Peer Community In
Mathematical &
Computational
Biology

RESEARCH ARTICLE

Open Access
Open Peer-Review
Open Data
Open Code

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

Samuele Soraggi[1,2], Johanna Rhodes[3], Isin Altinkaya[2,4,5], Oliver Tarrant[2], François Balloux[6], Matthew C. Fisher[3], & Matteo Fumagalli[2,7]

[1] Bioinformatics Research Center (BiRC), University of Aarhus, 8000 Aarhus, Denmark
[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

Cite as:
Soraggi S, Rhodes J, Altinkaya I, Tarrant O, Balloux F, Fisher MC, Fumagalli M (2022) HMMploidy: inference of ploidy levels from short-read sequencing data. bioRxiv, 2021.06.29.450340, ver. 6 peer-reviewed and recommended by Peer Community in Mathematical and Computational Biology. https://doi.org/10.1101/2021.06.29.450340

Correspondence:
samuele@birc.au.dk
m.fumagalli@qmul.ac.uk

Recommender:
Alan Rogers

Reviewers:
Nicolas Galtier, Barbara Holland and Benjamin Peter

This version of the article has been peer-reviewed and recommended by *Peer Community In Mathematical and Computational Biology* (https://doi.org/10.24072/pci.mcb.100010)

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 (Levy and Myers, 2016; Metzker, 2010). Many statistical and computational methods, and accompanying software, to process NGS data for genotype and variant calling have been proposed (Garrison and Marth, 2012; Li et al., 2009; Van der Auwera et al., 2013). Additionally, dedicated software have been developed to analyse low-coverage sequencing data (Fumagalli, Vieira, Linderoth, et al., 2014; Nielsen et al., 2011), a popular and cost-effective approach in population genomic studies (Lou et al., 2021). 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 (Van de Peer et al., 2017). Polyploidy is common to many organisms, and it can be the consequence of hybridisation or whole genome duplication (Fox et al., 2020). For instance, polyploidy plays a significant role in the evolution and speciation of plants (Sattler et al., 2016), as 34.5% of vascular plants (including leading commercial crop species) are shown to be polyploid (Wood et al., 2009).

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 (Coward and Harding, 2014), and it is a phenomenon commonly detected in cancer cells (Davoli and Lange, 2011) and several pathogenic fungi (i.e. *Cryptococcus neoformans*, *Candida albicans* and *Candida glabrata*) and monocellular parasites (Avramovska et al., 2021; Farrer et al., 2013; Fu et al., 2021; Morrow and Fraser, 2013; Stone et al., 2019; Yang et al., 2021; Zhu et al., 2018).

Among aneuploid species, *Cryptococcus neoformans* is a fungal pathogen capable of causing meningitis in immunocompromised individuals, particularly HIV/AIDS patients (May et al., 2016). 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 (Morrow and Fraser, 2013). Aneuploidy-driven heteroresistance to the frontline antifungal drug fluconazole has been described (Stone et al., 2019), resulting in treatment failure in patients. Within fluconazole resistant colonies, aneuploidy was common, particularly disomy of chromosome 1 which harbours the gene encoding the main drug target of fluconazole, *ERG11* (Stone et al., 2019). 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` (Weiß et al., 2018)), comparing frequencies and coverage to a reference data set (`ploidyNGS` (Augusto Corrêa dos Santos et al., 2017)), 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` (Bao et al., 2014), `sequenza` (Favero et al., 2015)). 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 (Augusto Corrêa dos Santos et al., 2017). 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 (Therkildsen and Palumbi, 2017), 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) (Rabiner, 1989) where the emissions are both sequencing depth levels and observed reads. The latter are translated into genotype likelihoods (Nielsen et al., 2011) 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` (Fumagalli, Vieira, Linderoth, et al., 2014), 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. $m = 1, \ldots, M$.

