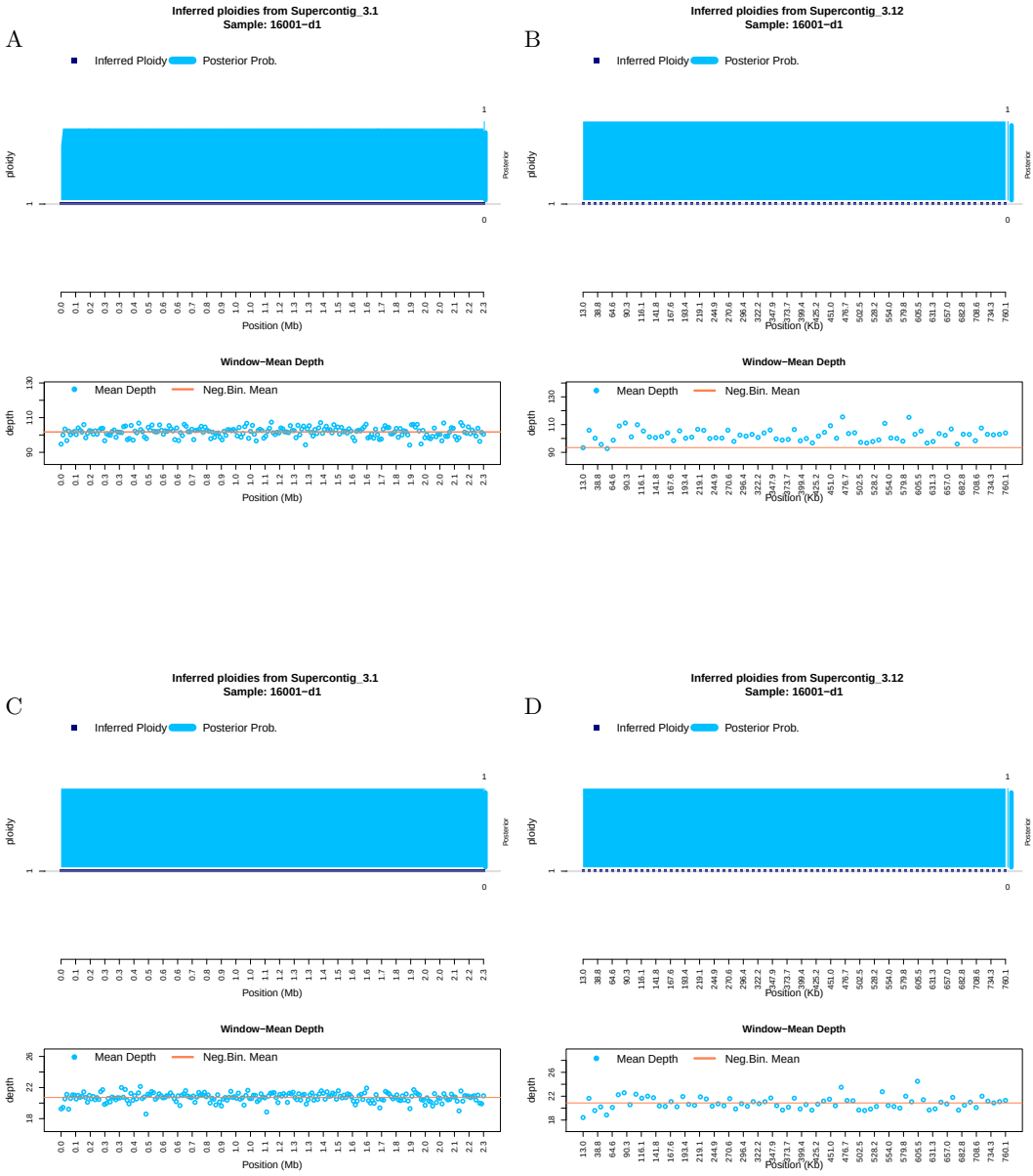

Figure S 7**: Ploidy inference on full and downsampled sequencing data.** Inferred ploidy levels from `HMMploidy` for chromosome 1 and 12 of isolate 16001-d1. **(A-B)** Results using the whole data on chromosomes 1 (A) and 12 (B). **(C-D)** Results using the data downsampled to 20% of its original depth on chromosomes 1 (C) and 12 (D).

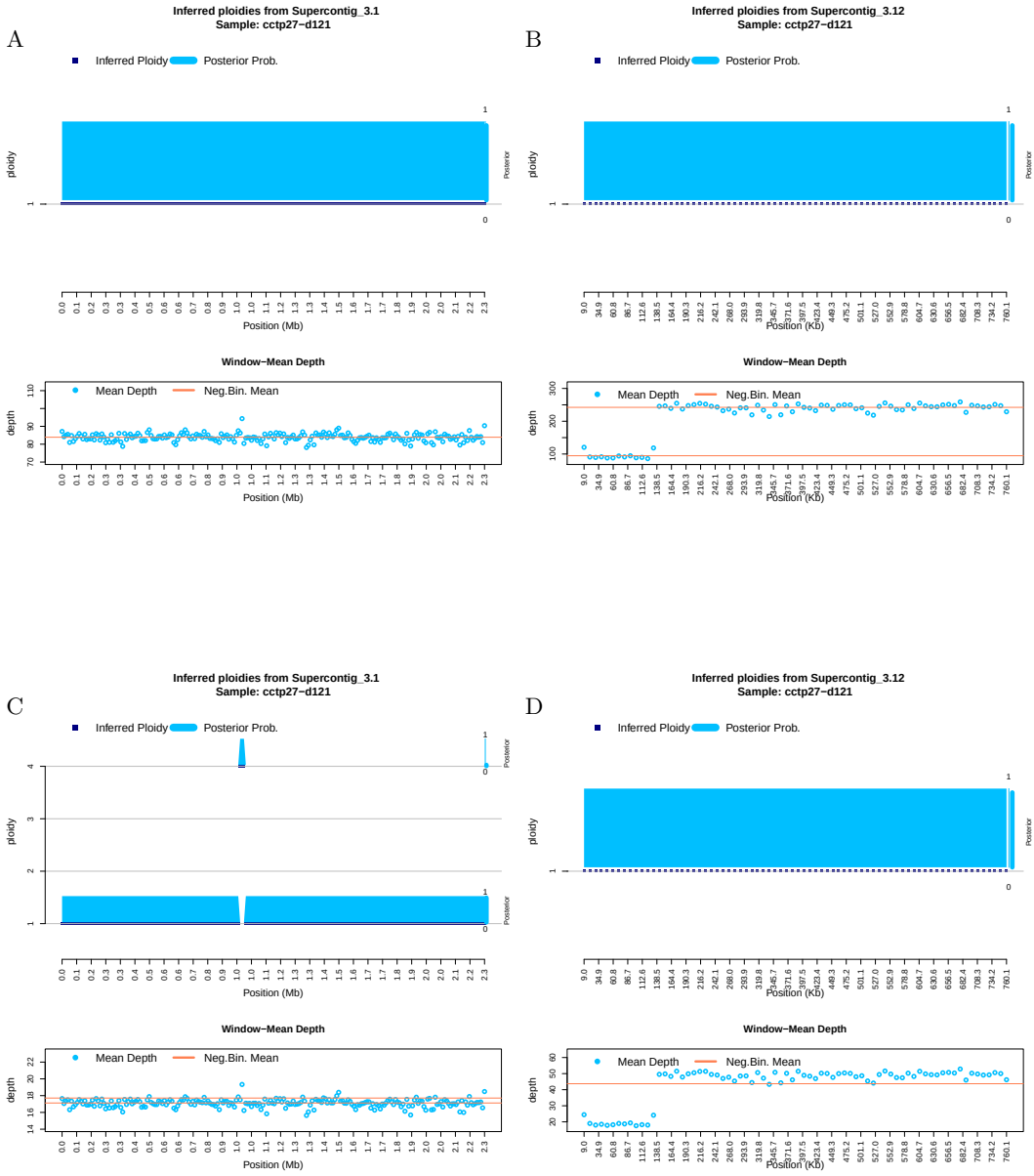

Figure S 8**: Ploidy inference on full and downsampled sequencing data.** Inferred ploidy levels from `HMMploidy` for chromosome 1 and 12 of isolate cctp27-d121. **(A-B)** Results using the whole data on chromosomes 1 (A) and 12 (B). **(C-D)** Results using the data downsampled to 20% of its original depth on chromosomes 1 (C) and 12 (D).

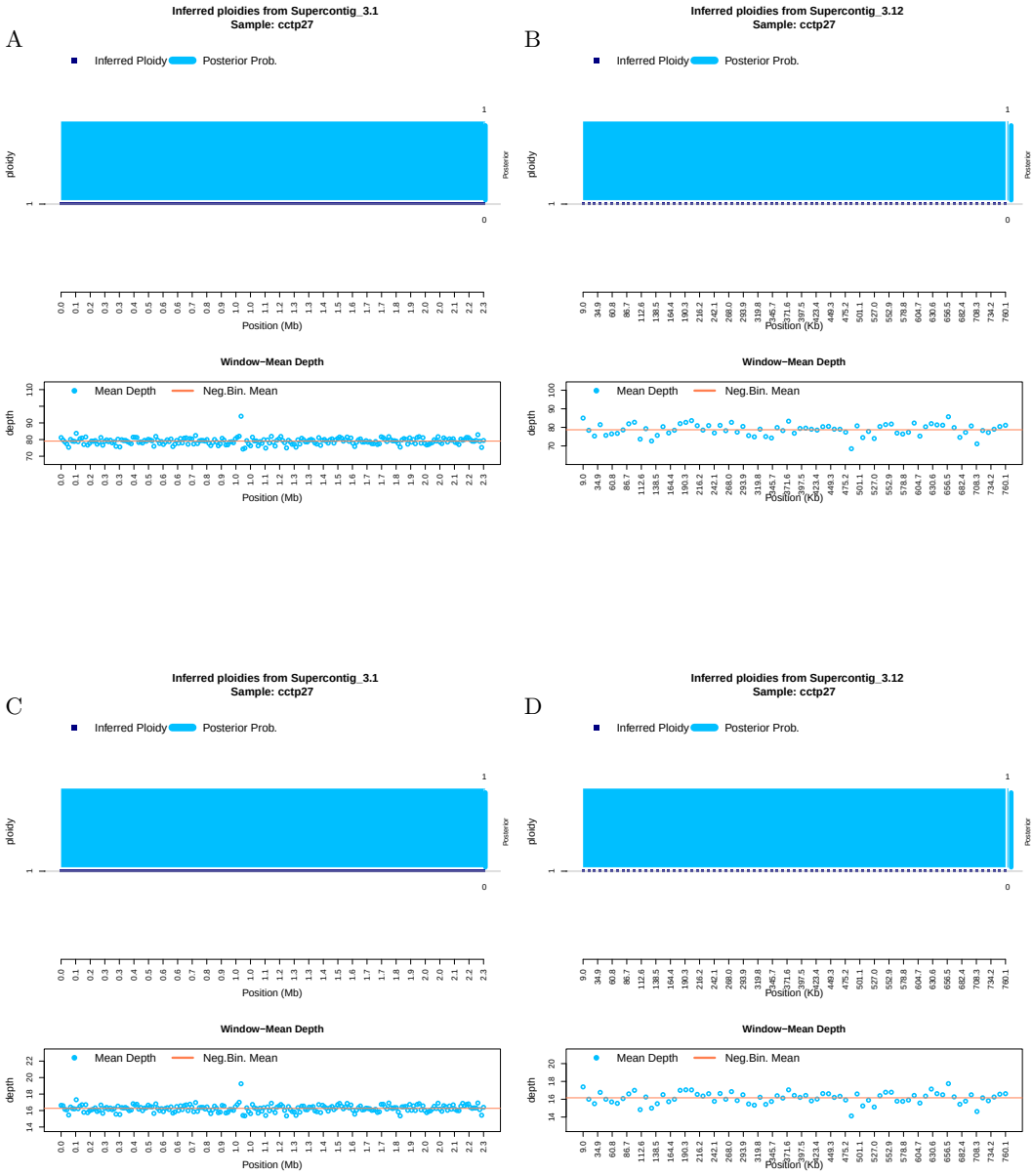

Figure S 9**: Ploidy inference on full and downsampled sequencing data.** Inferred ploidy levels from `HMMploidy` for chromosome 1 and 12 of isolate cctp27. **(A-B)** Results using the whole data on chromosomes 1 (A) and 12 (B). **(C-D)** Results using the data downsampled to 20% of its original depth on chromosomes 1 (C) and 12 (D).