Probability of sequenced data

Let $O = (O_1, \ldots, O_M)$ 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 $\{0, \ldots, Y\}$, where each value is the number of alternate (or derived) alleles of said genotype.

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

$$p(O|Y,F) = \sum_{G \in \{0,\ldots,Y\}} p(O|G,Y)p(G|Y,F), \tag{1}$$

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 (McKenna et al., 2010). Note that the only varying parameter in it is the genotype; therefore it is usually rewritten as $L(G|O,Y)$. The second factor is the probability of the genotype given the population allele 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 (Hardy, 1908; Weinberg, 1908). 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 (Chen et al., 2017; Lachance, 2009; Wittke-Thompson et al., 2005). 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 $Y_m^{(1)}, \dots, Y_m^{(K)}$, 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 (Bishop, 2006; Casella and Berger, 2002) (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 (Anders and Huber, 2010).

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.

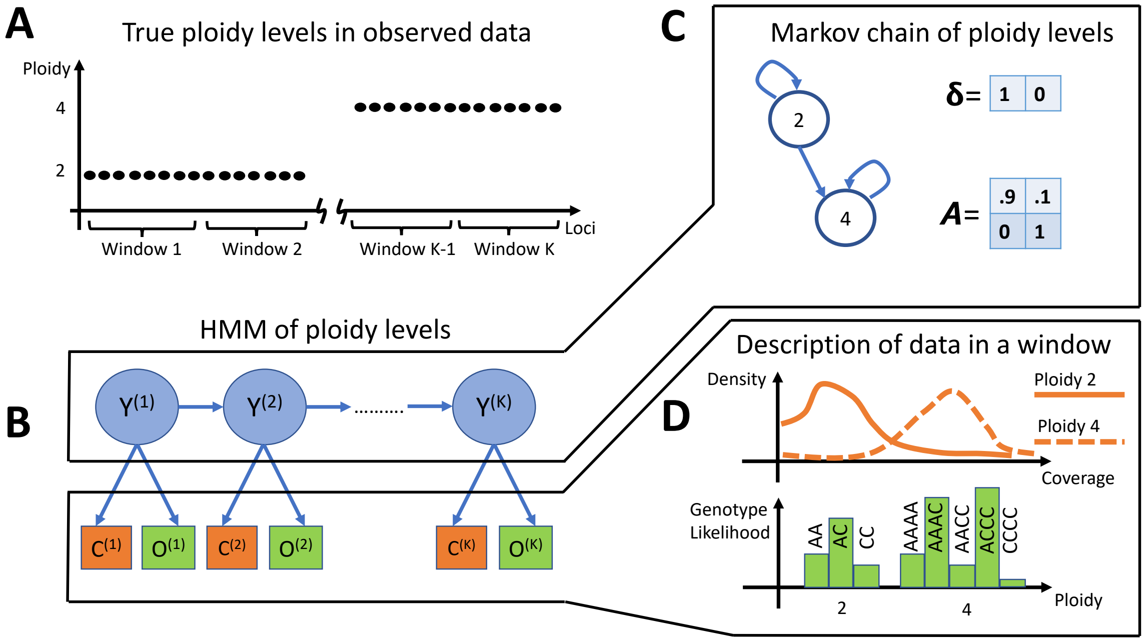
We propose a heuristic optimisation algorithm to automatically find the number of latent states of the HMM, and to assign them to the correct ploidy through the genotype likelihoods. Our implementation, described in the Supplementary Material, is a heuristic version of the well-known Expectation Conditional Maximisation (ECM) algorithm (Cappe et al., 2005).

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 (Fumagalli, Vieira, Korneliussen, et al., 2013). 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 combination of ploidy (from 1 to 5, constant along each genome),

Figure 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 $Y^{(1)}, \dots, Y^{(K)}$ - one for each window of loci - that can assume one of two values of the ploidies. Observations $C^{(1)}, \dots, C^{(K)}$ and $O^{(1)}, \dots, O^{(K)}$ 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.