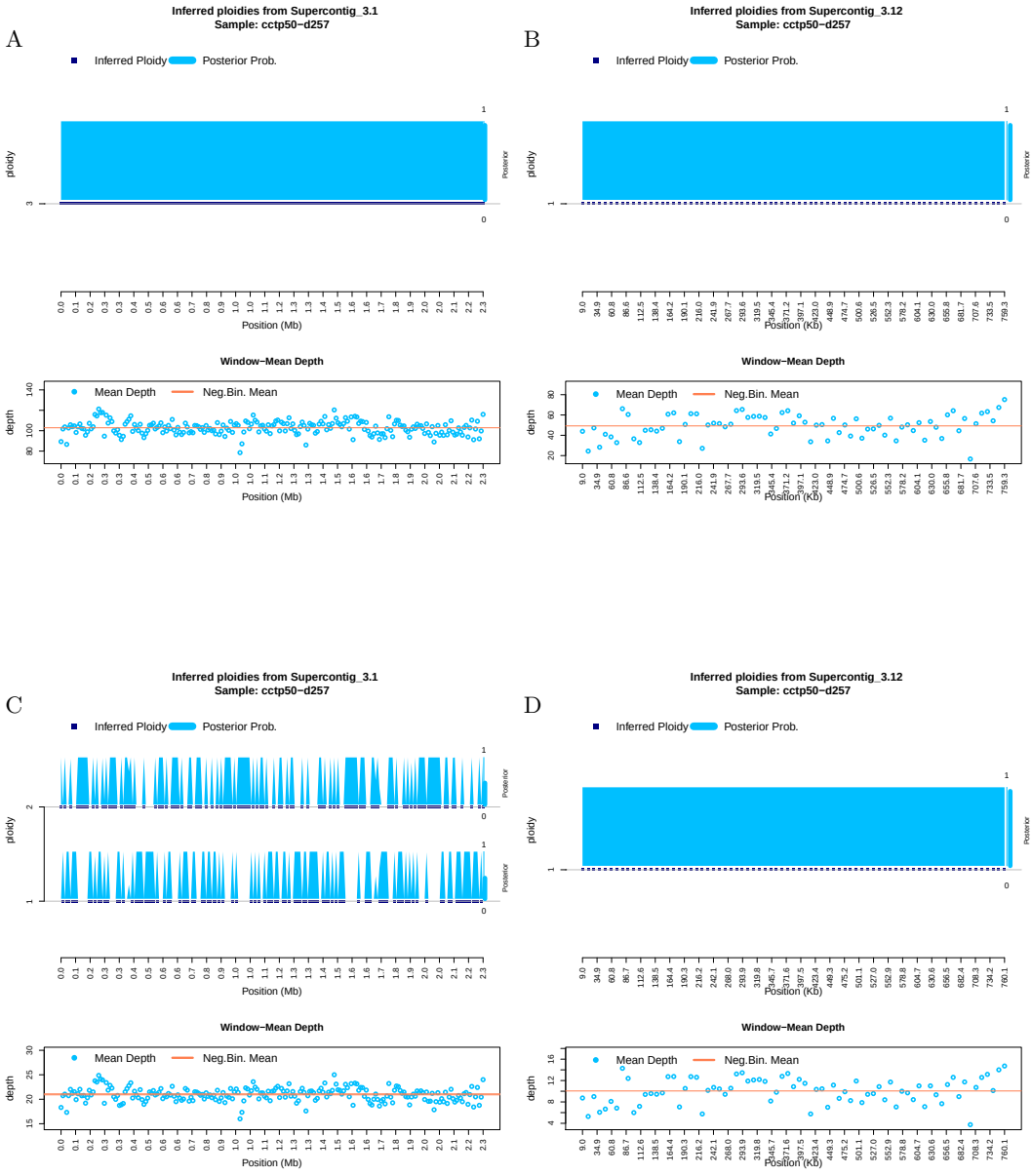

Figure S 10**: Ploidy inference on full and downsampled sequencing data.** Inferred ploidy levels from `HMMploidy` for chromosome 1 and 12 of isolate cctp50-d257. **(A-B)** Results using the whole data on chromosomes 1 (A) and 12 (B). **(C-D)** Results using the data downsampled to 20% of its original depth on chromosomes 1 (C) and 12 (D).

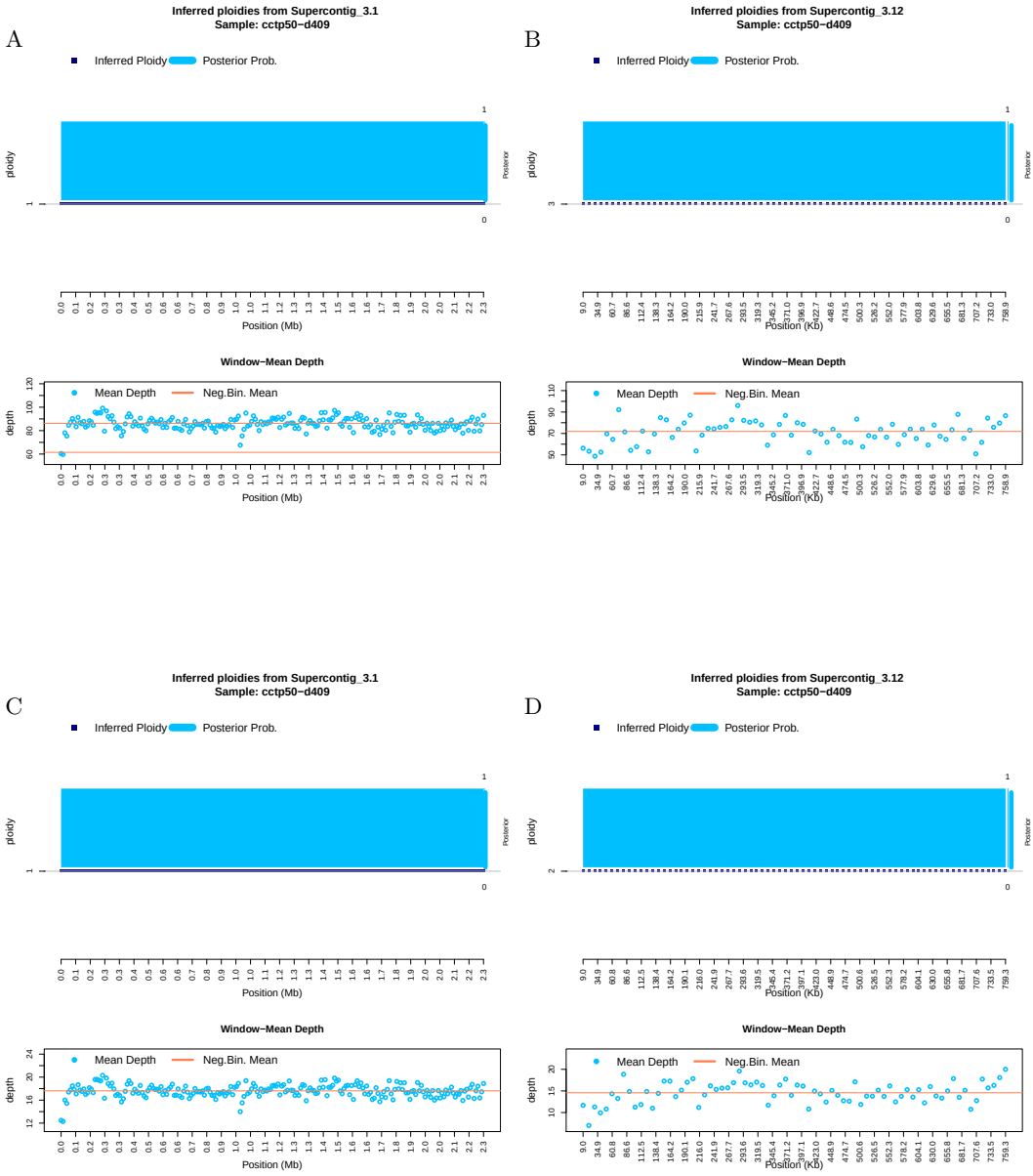

Figure S 11**: Ploidy inference on full and downsampled sequencing data.** Inferred ploidy levels from `HMMploidy` for chromosome 1 and 12 of isolate cctp50-d409. **(A-B)** Results using the whole data on chromosomes 1 (A) and 12 (B). **(C-D)** Results using the data downsampled to 20% of its original depth on chromosomes 1 (C) and 12 (D).

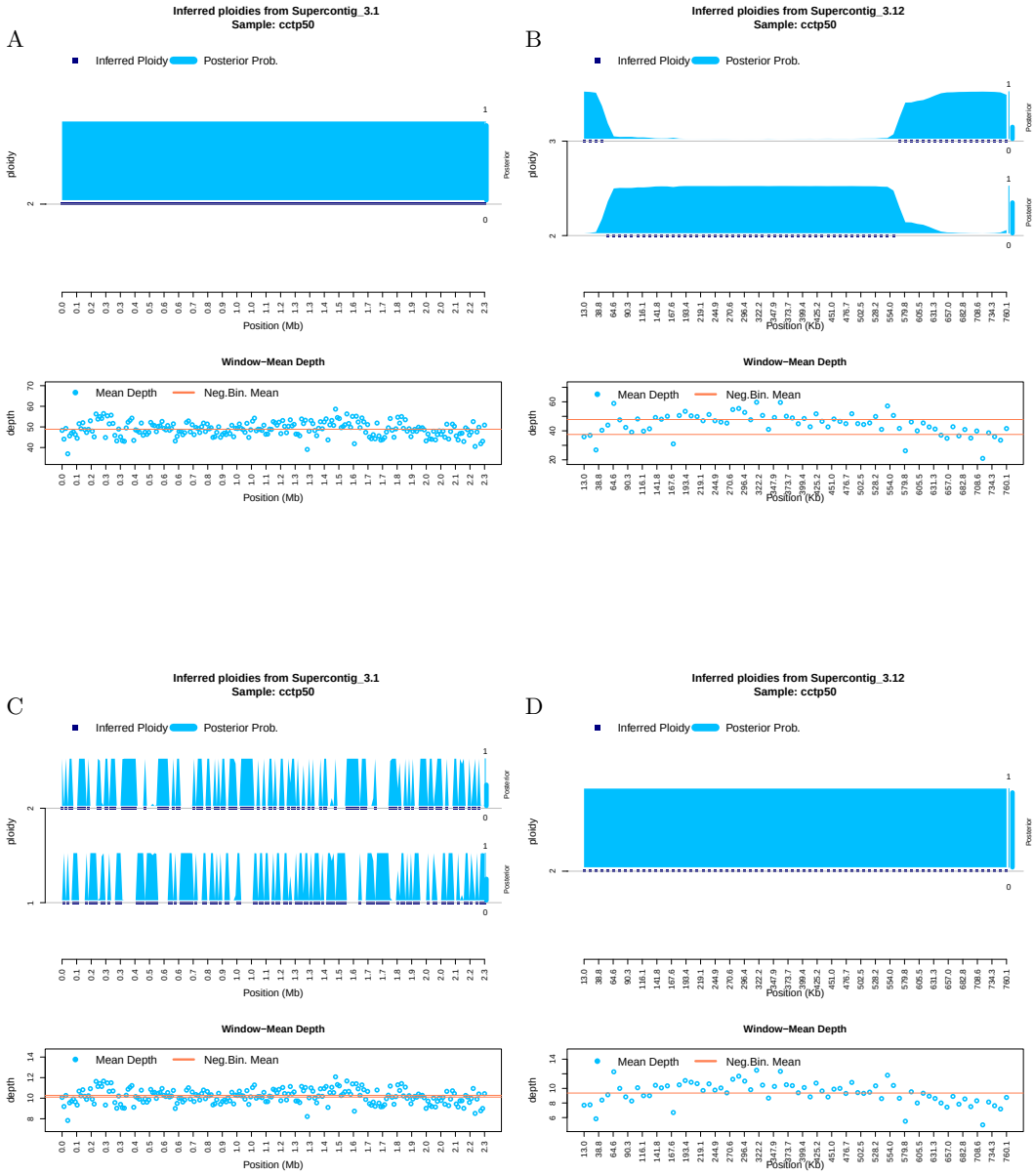

Figure S 12**: Ploidy inference on full and downsampled sequencing data.** Inferred ploidy levels from `HMMploidy` for chromosome 1 and 12 of isolate cctp50. **(A-B)** Results using the whole data on chromosomes 1 (A) and 12 (B). **(C-D)** Results using the data downsampled to 20% of its original depth on chromosomes 1 (C) and 12 (D).

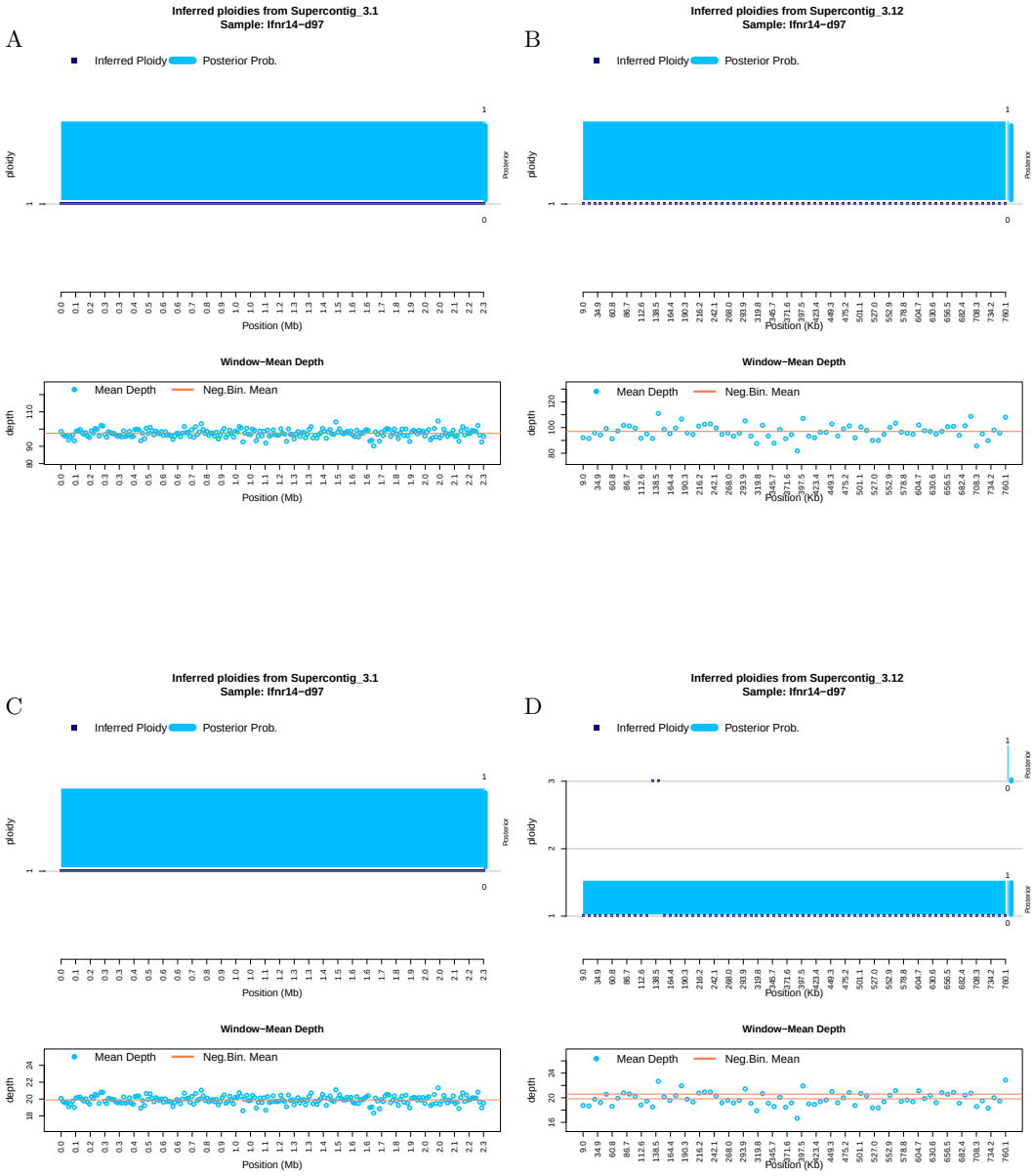

Figure S 13**: Ploidy inference on full and downsampled sequencing data.** Inferred ploidy levels from `HMMploidy` for chromosome 1 and 12 of isolate ifnr14-d97. **(A-B)** Results using the whole data on chromosomes 1 (A) and 12 (B). **(C-D)** Results using the data downsampled to 20% of its original depth on chromosomes 1 (C) and 12 (D).