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 (Rhodes, Beale, Vanhove, et al., 2017). Whole-genome sequencing data was performed on an Illumina machine following an established protocol for sample preparation (Rhodes, Beale, and Fisher, 2014) and data processing (Rhodes, Beale, Vanhove, et al., 2017). Reads are mapped onto *C. neoformans* H99 reference genome (Loftus et al., 2005), 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 (Augusto Corrêa dos Santos et al., 2017) and nQuire (including a version with denoising option, nQuire.Den) (Weiß et al., 2018). 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.

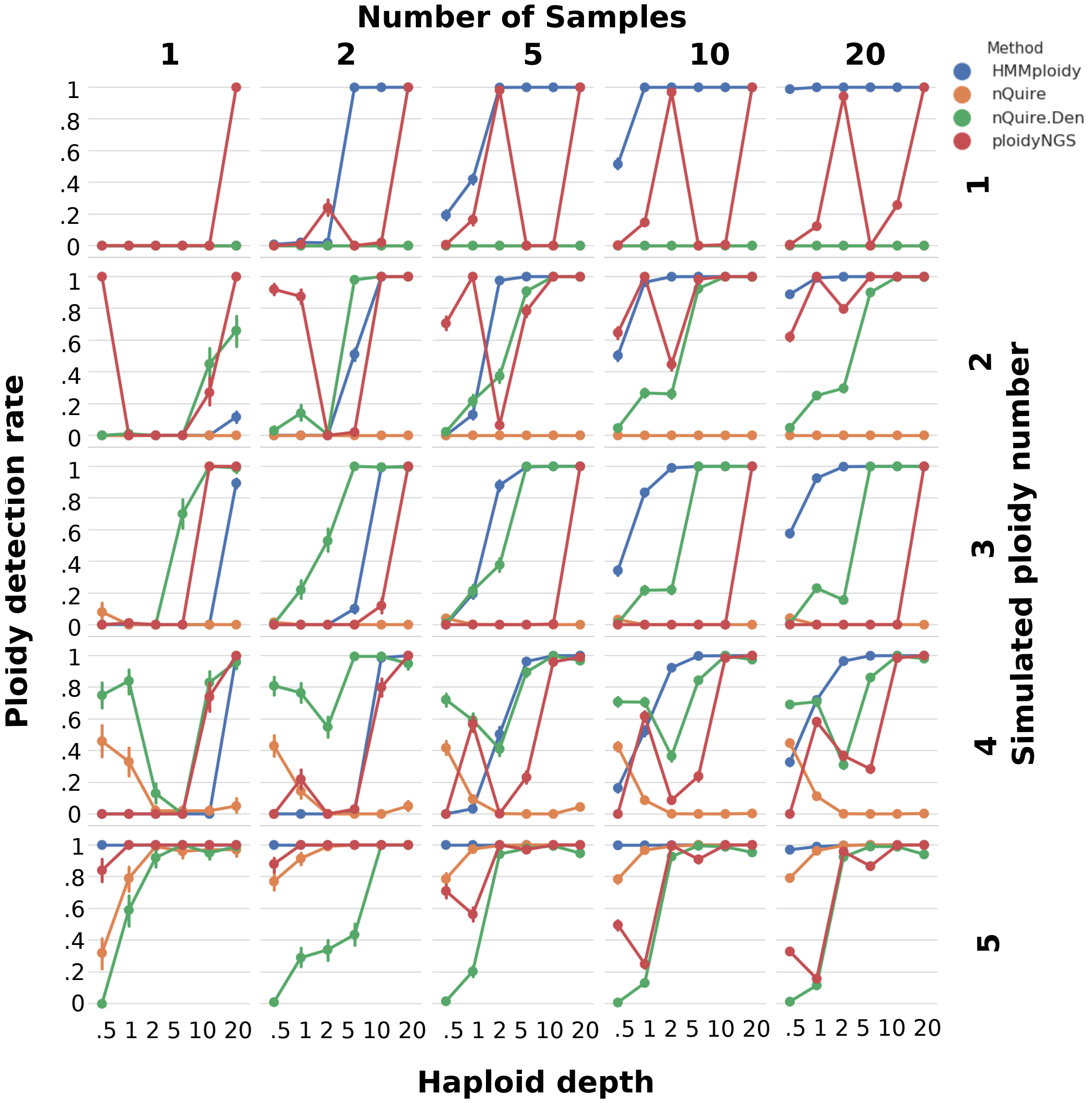
Note also that none of the methods performs well with a single haploid sample. This happens because many loci show only one possible genotype, and even with the genotype likelihoods it is impossible to determine the multiplicity of the ploidy. With more samples it is possible to exploit loci with at 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 $O(|\mathcal{Y}|^2K)$ 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.