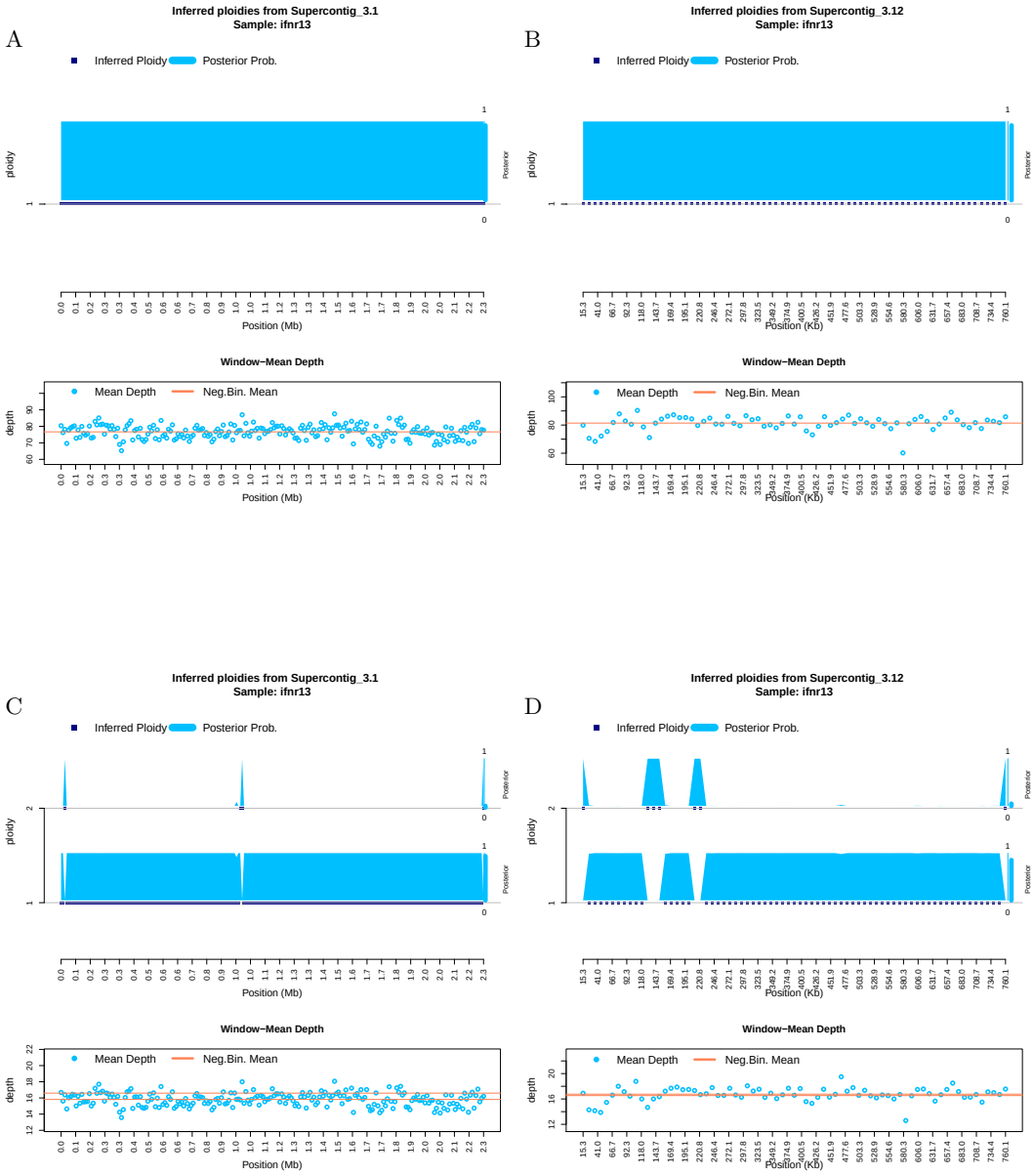

Figure S 14**: Ploidy inference on full and downsampled sequencing data.** Inferred ploidy levels from `HMMploidy` for chromosome 1 and 12 of isolate ifnr13. **(A-B)** Results using the whole data on chromosomes 1 (A) and 12 (B). **(C-D)** Results using the data downsampled to 20% of its original depth on chromosomes 1 (C) and 12 (D).

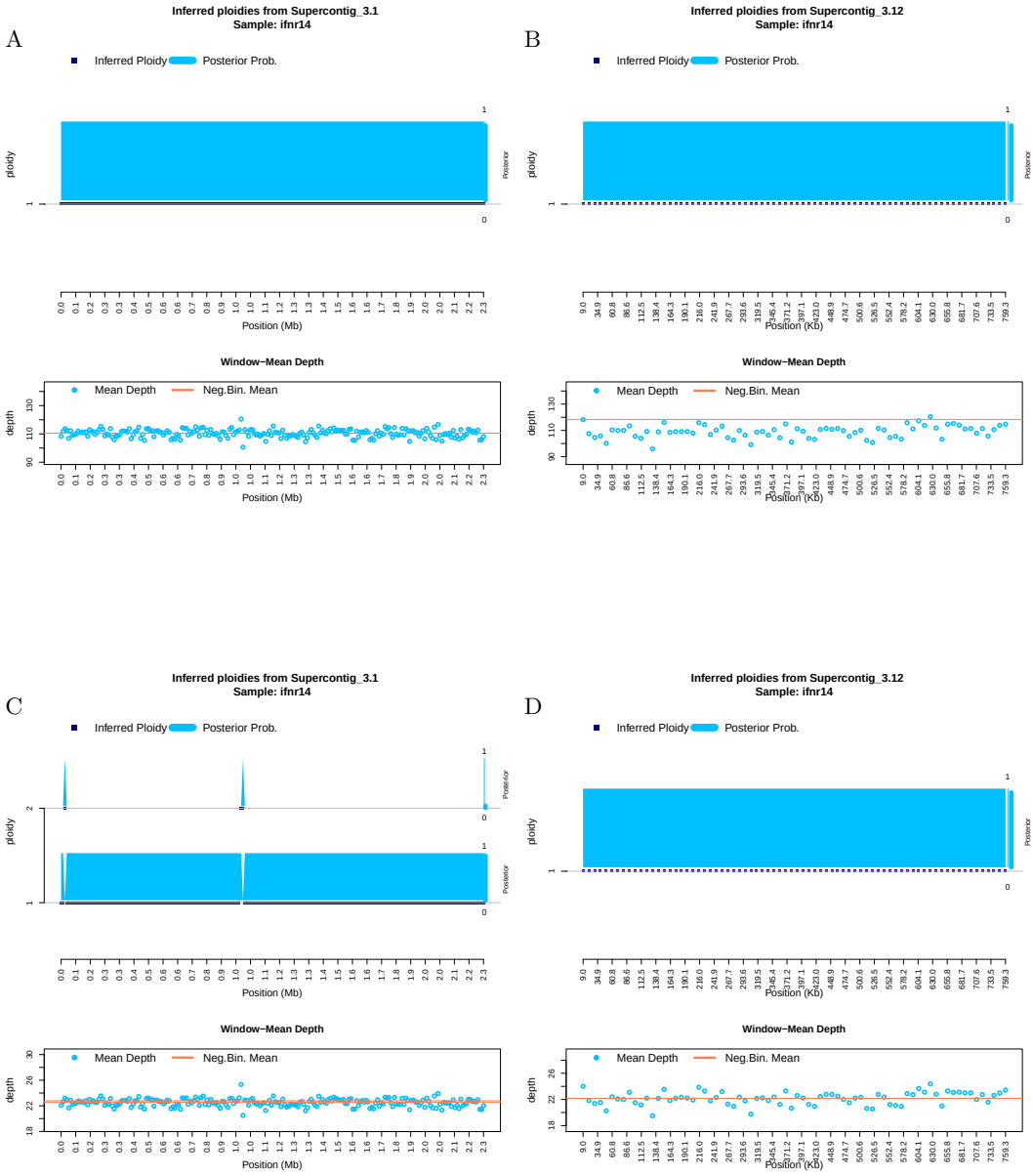

Figure S 15**: Ploidy inference on full and downsampled sequencing data.** Inferred ploidy levels from `HMMploidy` for chromosome 1 and 12 of isolate ifnr14. **(A-B)** Results using the whole data on chromosomes 1 (A) and 12 (B). **(C-D)** Results using the data downsampled to 20% of its original depth on chromosomes 1 (C) and 12 (D).

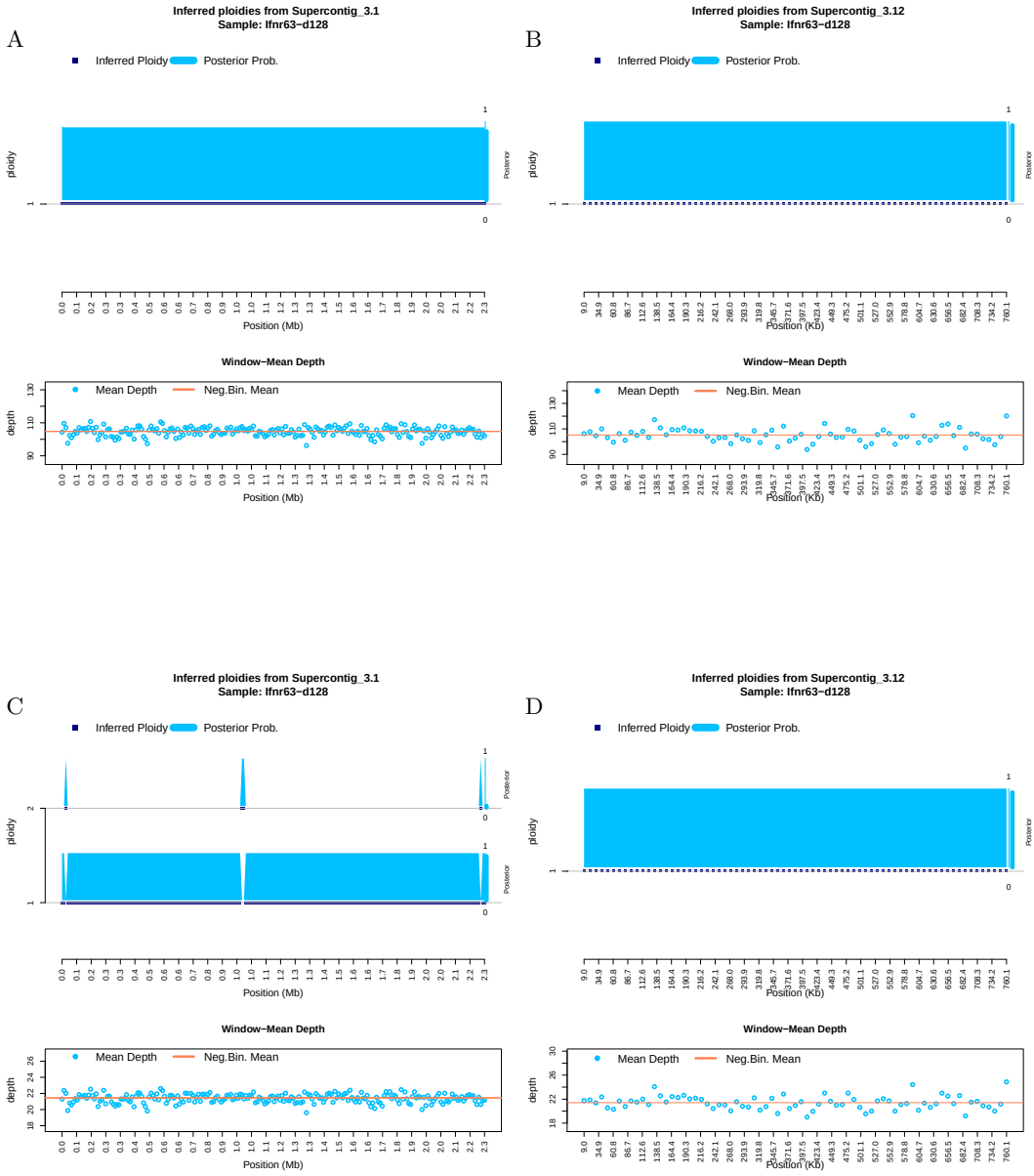

Figure S 16**: Ploidy inference on full and downsampled sequencing data.** Inferred ploidy levels from `HMMploidy` for chromosome 1 and 12 of isolate ifnr63-d128. **(A-B)** Results using the whole data on chromosomes 1 (A) and 12 (B). **(C-D)** Results using the data downsampled to 20% of its original depth on chromosomes 1 (C) and 12 (D).