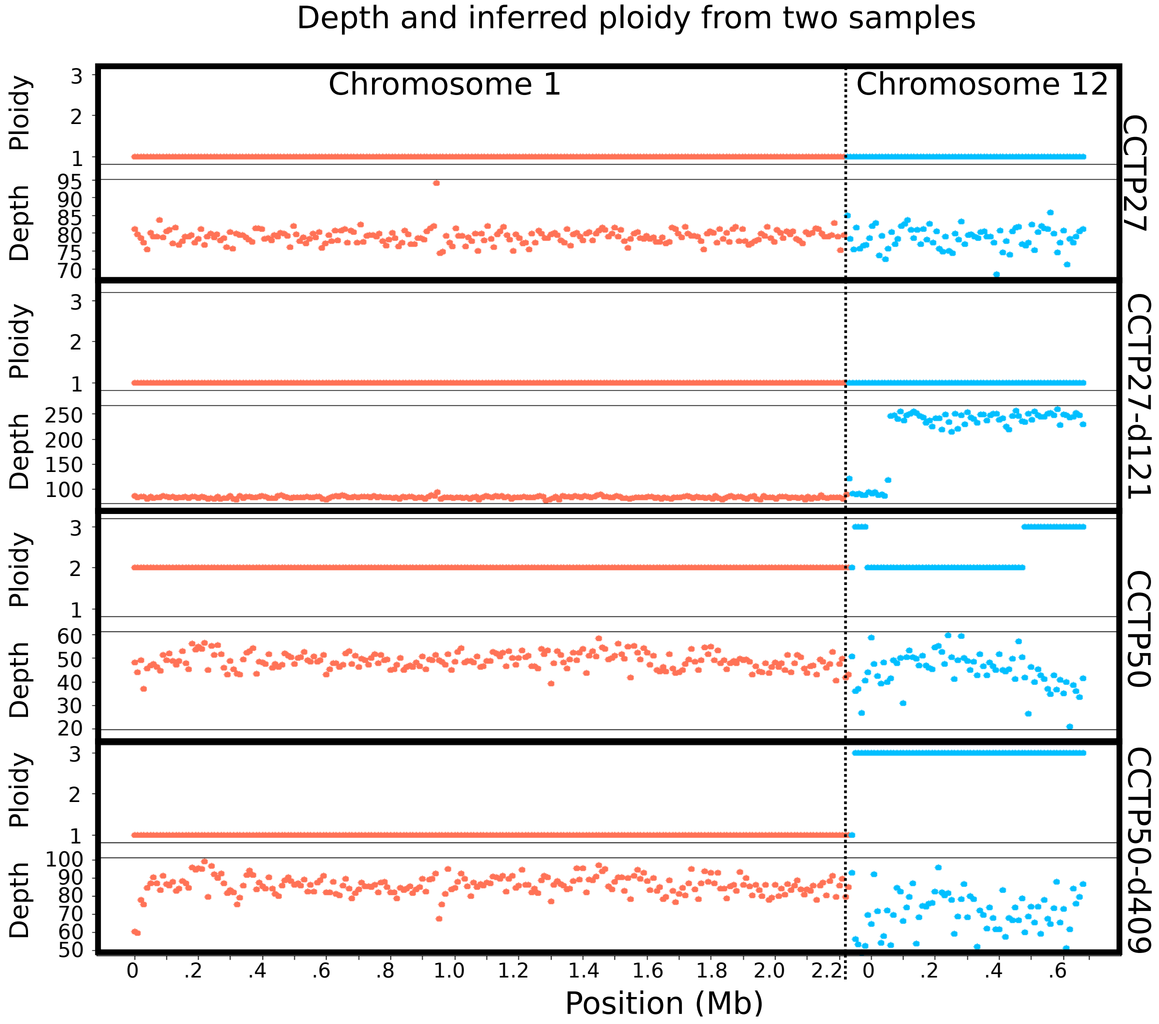
Figure 2. : Comparison of ploidy detection rates for different methods at various experimental scenarios. The rate of detecting the correct ploidy (y-axis) is shown against the haploid sequencing depth (x-axis) for different sample sizes (on columns) and ploidy levels (on rows). For every simulated ploidy 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.



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.

Figure 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.



Application to real data

We used HMMploidy to infer ploidy variation in 23 isolates of *Cryptococcus neoformans* recovered from HIV-infected patients (Rhodes, Beale, Vanhove, et al., 2017). By analysing variation in normalised sequencing coverage, Rhodes and coworkers identified extensive instances of aneuploidy, especially on chromosome 12, in several pairs of isolates (Rhodes, Beale, Vanhove, et al., 2017), in line with previous findings using karyotypic analysis (Ormerod et al., 2013). 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 (Rhodes, Beale, Vanhove, et al., 2017), 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 downsampling the sequencing data to 20% of the original data set (Fig. S6-S28). Interestingly, `ploidyNGS`, `nQuire` and `nQuire.Den` infer 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 (May et al., 2016). 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 (Longley et al., 2008). Resistance to antifungal drugs is thought to be responsible for such poor outcomes and relapse episodes, but its molecular mechanisms are not yet understood (Stone et al., 2019). Resistance to oral fluconazole antifungal drugs in *Cryptococcus neoformans* was associated with aneuploidy (Sionov et al., 2013). Recent genomic studies identified multiple occurrences of aneuploidy in resistant and relapse isolates (Stone et al., 2019). 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 (Rhodes, Desjardins, et al., 2017) 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 (Ben-David and Amon, 2020), or partial changes of copy numbers in polyploid genomes due to deletions or duplications (Vu et al., 2017).

Acknowledgements

Preprint version 6 of this article has been peer-reviewed and recommended by Peer Community In Mathematical and Computational Biology (https://doi.org/10.24072/pci.mcb.100010). We are grateful to Alan Rogers, Barbara Holland, Benjamin Peter, Nicolas Galtier, and several anonymous reviewers for improving the manuscript. We wish to thank Anders Albrechtsen for providing feedback on the formulation of genotype likelihoods. We would like to thank GenomeDK and Aarhus University for providing computational resources and support.

Fundings

SS was supported by a grant from Novo Nordisk Foundation (NNF20OC0063268). 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.

Conflict of interest disclosure

The authors declare that they comply with the PCI rule of having no financial conflicts of interest in relation to the content of the article. The authors declare the following non-financial conflict of interest: Matteo Fumagalli and François Balloux are recommenders of PCI.

Data, script and code availability

Data are available online: https://doi.org/10.17605/OSF.IO/5F7AR
Script and codes are available online: https://doi.org/10.5281/zenodo.7116023

Supplementary information availability

Supplementary information is available online: https://doi.org/10.1101/2021.06.29.450340

References

Anders S and W Huber (2010). Differential expression analysis for sequence count data. *Genome biology* 11. gb-2010-11-10-r106[PII], R106–R106. https://doi.org/10.1186/gb-2010-11-10-r106.