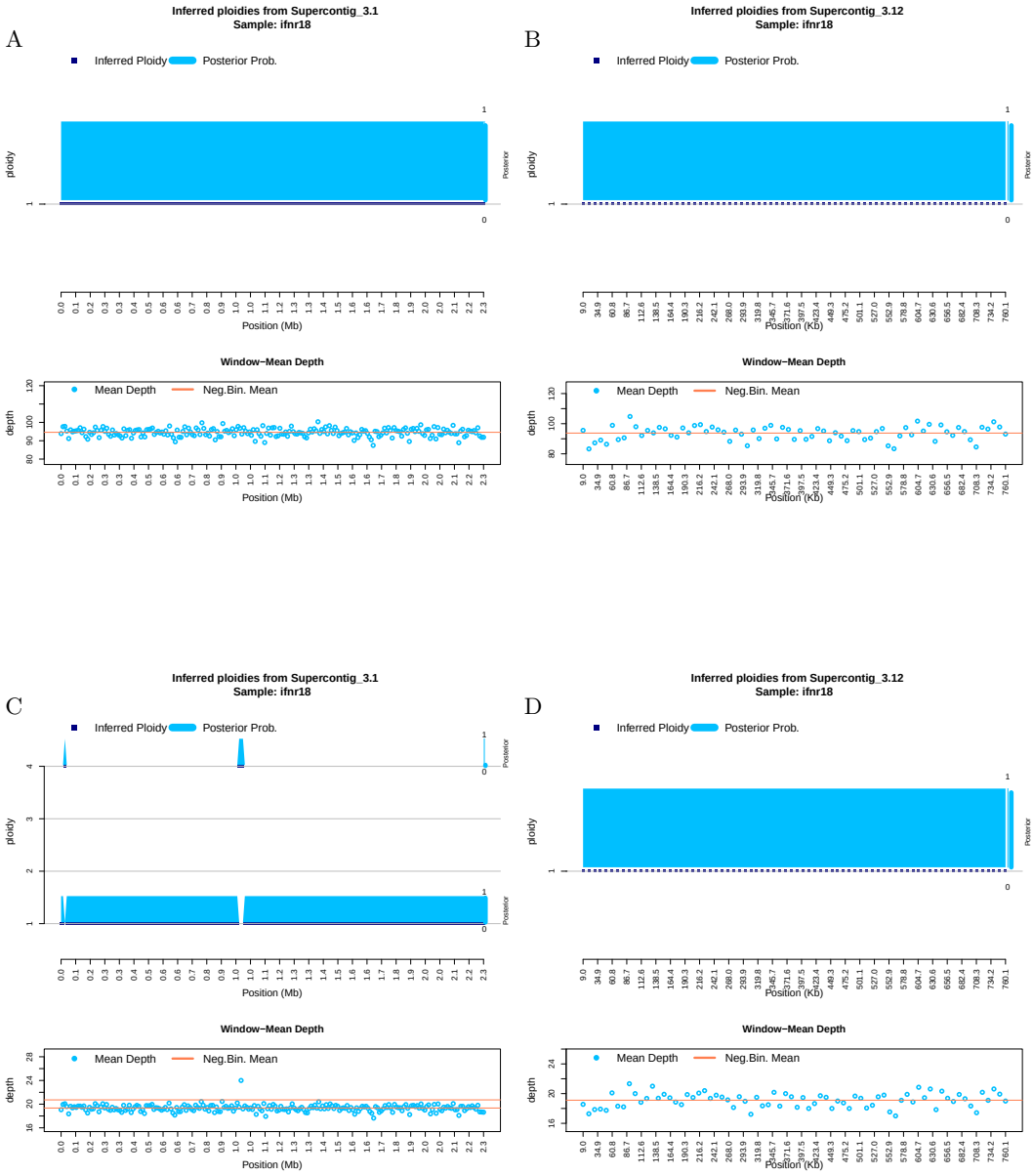

Figure S 17**: Ploidy inference on full and downsampled sequencing data.** Inferred ploidy levels from `HMMploidy` for chromosome 1 and 12 of isolate ifnr18. **(A-B)** Results using the whole data on chromosomes 1 (A) and 12 (B). **(C-D)** Results using the data downsampled to 20% of its original depth on chromosomes 1 (C) and 12 (D).

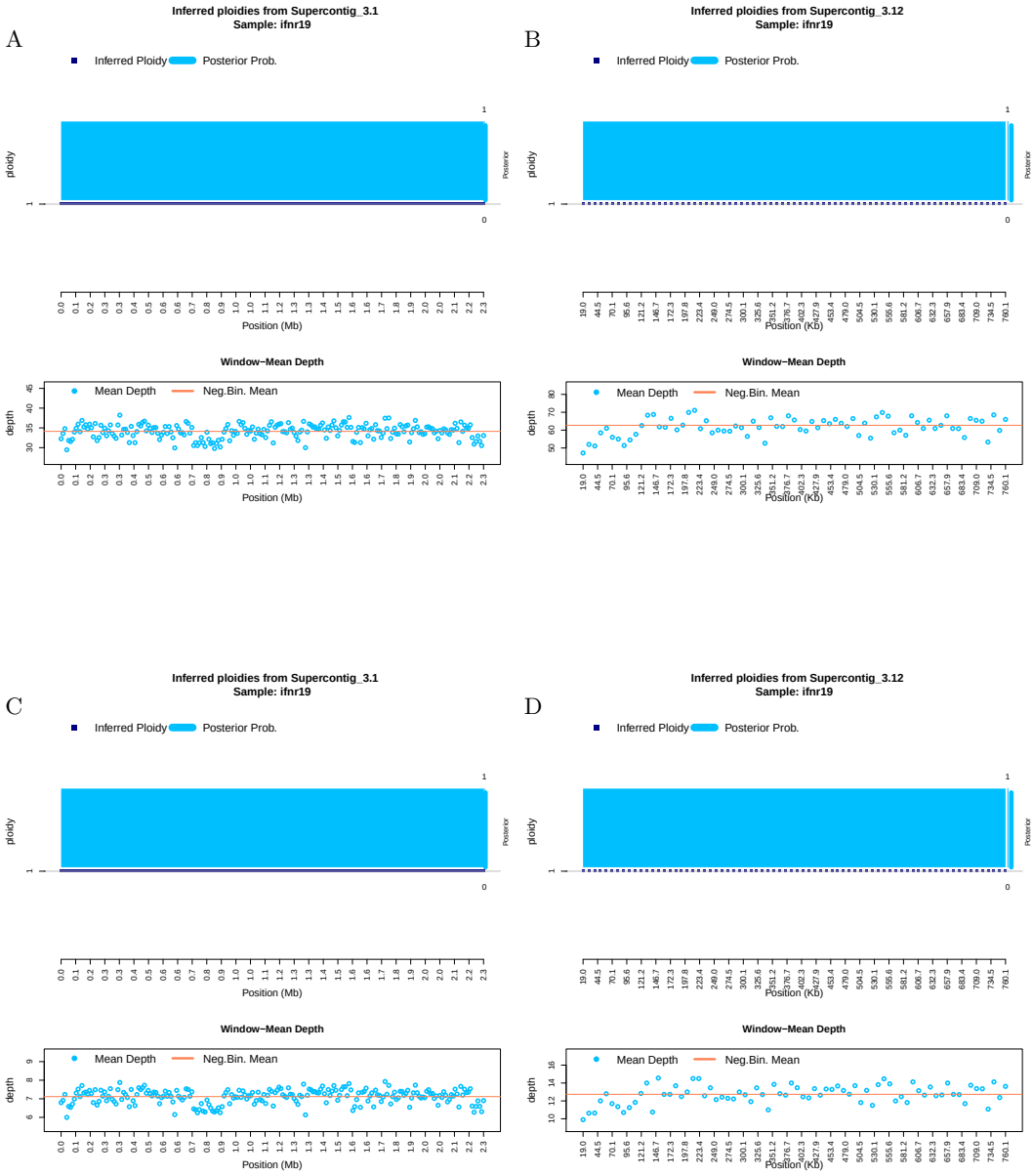

Figure S 18**: Ploidy inference on full and downsampled sequencing data.** Inferred ploidy levels from `HMMploidy` for chromosome 1 and 12 of isolate ifnr19. **(A-B)** Results using the whole data on chromosomes 1 (A) and 12 (B). **(C-D)** Results using the data downsampled to 20% of its original depth on chromosomes 1 (C) and 12 (D).

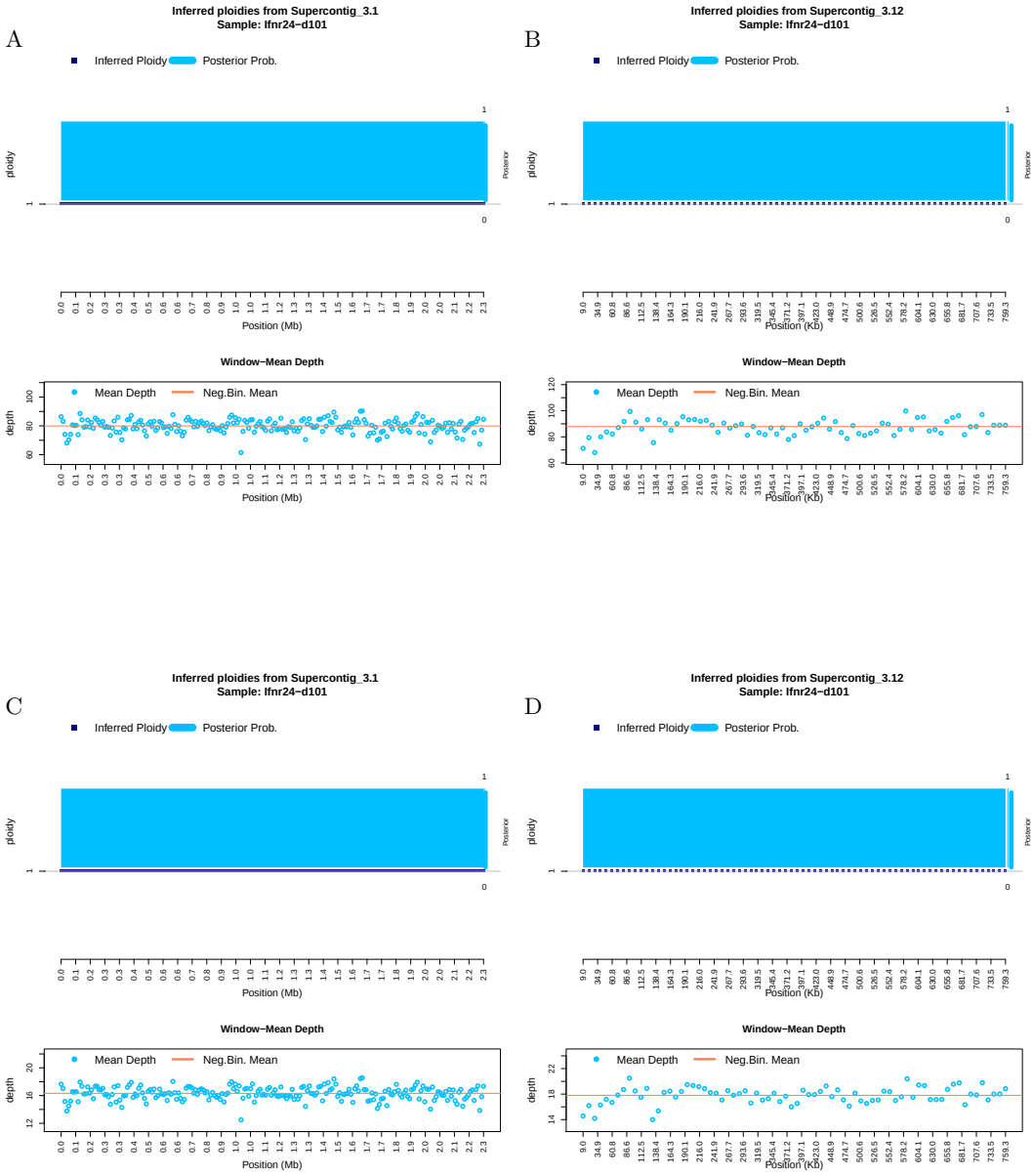

Figure S 19**: Ploidy inference on full and downsampled sequencing data.** Inferred ploidy levels from `HMMploidy` for chromosome 1 and 12 of isolate ifnr24-d101. **(A-B)** Results using the whole data on chromosomes 1 (A) and 12 (B). **(C-D)** Results using the data downsampled to 20% of its original depth on chromosomes 1 (C) and 12 (D).