Augusto Corrêa dos Santos R, GH Goldman, and DM Riaño-Pachón (2017). ploidyNGS: visually exploring ploidy with Next Generation Sequencing data. *Bioinformatics* 33, 2575–2576. https://doi.org/10.1093/bioinformatics/btx204.

Avramovska O, E Rego, and MA Hickman (2021). Tetraploidy accelerates adaption under drug-selection in a fungal pathogen. *bioRxiv*. https://doi.org/10.1101/2021.02.28.433243.

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

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

Bishop C (2006). *Pattern Recognition and Machine Learning*. Springer.

Cappe O, E Moulines, and T Ryden (2005). *Inference in Hidden Markov Models*. Springer Science+Business Media, Inc.

Casella G and RL Berger (2002). *Statistical inference*. Thomson Learning, p. 660.

Chen B, JW Cole, and C Grond-Ginsbach (2017). Departure from Hardy Weinberg Equilibrium and Genotyping Error. *Frontiers in Genetics* 8. https://doi.org/10.3389/fgene.2017.00167.

Coward J and A Harding (2014). Size Does Matter: Why Polyploid Tumor Cells are Critical Drug Targets in the War on Cancer. *Frontiers in Oncology* 4, 123. https://doi.org/10.3389/fonc.2014.00123.

Davoli T and T de Lange (2011). The Causes and Consequences of Polyploidy in Normal Development and Cancer. *Annual Review of Cell and Developmental Biology* 27. PMID: 21801013, 585–610. https://doi.org/10.1146/annurev-cellbio-092910-154234.

Farrer RA, DA Henk, TWJ Garner, F Balloux, DC Woodhams, and MC Fisher (2013). Chromosomal Copy Number Variation, Selection and Uneven Rates of Recombination Reveal Cryptic Genome Diversity Linked to Pathogenicity. *PLoS Genetics* 9. Ed. by Heitman J, e1003703. https://doi.org/10.1371/journal.pgen.1003703.

Favero F, T Joshi, AM Marquard, NJ Birkbak, M Krzystanek, Q Li, Z Szallasi, and AC Eklund (2015). Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Annals of Oncology* 26, 64–70. https://doi.org/10.1093/annonc/mdu479.

Fox DT, DE Soltis, PS Soltis, TL Ashman, and Y Van de Peer (2020). Polyploidy: A Biological Force From Cells to Ecosystems. *Trends in Cell Biology* 30, 688–694. https://doi.org/10.1016/j.tcb.2020.06.006.

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

Fumagalli M, FG Vieira, TS Korneliussen, T Linderoth, E Huerta-Sánchez, A Albrechtsen, and R Nielsen (2013). Quantifying Population Genetic Differentiation from Next-Generation Sequencing Data. *Genetics* 195, 979–992. https://doi.org/10.1534/genetics.113.154740.

Fumagalli M, FG Vieira, T Linderoth, and R Nielsen (2014). ngsTools: methods for population genetics analyses from next-generation sequencing data. *Bioinformatics* 30, 1486–1487. https://doi.org/10.1093/bioinformatics/btu041.

Garrison E and G Marth (2012). Haplotype-based variant detection from short-read sequencing. https://doi.org/10.48550/ARXIV.1207.3907.

Hardy GH (1908). Mendelian Proportions in a Mixed Population. *Science* 28, 49–50. https://doi.org/10.1126/science.28.706.49.

Lachance J (2009). Detecting selection-induced departures from Hardy-Weinberg proportions. *Genetics Selection Evolution* 41, 15. https://doi.org/10.1186/1297-9686-41-15.

Levy SE and RM Myers (2016). Advancements in Next-Generation Sequencing. *Annual Review of Genomics and Human Genetics* 17. PMID: 27362342, 95–115. https://doi.org/10.1146/annurev-genom-083115-022413.