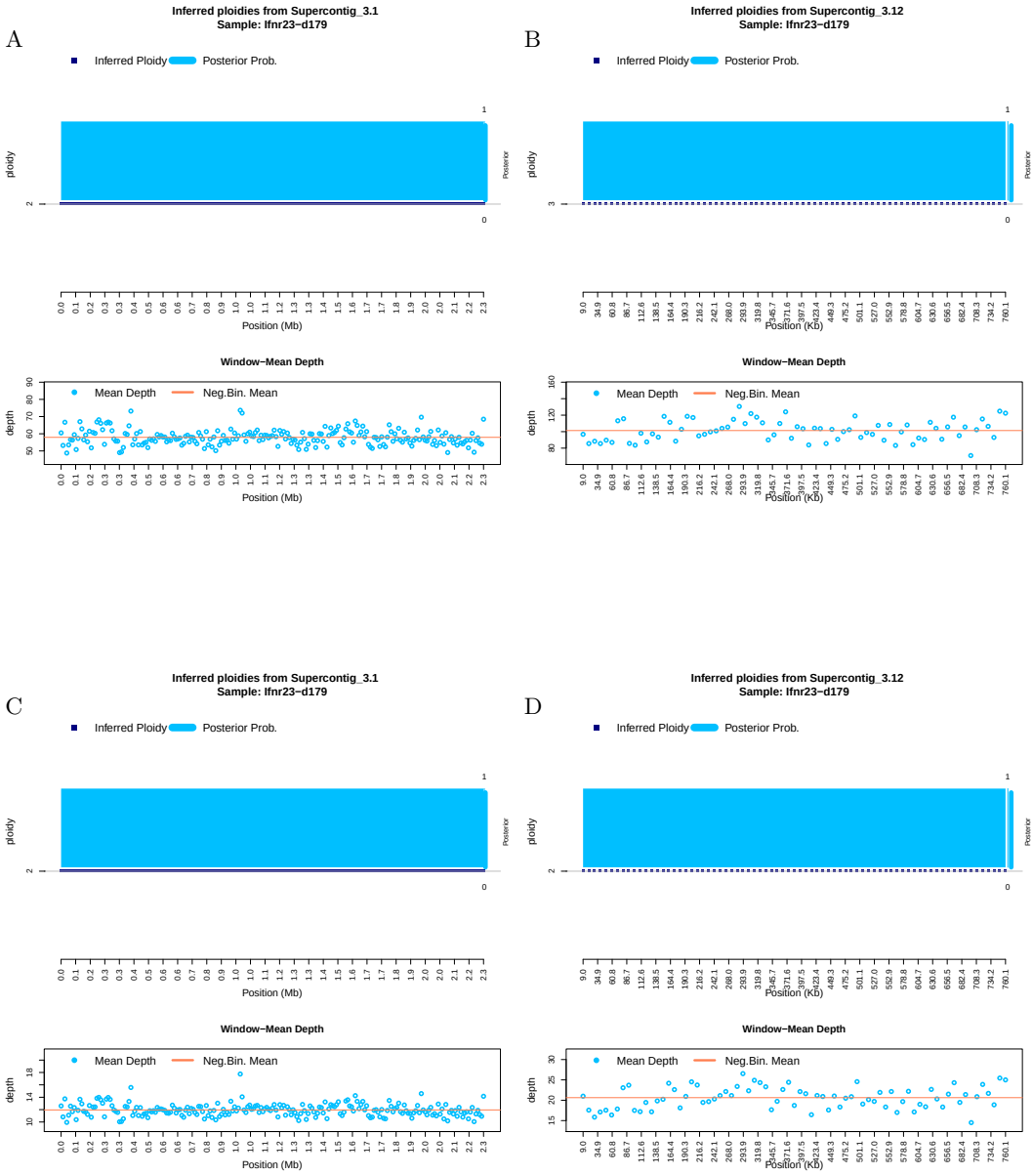

Figure S 20**: Ploidy inference on full and downsampled sequencing data.** Inferred ploidy levels from `HMMploidy` for chromosome 1 and 12 of isolate ifnr23-d179. **(A-B)** Results using the whole data on chromosomes 1 (A) and 12 (B). **(C-D)** Results using the data downsampled to 20% of its original depth on chromosomes 1 (C) and 12 (D).

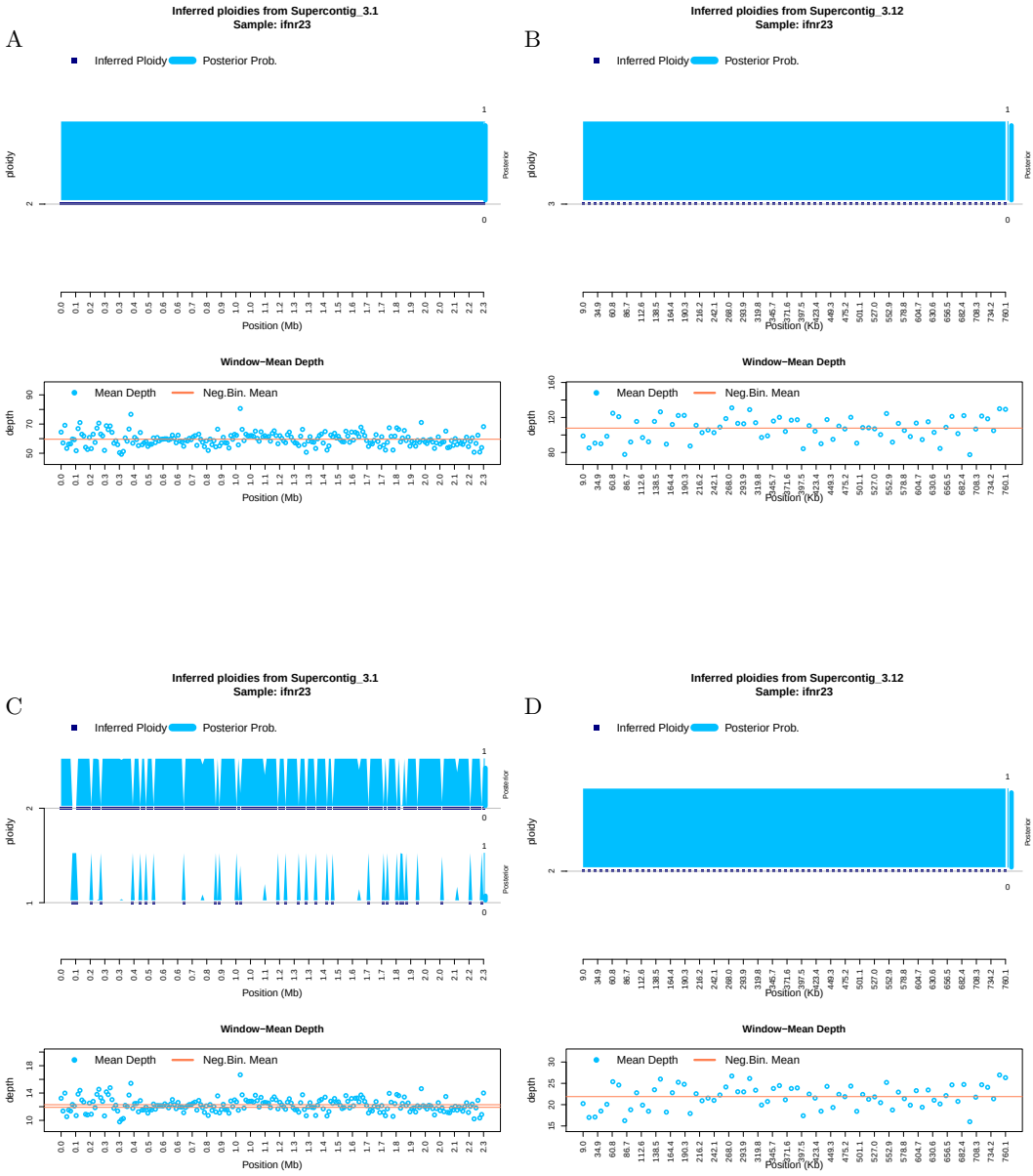

Figure S 21**: Ploidy inference on full and downsampled sequencing data.** Inferred ploidy levels from `HMMploidy` for chromosome 1 and 12 of isolate ifnr23. **(A-B)** Results using the whole data on chromosomes 1 (A) and 12 (B). **(C-D)** Results using the data downsampled to 20% of its original depth on chromosomes 1 (C) and 12 (D).

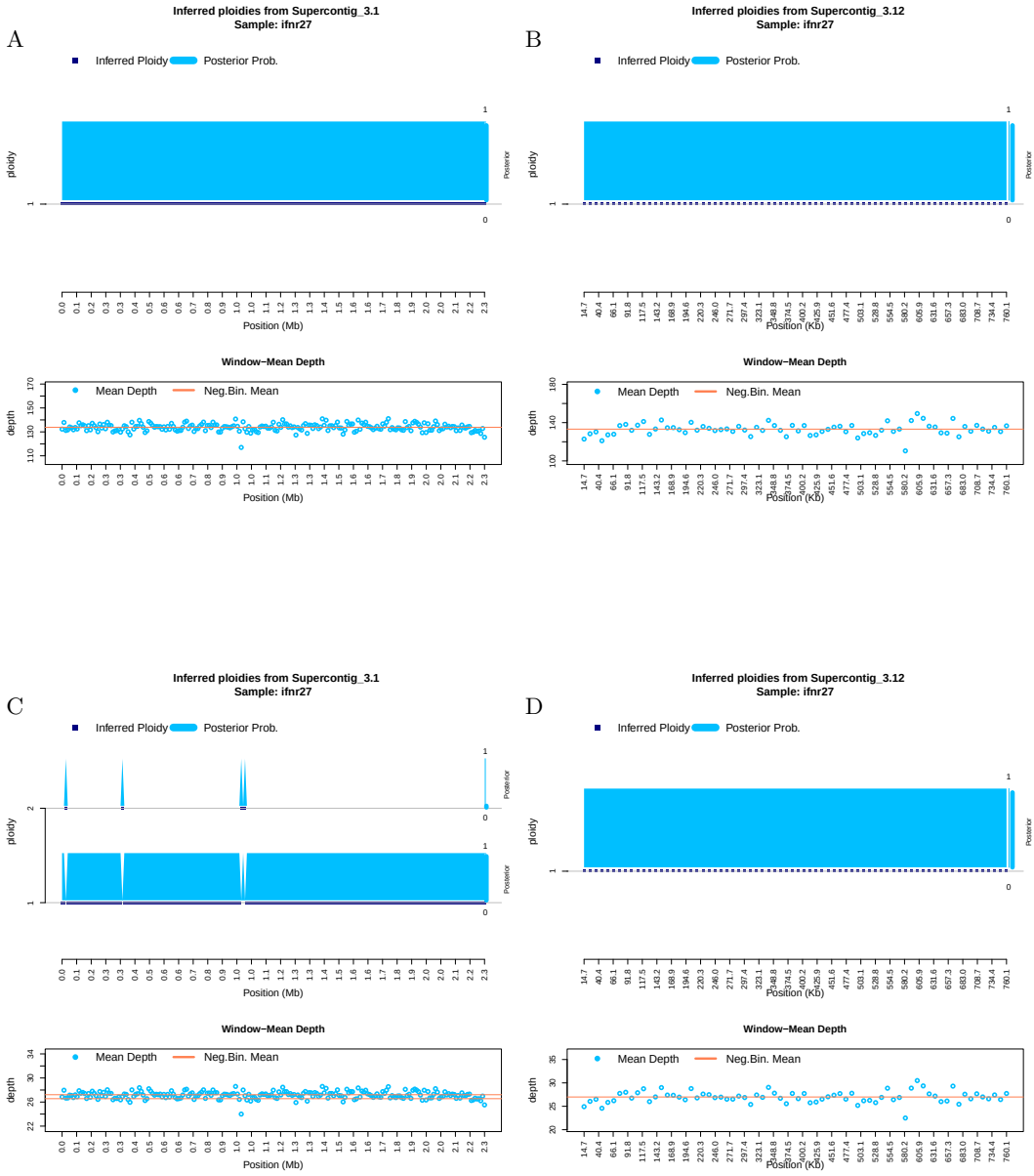

Figure S 22**: Ploidy inference on full and downsampled sequencing data.** Inferred ploidy levels from `HMMploidy` for chromosome 1 and 12 of isolate ifnr27. **(A-B)** Results using the whole data on chromosomes 1 (A) and 12 (B). **(C-D)** Results using the data downsampled to 20% of its original depth on chromosomes 1 (C) and 12 (D).