Li H, B Handsaker, A Wysoker, T Fennell, J Ruan, N Homer, G Marth, G Abecasis, R Durbin, and 1GPDP Subgroup (June 2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

Loftus BJ, E Fung, P Roncaglia, D Rowley, P Amedeo, D Bruno, J Vamathevan, M Miranda, IJ Anderson, JA Fraser, JE Allen, IE Bosdet, MR Brent, R Chiu, TL Doering, MJ Donlin, CA D'Souza, DS Fox, V Grinberg, J Fu, M Fukushima, BJ Haas, JC Huang, G Janbon, SJM Jones, HL Koo, MI Krzywinski, JK Kwon-Chung, KB Lengeler, R Maiti, MA Marra, RE Marra, CA Mathewson, TG Mitchell, M Pertea, FR Riggs, SL Salzberg, JE Schein, A Shvartsbeyn, H Shin, M Shumway, CA Specht, BB Suh, A Tenney, TR Utterback, BL Wickes, JR Wortman, NH Wye, JW Kronstad, JK Lodge, J Heitman, RW Davis, CM Fraser, and RW Hyman (2005). The Genome of the Basidiomycetous Yeast and Human Pathogen Cryptococcus neoformans. *Science* 307, 1321–1324. https://doi.org/10.1126/science.1103773.

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

Lou RN, A Jacobs, AP Wilder, and NO Therkildsen (2021). A beginner's guide to low-coverage whole genome sequencing for population genomics. *Molecular Ecology* 30, 5966–5993. https://doi.org/https://doi.org/10.1111/mec.16077.

May RC, NR Stone, DL Wiesner, T Bicanic, and K Nielsen (2016). Cryptococcus: from environmental saprophyte to global pathogen. *Nature Reviews Microbiology* 14, 106–117. https://doi.org/10.1038/nrmicro.2015.6.

McKenna A, M Hanna, E Banks, A Sivachenko, K Cibulskis, A Kernytsky, K Garimella, D Altshuler, S Gabriel, M Daly, and MA DePristo (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* 20, 1297–303. https://doi.org/10.1101/gr.107524.110.

Metzker ML (2010). Sequencing technologies — the next generation. *Nature Reviews Genetics* 11, 31–46. https://doi.org/10.1038/nrg2626.

Morrow CA and JA Fraser (2013). Ploidy variation as an adaptive mechanism in human pathogenic fungi. *Seminars in Cell and Developmental Biology* 24, 339–346.

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

Ormerod KL, CA Morrow, EWL Chow, IR Lee, SDM Arras, HJ Schirra, GM Cox, BC Fries, and JA Fraser (2013). Comparative Genomics of Serial Isolates of Cryptococcus neoformans Reveals Gene Associated With Carbon Utilization and Virulence. *G3 Genes|Genomes|Genetics* 3, 675–686. https://doi.org/10.1534/g3.113.005660.

Rabiner L (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* 77, 257–286. https://doi.org/10.1109/5.18626.

Rhodes J, MA Beale, M Vanhove, JN Jarvis, S Kannambath, JA Simpson, A Ryan, G Meintjes, TS Harrison, MC Fisher, and T Bicanic (2017). A Population Genomics Approach to Assessing the Genetic Basis of Within-Host Microevolution Underlying Recurrent Cryptococcal Meningitis Infection. *G3 Genes|Genomes|Genetics*. https://doi.org/10.1534/g3.116.037499.

Rhodes J, MA Beale, and MC Fisher (2014). Illuminating Choices for Library Prep: A Comparison of Library Preparation Methods for Whole Genome Sequencing of Cryptococcus neoformans Using Illumina HiSeq. *PLOS ONE* 9, 1–9. https://doi.org/10.1371/journal.pone.0113501.