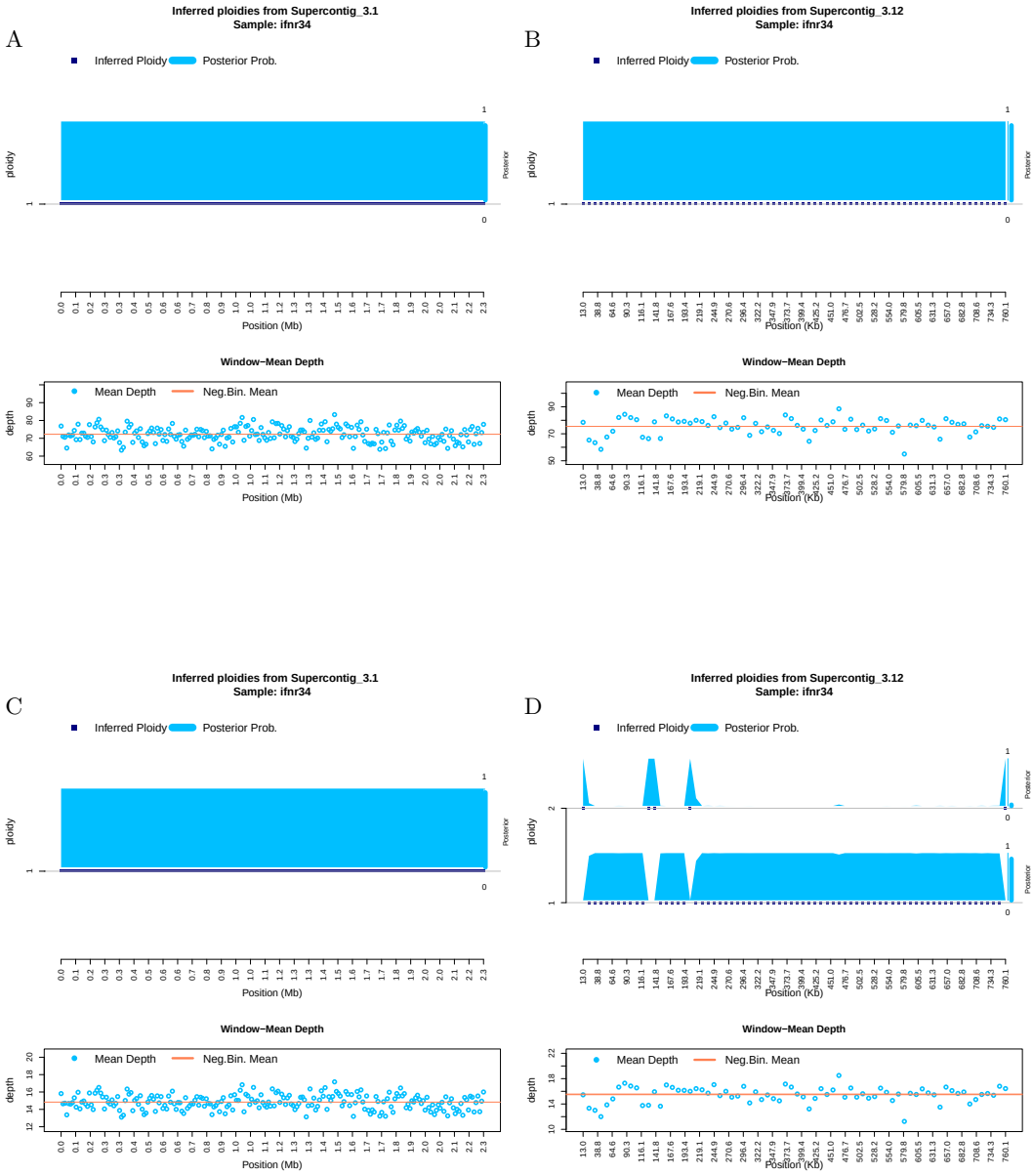

Figure S 23**: Ploidy inference on full and downsampled sequencing data.** Inferred ploidy levels from `HMMploidy` for chromosome 1 and 12 of isolate ifnr34. **(A-B)** Results using the whole data on chromosomes 1 (A) and 12 (B). **(C-D)** Results using the data downsampled to 20% of its original depth on chromosomes 1 (C) and 12 (D).

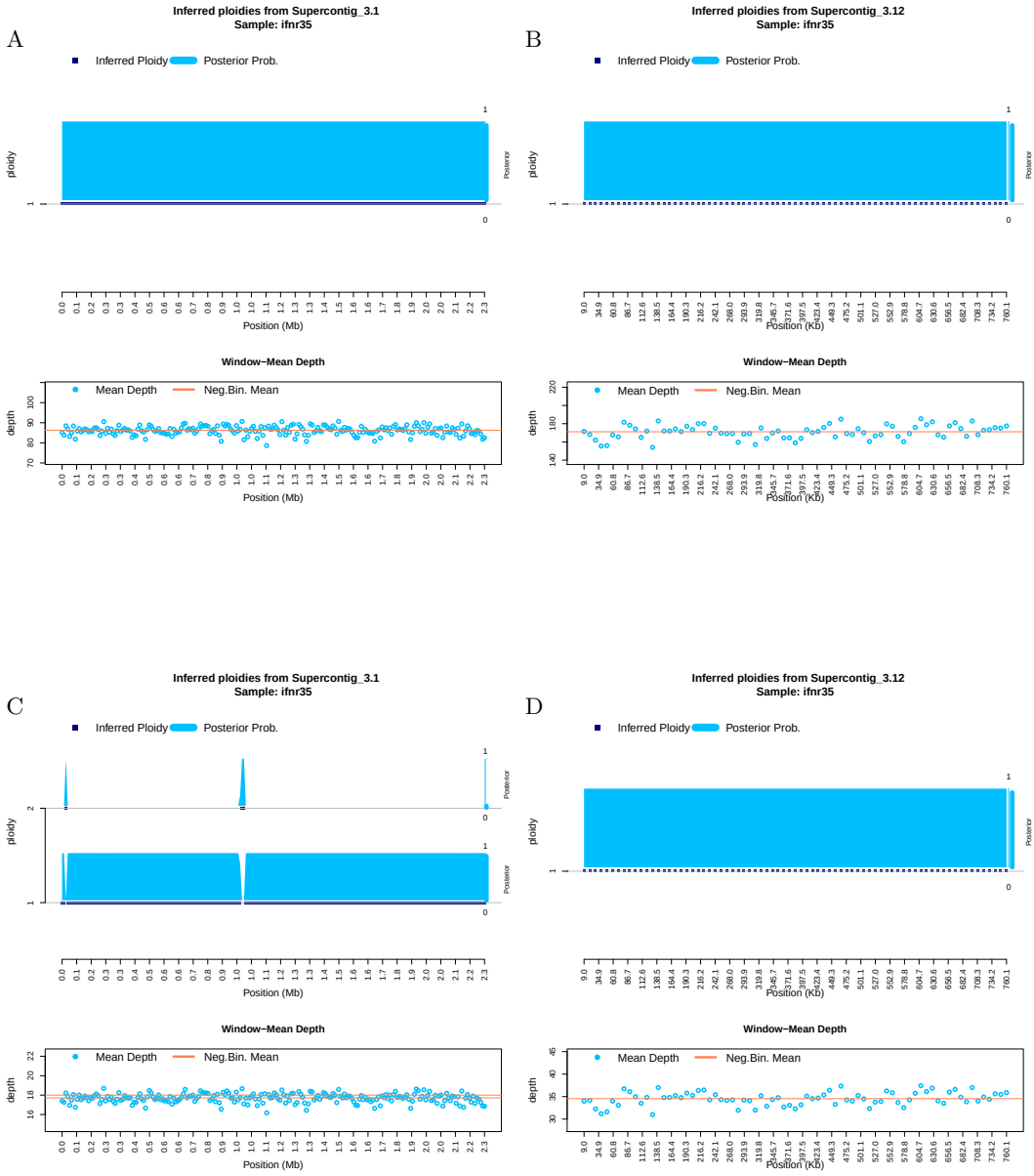

Figure S 24**: Ploidy inference on full and downsampled sequencing data.** Inferred ploidy levels from `HMMploidy` for chromosome 1 and 12 of isolate ifnr35. **(A-B)** Results using the whole data on chromosomes 1 (A) and 12 (B). **(C-D)** Results using the data downsampled to 20% of its original depth on chromosomes 1 (C) and 12 (D).

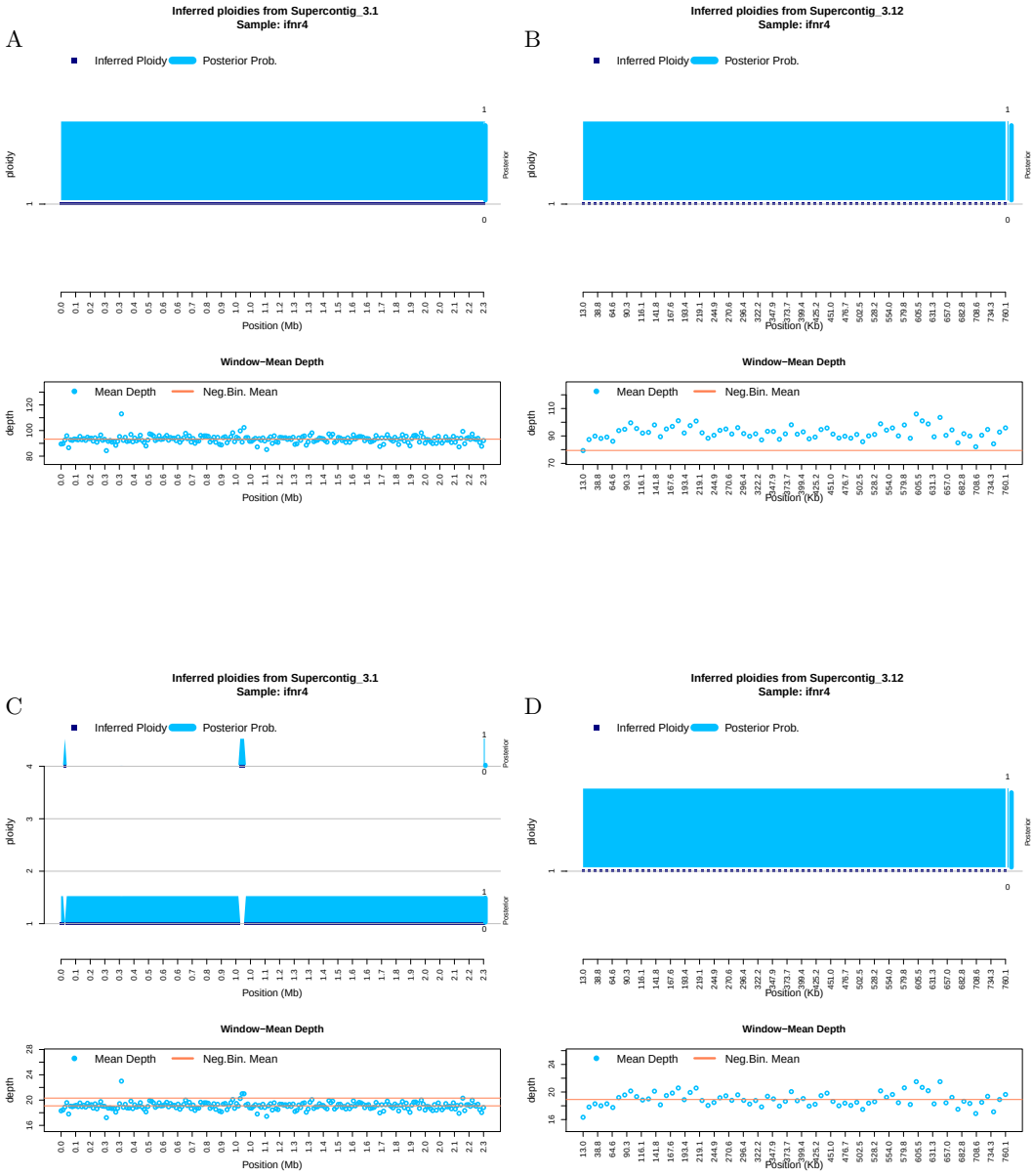

Figure S 25**: Ploidy inference on full and downsampled sequencing data.** Inferred ploidy levels from `HMMploidy` for chromosome 1 and 12 of isolate ifnr4. **(A-B)** Results using the whole data on chromosomes 1 (A) and 12 (B). **(C-D)** Results using the data downsampled to 20% of its original depth on chromosomes 1 (C) and 12 (D).

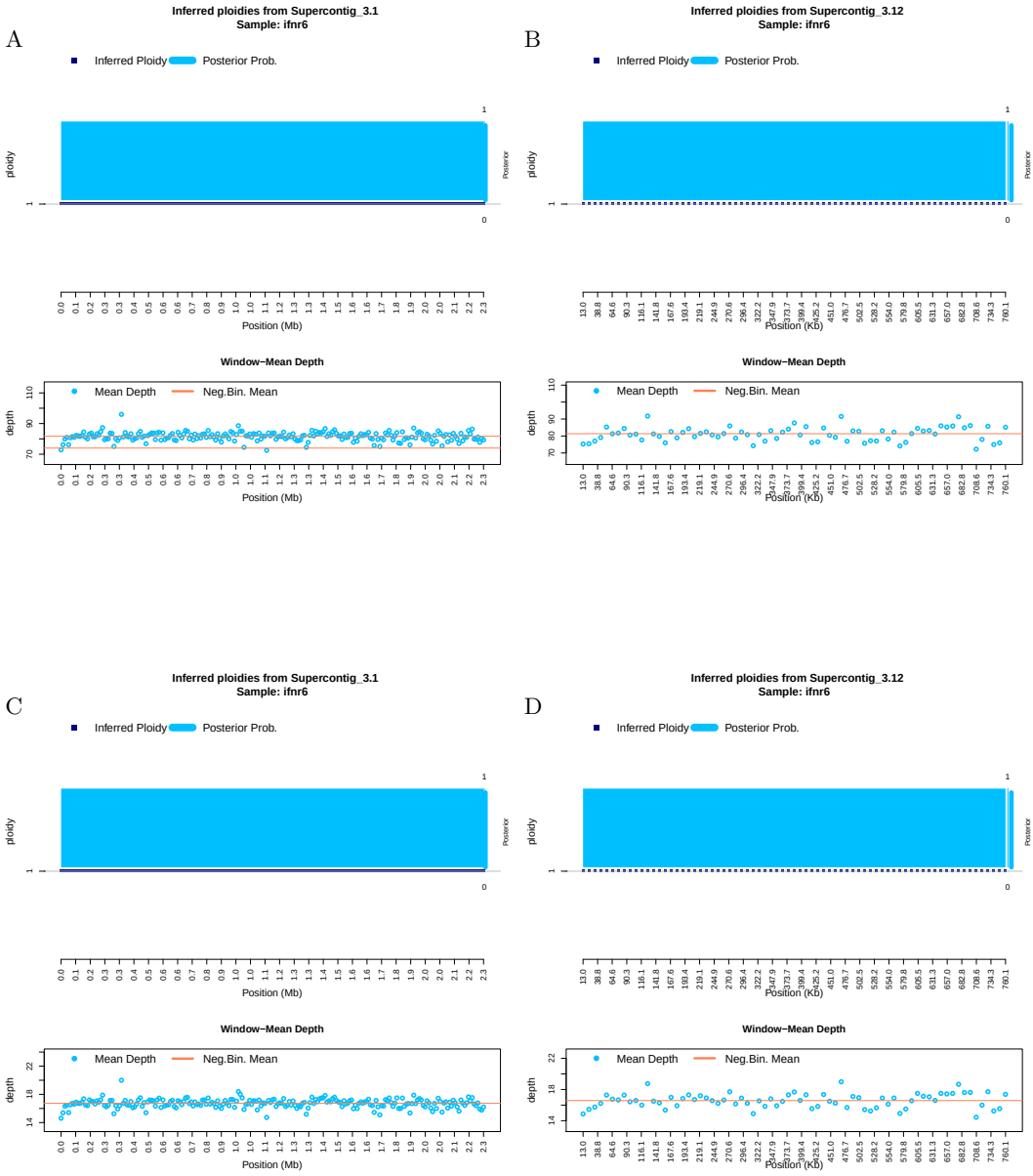

Figure S 26**: Ploidy inference on full and downsampled sequencing data.** Inferred ploidy levels from `HMMploidy` for chromosome 1 and 12 of isolate ifnr6. **(A-B)** Results using the whole data on chromosomes 1 (A) and 12 (B). **(C-D)** Results using the data downsampled to 20% of its original depth on chromosomes 1 (C) and 12 (D).