Rhodes J, CA Desjardins, SM Sykes, MA Beale, M Vanhove, S Sakthikumar, Y Chen, S Gujja, S Saif, A Chowdhary, DJ Lawson, V Ponzio, AL Colombo, W Meyer, DM Engelthaler, F Hagen, MT Illnait-Zaragozi, A Alanio, JM Vreulink, J Heitman, JR Perfect, AP Litvintseva, T Bicanic, TS Harrison, MC Fisher, and CA Cuomo (July 2017). Tracing Genetic Exchange and Biogeography of Cryptococcus neoformans var. grubii at the Global Population Level. *Genetics* 207, 327–346. https://doi.org/10.1534/genetics.117.203836.

Sattler MC, CR Carvalho, and WR Clarindo (2016). The polyploidy and its key role in plant breeding. *Planta* 243, 281–296. https://doi.org/10.1007/s00425-015-2450-x.

Sionov E, YC Chang, and KJ Kwon-Chung (2013). Azole Heteroresistance in Cryptococcus neoformans: Emergence of Resistant Clones with Chromosomal Disomy in the Mouse Brain during Fluconazole Treatment. *Antimicrobial Agents and Chemotherapy* 57, 5127–5130. https://doi.org/10.1128/AAC.00694-13.

Stone NR, J Rhodes, MC Fisher, S Mfinanga, S Kivuyo, J Rugemalila, ES Segal, L Needleman, SF Molloy, J Kwon-Chung, TS Harrison, W Hope, J Berman, and T Bicanic (2019). Dynamic ploidy changes drive fluconazole resistance in human cryptococcal meningitis. *The Journal of Clinical Investigation* 129, 999–1014. https://doi.org/10.1172/JCI124516.

Therkildsen NO and SR Palumbi (2017). Practical low-coverage genomewide sequencing of hundreds of individually barcoded samples for population and evolutionary genomics in nonmodel species. *Molecular Ecology Resources* 17, 194–208. https://doi.org/https://doi.org/10.1111/1755-0998.12593.

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

Van der Auwera GA, MO Carneiro, C Hartl, R Poplin, G del Angel, A Levy-Moonshine, T Jordan, K Shakir, D Roazen, J Thibault, E Banks, KV Garimella, D Altshuler, S Gabriel, and MA DePristo (2013). From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Current Protocols in Bioinformatics* 43, 11.10.1–11.10.33. https://doi.org/https://doi.org/10.1002/0471250953.bi1110s43.

Vu GTH, HX Cao, B Reiss, and I Schubert (2017). Deletion-bias in DNA double-strand break repair differentially contributes to plant genome shrinkage. *New Phytologist* 214, 1712–1721. https://doi.org/https://doi.org/10.1111/nph.14490.

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

Weiß CL, M Pais, LM Cano, S Kamoun, and HA Burbano (2018). nQuire: a statistical framework for ploidy estimation using next generation sequencing. *BMC Bioinformatics* 19, 122. https://doi.org/10.1186/s12859-018-2128-z.

Wittke-Thompson JK, A Pluzhnikov, and NJ Cox (2005). Rational Inferences about Departures from Hardy-Weinberg Equilibrium. *The American Journal of Human Genetics* 76, 967–986. https://doi.org/https://doi.org/10.1086/430507.

Wood TE, N Takebayashi, MS Barker, I Mayrose, PB Greenspoon, and LH Rieseberg (2009). The frequency of polyploid speciation in vascular plants. *Proceedings of the National Academy of Sciences* 106, 13875–13879. https://doi.org/10.1073/pnas.0811575106.

Yang F, V Gritsenko, H Lu, C Zhen, L Gao, J Berman, Yy Jiang, and A Alanio (2021). Adaptation to Fluconazole via Aneuploidy Enables Cross-Adaptation to Amphotericin B and Flucytosine in Cryptococcus neoformans. *Microbiology Spectrum* 9, e00723–21. https://doi.org/10.1128/Spectrum.00723-21.

Zhu J, HJ Tsai, MR Gordon, and R Li (2018). Cellular Stress Associated with Aneuploidy. *Developmental cell* 44, 420–431. https://doi.org/10.1016/j.devcel.2018.02.002.