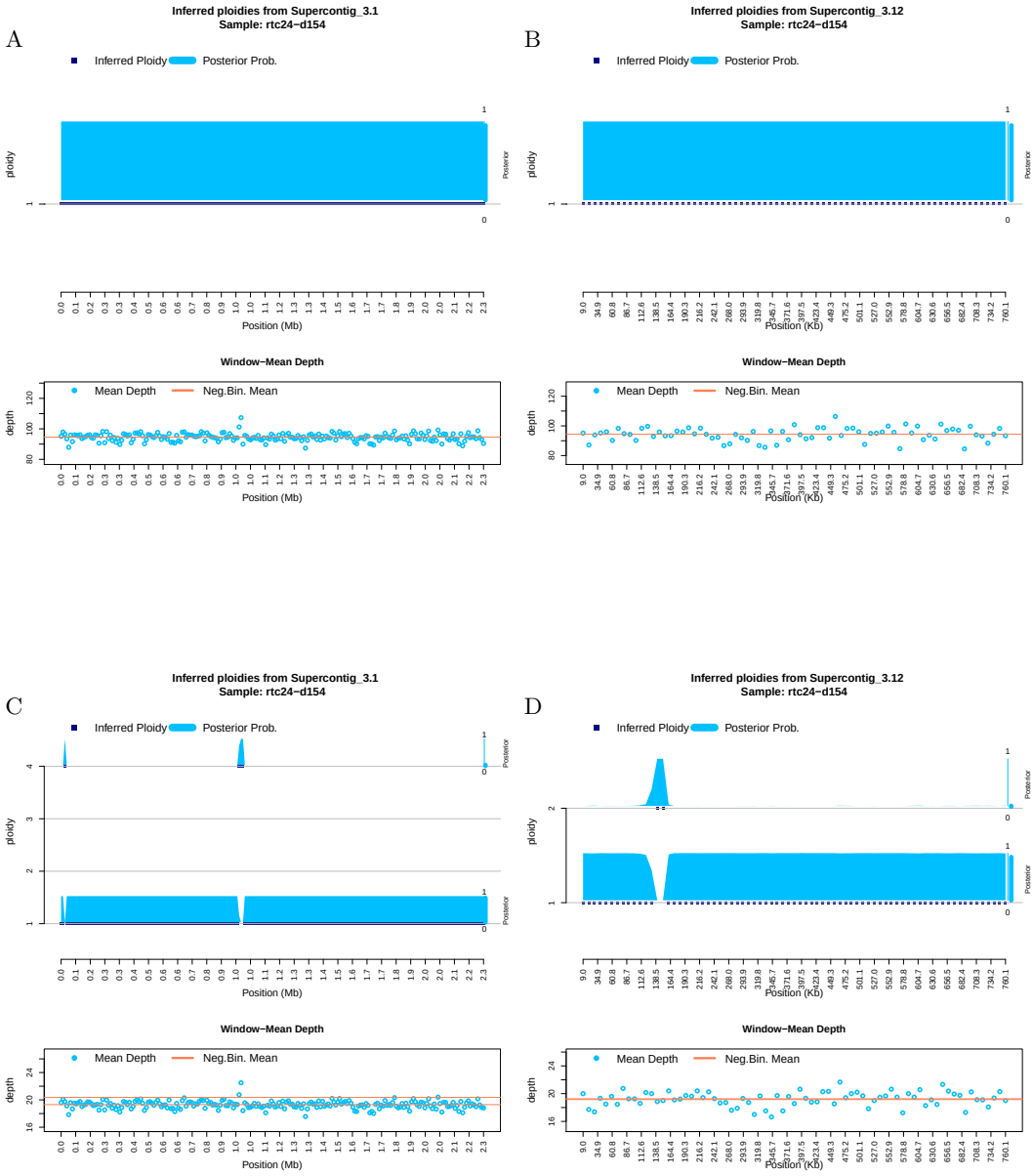

Figure S 27**: Ploidy inference on full and downsampled sequencing data.** Inferred ploidy levels from `HMMploidy` for chromosome 1 and 12 of isolate rtc24-d154. **(A-B)** Results using the whole data on chromosomes 1 (A) and 12 (B). **(C-D)** Results using the data downsampled to 20% of its original depth on chromosomes 1 (C) and 12 (D).

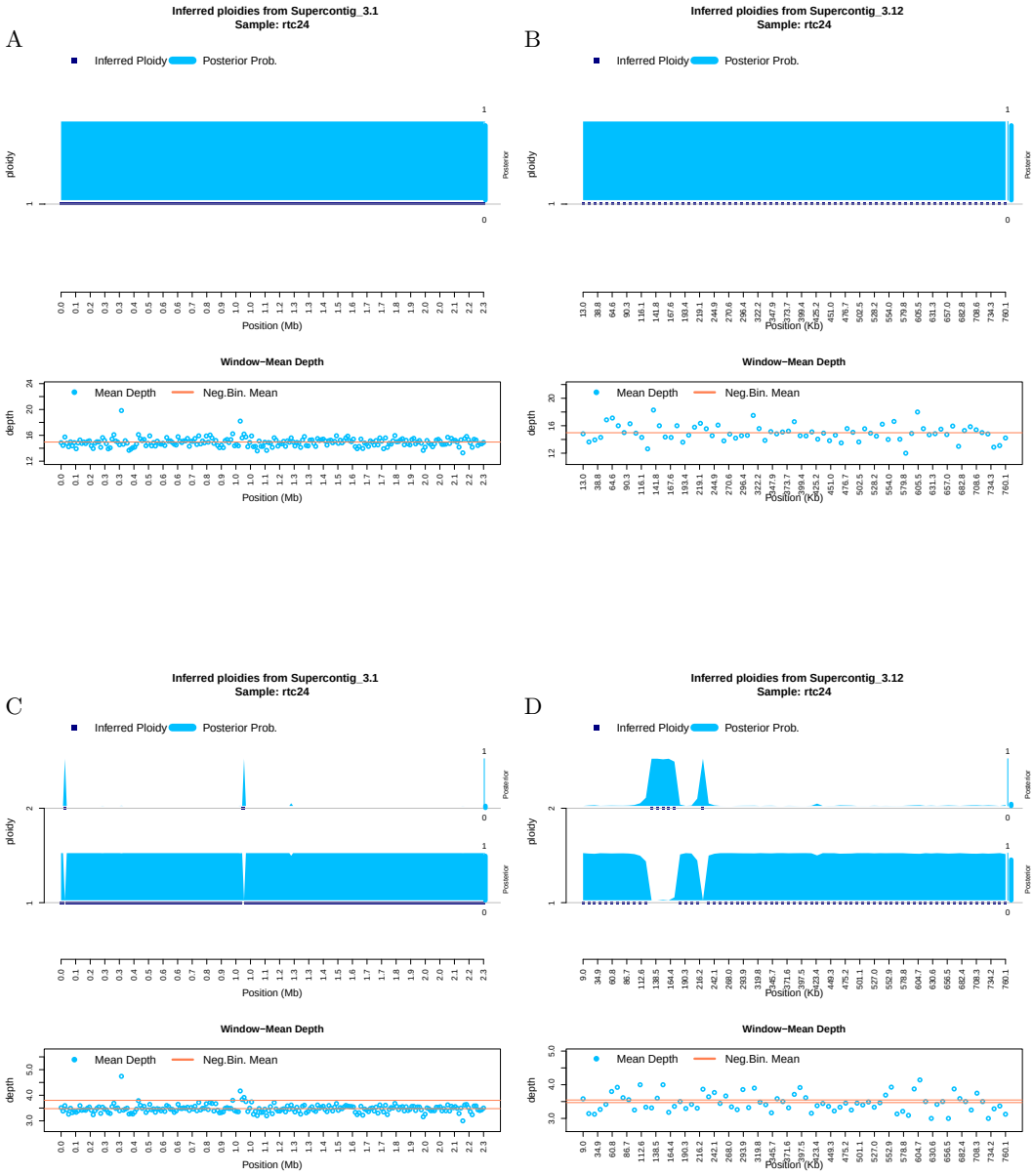

Figure S 28**: Ploidy inference on full and downsampled sequencing data.** Inferred ploidy levels from `HMMploidy` for chromosome 1 and 12 of isolate rtc24. **(A-B)** Results using the whole data on chromosomes 1 (A) and 12 (B). **(C-D)** Results using the data downsampled to 20% of its original depth on chromosomes 1 (C) and 12 (D).

Bibliography

[1] Anders, S., Huber, W.: Differential expression analysis for sequence count data. Genome biology **11**(10), R106–R106 (2010). https://doi.org/10.1186/gb-2010-11-10-r106, https://pubmed.ncbi.nlm.nih.gov/20979621, 20979621[pmid]

[2] Augusto Corrêa dos Santos, R., Goldman, G.H., Riaño-Pachón, D.M.: ploidyNGS: visually exploring ploidy with Next Generation Sequencing data. Bioinformatics **33**(16), 2575–2576 (aug 2017). https://doi.org/10.1093/bioinformatics/btx204, http://www.ncbi.nlm.nih.gov/pubmed/28383704

[3] Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K.V., Altshuler, D., Gabriel, S., DePristo, M.A.: From fastq data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. Current Protocols in Bioinformatics **43**(1), 11.10.1–11.10.33 (2013). https://doi.org/https://doi.org/10.1002/0471250953.bi1110s43, https://currentprotocols.onlinelibrary.wiley.com/doi/abs/10.1002/0471250953.bi1110s43

[4] Avramovska, O., Rego, E., Hickman, M.A.: Tetraploidy accelerates adaption under drug-selection in a fungal pathogen. bioRxiv (2021). https://doi.org/10.1101/2021.02.28.433243, https://www.biorxiv.org/content/early/2021/02/28/2021.02.28.433243

[5] Bao, L., Pu, M., Messer, K.: AbsCN-seq: a statistical method to estimate tumor purity, ploidy and absolute copy numbers from next-generation sequencing data. Bioinformatics **30**(8), 1056–1063 (apr 2014). https://doi.org/10.1093/bioinformatics/btt759, https://academic.oup.com/bioinformatics/ARTICLE-lookup/doi/10.1093/bioinformatics/btt759

[6] Ben-David, U., Amon, A.: Context is everything: aneuploidy in cancer. Nature Reviews Genetics **21**(1), 44–62 (Jan 2020). https://doi.org/10.1038/s41576-019-0171-x, https://doi.org/10.1038/s41576-019-0171-x

[7] Bishop, C.M.: Pattern recognition and machine learning. Springer (2006)

[8] Cappe, O., Moulines, E., Ryden, T.: Inference in Hidden Markov Models. Springer Science+Business Media, Inc (2005)

[9] Casella, G., Berger, R.L.: Statistical inference. Thomson Learning (2002)

[10] Chen, B., Cole, J.W., Grond-Ginsbach, C.: Departure from Hardy Weinberg Equilibrium and Genotyping Error. Front Genet. (8) (2017)

[11] Choudhary, S., Satija, R.: Comparison and evaluation of statistical error models for scrna-seq. Genome Biology **23**(1), 27 (Jan 2022). https://doi.org/10.1186/s13059-021-02584-9, https://doi.org/10.1186/s13059-021-02584-9

[12] Coward, J., Harding, A.: Size does matter: Why polyploid tumor cells are critical drug targets in the war on cancer. Frontiers in Oncology **4**, 123 (2014). https://doi.org/10.3389/fonc.2014.00123, https://www.frontiersin.org/article/10.3389/fonc.2014.00123

[13] Davoli, T., de Lange, T.: The causes and consequences of polyploidy in normal development and cancer. Annual Review of Cell and Developmental Biology **27**(1), 585–610 (2011). https://doi.org/10.1146/annurev-cellbio-092910-154234, https://doi.org/10.1146/annurev-cellbio-092910-154234, pMID: 21801013

[14] Ewing, B., Hillier, L., Wendl, M.C., Green, P.: Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome research **8**(3), 175–85 (mar 1998), http://www.ncbi.nlm.nih.gov/pubmed/9521921

[15] Farrer, R.A., Henk, D.A., Garner, T.W.J., Balloux, F., Woodhams, D.C., Fisher, M.C.: Chromosomal Copy Number Variation, Selection and Uneven Rates of Recombination Reveal Cryptic Genome Diversity Linked to Pathogenicity. PLoS Genetics **9**(8), e1003703 (aug 2013). https://doi.org/10.1371/journal.pgen.1003703, http://dx.plos.org/10.1371/journal.pgen.1003703

[16] Favero, F., Joshi, T., Marquard, A.M., Birkbak, N.J., Krzystanek, M., Li, Q., Szallasi, Z., Eklund, A.C.: Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. Annals of Oncology **26**(1), 64–70 (jan 2015). https://doi.org/10.1093/annonc/mdu479, http://www.ncbi.nlm.nih.gov/pubmed/25319062

[17] Fox, D.T., Soltis, D.E., Soltis, P.S., Ashman, T.L., Van de Peer, Y.: Polyploidy: A biological force from cells to ecosystems. Trends in Cell Biology **30**(9), 688–694 (Sep 2020). https://doi.org/10.1016/j.tcb.2020.06.006, https://doi.org/10.1016/j.tcb.2020.06.006

[18] Fu, C., Davy, A., Holmes, S., Sun, S., Yadav, V., Gusa, A., Coelho, M.A., Heitman, J.: Dynamic genome plasticity during unisexual reproduction in the human fungal pathogen cryptococcus deneoformans. PLOS Genetics **17**(11), 1–31 (11 2021). https://doi.org/10.1371/journal.pgen.1009935, https://doi.org/10.1371/journal.pgen.1009935

[19] Fumagalli, M., Vieira, F.G., Linderoth, T., Nielsen, R.: ngsTools: methods for population genetics analyses from next-generation sequencing data. Bioinformatics **30**(10), 1486–1487 (May 2014)

[20] Fumagalli, M.: Assessing the effect of sequencing depth and sample size in population genetics inferences. PLOS ONE **8**(11), 1–11 (11 2013). https://doi.org/10.1371/journal.pone.0079667, https://doi.org/10.1371/journal.pone.0079667

[21] Fumagalli, M., Vieira, F.G., Korneliussen, T.S., Linderoth, T., Huerta-Sánchez, E., Albrechtsen, A., Nielsen, R.: Quantifying population genetic differentiation from next-generation sequencing data. Genetics **195**(3), 979–992 (2013). https://doi.org/10.1534/genetics.113.154740, https://www.genetics.org/content/195/3/979

[22] Fumagalli, M., Vieira, F.G., Linderoth, T., Nielsen, R.: ngsTools: methods for population genetics analyses from next-generation

sequencing data. Bioinformatics **30**(10), 1486–1487 (01 2014). https://doi.org/10.1093/bioinformatics/btu041, https://doi.org/10.1093/bioinformatics/btu041

[23] Garrison, E., Marth, G.: Haplotype-based variant detection from short-read sequencing (2012)

[24] Hardy, G.H.: Mendelian Proportions in a Mixed Population. Science, New Series **28**(706), 49–50 (1908)

[25] Jacqueline, K.W., Anna, P., Nancy, J.C.: Rational Inferences about Departures from Hardy-Weinberg Equilibrium. The American Journal of Human Genetics (6), 967–986 (2005)

[26] Lachance, J.: Detecting selection-induced departures from hardy-weinberg proportions. Genetics Selection Evolution (1), 15 (2009)

[27] Levy, S.E., Myers, R.M.: Advancements in next-generation sequencing. Annual Review of Genomics and Human Genetics **17**(1), 95–115 (2016). https://doi.org/10.1146/annurev-genom-083115-022413, https://doi.org/10.1146/annurev-genom-083115-022413, pMID: 27362342

[28] Li, C., Biswas, G.: Temporal Pattern Generation Using Hidden Markov Model Based Unsupervised Classification. In: IDA 1999: Advances in Intelligent Data Analysis, pp. 245–256. Springer, Berlin, Heidelberg (1999). https://doi.org/10.1007/3-540-48412-$4_2$1

[29] Li, H., J, R., Durbin, R.: Mapping short dna sequencing reads and calling variants using mapping quality scores. Genome research **11**(18), 1851–1858 (2008). https://doi.org/10.1101/gr.078212.108

[30] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., Subgroup, .G.P.D.P.: The Sequence Alignment/Map format and SAMtools. Bioinformatics **25**(16), 2078–2079 (06 2009). https://doi.org/10.1093/bioinformatics/btp352, https://doi.org/10.1093/bioinformatics/btp352

[31] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., Subgroup, .G.P.D.P.: The Sequence Alignment/Map format and SAMtools. Bioinformatics **25**(16), 2078–2079 (06 2009). https://doi.org/10.1093/bioinformatics/btp352, https://doi.org/10.1093/bioinformatics/btp352

[32] Loftus, B.J., Fung, E., Roncaglia, P., Rowley, D., Amedeo, P., Bruno, D., Vamathevan, J., Miranda, M., Anderson, I.J., Fraser, J.A., Allen, J.E., Bosdet, I.E., Brent, M.R., Chiu, R., Doering, T.L., Donlin, M.J., D'Souza, C.A., Fox, D.S., Grinberg, V., Fu, J., Fukushima, M., Haas, B.J., Huang, J.C., Janbon, G., Jones, S.J.M., Koo, H.L., Krzywinski, M.I., Kwon-Chung, J.K., Lengeler, K.B., Maiti, R., Marra, M.A., Marra, R.E., Mathewson, C.A., Mitchell, T.G., Pertea, M., Riggs, F.R., Salzberg, S.L., Schein, J.E., Shvartsbeyn, A., Shin, H., Shumway, M., Specht, C.A., Suh, B.B., Tenney, A., Utterback, T.R., Wickes, B.L., Wortman, J.R., Wye, N.H., Kronstad, J.W., Lodge, J.K., Heitman, J., Davis, R.W., Fraser, C.M., Hyman, R.W.: The genome of the basidiomycetous yeast and human pathogen ¡i¿cryptococcus neoformans¡/i¿. Science **307**(5713), 1321–1324 (2005). https://doi.org/10.1126/science.1103773

[33] Longley, N., Muzoora, C., Taseera, K., Mwesigye, J., Rwebembera, J., Chakera, A., Wall, E., Andia, I., Jaffar, S., Harrison, T.S.: Dose Response Effect of High-Dose Fluconazole for HIV-Associated Cryptococcal Meningitis in Southwestern Uganda. Clinical Infectious Diseases **47**(12), 1556–1561 (12 2008). https://doi.org/10.1086/593194, https://doi.org/10.1086/593194

[34] Lou, R.N., Jacobs, A., Wilder, A., Therkildsen, N.O.: A beginner's guide to low-coverage whole genome sequencing for population genomics. Molecular Ecology **n/a**(n/a) (2021). https://doi.org/https://doi.org/10.1111/mec.16077, https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.16077

[35] May, R.C., Stone, N.R., Wiesner, D.L., Bicanic, T., Nielsen, K.: Cryptococcus: from environmental saprophyte to global pathogen. Nature Reviews Microbiology **14**(2), 106–117 (Feb 2016). https://doi.org/10.1038/nrmicro.2015.6, https://doi.org/10.1038/nrmicro.2015.6

[36] McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., DePristo, M.A.: The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research **20**(9), 1297–303 (sep 2010). https://doi.org/10.1101/gr.107524.110, http://www.ncbi.nlm.nih.gov/pubmed/20644199

[37] McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., DePristo, M.A.: The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research **20**(9), 1297–303 (sep 2010). https://doi.org/10.1101/gr.107524.110, http://www.ncbi.nlm.nih.gov/pubmed/20644199

[38] Metzker, M.L.: Sequencing technologies — the next generation. Nature Reviews Genetics **11**(1), 31–46 (jan 2010). https://doi.org/10.1038/nrg2626, http://www.ncbi.nlm.nih.gov/pubmed/19997069

[39] Morrow, C.A., Fraser, J.A.: Ploidy variation as an adaptive mechanism in human pathogenic fungi. Seminars in Cell and Developmental Biology **24**(4), 339–346 (apr 2013)

[40] Nielsen, R., Paul, J., Albrechtsen, A., Song, Y.: Genotype and snp calling from next-generation sequencing data. Nature Reviews. Genetics **12**(6), 443–451 (2011). https://doi.org/10.1038/nrg2986

[41] Ormerod, K.L., Morrow, C.A., Chow, E.W.L., Lee, I.R., Arras, S.D.M., Schirra, H.J., Cox, G.M., Fries, B.C., Fraser, J.A.: Comparative genomics of serial isolates of cryptococcus neoformans reveals gene associated with carbon utilization and virulence. G3 (Bethesda, Md.) **3**(4), 675–686 (Apr 2013). https://doi.org/10.1534/g3.113.005660, https://pubmed.ncbi.nlm.nih.gov/23550133, 23550133[pmid]

[42] Van de Peer, Y., Mizrachi, E., Marchal, K.: The evolutionary significance of polyploidy. Nature Reviews Genetics **18**(7), 411–424 (Jul 2017). https://doi.org/10.1038/nrg.2017.26, https://doi.org/10.1038/nrg.2017.26

[43] Rabiner, L.: A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE **77**(2), 257–286 (1989). https://doi.org/10.1109/5.18626, http://ieeexplore.ieee.org/document/18626/

[44] Rhodes, J., Beale, M.A., Fisher, M.C.: Illuminating choices for library prep: A comparison of library preparation methods for whole genome sequencing of cryptococcus neoformans using illumina hiseq. PLOS ONE **9**(11), 1–9 (11 2014). https://doi.org/10.1371/journal.pone.0113501, https://doi.org/10.1371/journal.pone.0113501

[45] Rhodes, J., Beale, M.A., Vanhove, M., Jarvis, J.N., Kannambath, S., Simpson, J.A., Ryan, A., Meintjes, G., Harrison, T.S., Fisher, M.C., Bicanic, T.: A Population Genomics Approach to Assessing the Genetic Basis of Within-Host Microevolution Underlying Recurrent Cryptococcal Meningitis Infection. G3 Genes—Genomes—Genetics (2017). https://doi.org/10.1534/g3.116.037499, https://doi.org/10.1534/g3.116.037499

[46] Rhodes, J., Desjardins, C.A., Sykes, S.M., Beale, M.A., Vanhove, M., Sakthikumar, S., Chen, Y., Gujja, S., Saif, S., Chowdhary, A., Lawson, D.J., Ponzio, V., Colombo, A.L., Meyer, W., Engelthaler, D.M., Hagen, F., Illnait-Zaragozi, M.T., Alanio, A., Vreulink, J.M., Heitman, J., Perfect, J.R., Litvintseva, A.P., Bicanic, T., Harrison, T.S., Fisher, M.C., Cuomo, C.A.: Tracing Genetic Exchange and Biogeography of Cryptococcus neoformans var. grubii at the Global Population Level. Genetics **207**(1), 327–346 (07 2017). https://doi.org/10.1534/genetics.117.203836, https://doi.org/10.1534/genetics.117.203836

[47] Sattler, M.C., Carvalho, C.R., Clarindo, W.R.: The polyploidy and its key role in plant breeding. Planta (243), 281–296 (2016)

[48] Sionov, E., Chang, Y.C., Kwon-Chung, K.J.: Azole heteroresistance in cryptococcus neoformans: Emergence of resistant clones with chromosomal disomy in the mouse brain during fluconazole treatment. Antimicrobial Agents and Chemotherapy **57**(10), 5127–5130 (2013). https://doi.org/10.1128/AAC.00694-13, https://journals.asm.org/doi/abs/10.1128/AAC.00694-13

[49] Stone, N.R., Rhodes, J., Fisher, M.C., Mfinanga, S., Kivuyo, S., Rugemalila, J., Segal, E.S., Needleman, L., Molloy, S.F., Kwon-Chung, J., Harrison, T.S., Hope, W., Berman, J., Bicanic, T.: Dynamic ploidy changes drive fluconazole resistance in human cryptococcal meningitis. Journal of Clinical Investigation **129**(3), 999–1014 (mar 2019)

[50] Therkildsen, N.O., Palumbi, S.R.: Practical low-coverage genomewide sequencing of hundreds of individually barcoded samples for population and evolutionary genomics in nonmodel species. Molecular Ecology Resources **17**(2), 194–208 (2017). https://doi.org/https://doi.org/10.1111/1755-0998.12593, https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.12593

[51] Viterbi, A., A.: Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Transactions on Information The-

ory **13**(2), 260–269 (apr 1967). https://doi.org/10.1109/TIT.1967.1054010, http://ieeexplore.ieee.org/document/1054010/

[52] Vu, G.T.H., Cao, H.X., Reiss, B., Schubert, I.: Deletion-bias in dna double-strand break repair differentially contributes to plant genome shrinkage. New Phytologist **214**(4), 1712–1721 (2017). https://doi.org/https://doi.org/10.1111/nph.14490, https://nph.onlinelibrary.wiley.com/doi/abs/10.1111/nph.14490

[53] Weinberg, W.: Über den Nachweis der Vererbung beim Menschen. Jahresh. Ver. Vaterl. Naturkd. Württemb. **64**, 369–382 (1908)

[54] Weiß, C.L., Pais, M., Cano, L.M., Kamoun, S., Burbano, H.A.: nQuire: a statistical framework for ploidy estimation using next generation sequencing (2018). https://doi.org/10.1186/s12859-018-2128-z, https://doi.org/10.1186/s12859-018-2128-z

[55] Wood, T.E., Takebayashi, N., Barker, M.S., Mayrose, I., Greenspoon, P.B., H, R.L.: The frequency of polyploid speciation in vascular plants. Proc Natl Acad Sci USA (106), 13875–13879 (2009)

[56] Yang, F., Gritsenko, V., Lu, H., Zhen, C., Gao, L., Berman, J., ying Jiang, Y., Alanio, A.: Adaptation to fluconazole via aneuploidy enables cross-adaptation to amphotericin b and flucytosine in cryptococcus neoformans. Microbiology Spectrum **9**(2), e00723–21 (2021). https://doi.org/10.1128/Spectrum.00723-21, https://journals.asm.org/doi/abs/10.1128/Spectrum.00723-21

[57] Zhu, J., Tsai, H.J., Gordon, M.R., Li, R.: Cellular Stress Associated with Aneuploidy. Developmental cell **44**(4), 420–431 (feb 2018). https://doi.org/10.1016/j.devcel.2018.02.002, http://www.ncbi.nlm.nih.gov/pubmed/29486194