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 · poliploidy · 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 (34; 25). Many statistical and computational methods, and accompanying software, to process NGS data for genotype and SNP calling have been proposed (27; 21; 3). Additionally, dedicated software have been developed to analyse low-coverage sequencing data (36; 17), a popular and cost-effective approach in population genomic studies (30). 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 (37). Polyploidy is common to many organisms at different genic and cellular levels, and it can be the consequence of hybridisation or whole genome duplication (15). For instance, polyploidy plays a significant role in the evolution and speciation of plants (42), as 34.5% of vascular plants (including leading commercial crop species) are shown to be polyploid (49).

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 (10), and it is a phenomenon commonly detected in cancer cells (11) and several pathogenic fungi (i.e. *Cryptococcus neoformans*, *Candida albicans* and *Candida glabrata*) and monocellular parasites (44; 35; 13; 51; 50; 16; 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 (35). Aneuploidy-driven heteroresistance to the frontline antifungal drug fluconazole has been described (44), 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* (44). 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` (48)), comparing frequencies and coverage to a reference data set (`ploidyNGS` (2)), or using inferred genotypes and information on GC content - although this is an approach specific for detecting aberrations in cancer genomes (e.g. `AbsCN-seq` (5), `sequenza` (14)). 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), often need a reference data at known ploidy set to be compared to, and 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 popular and cost-effective strategy in the population genetics of both model and non-model species (45), a tool that incorporate 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) (38) where the emissions are both sequencing depth levels and observed reads. The latter are translated into genotype likelihoods (36) 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` is written in `R/C++` and `python`. Source code is freely available at https://github.com/SamueleSoraggi/HMMploidy, integrated into `ngsTools` (20), 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, 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 Next Generation Sequencing (NGS) data for M sequenced genomes at N sites. Consider a fixed m-th genome and n-th locus. For such genome and locus define $Y_{m,n}$, $G_{m,n}$ and $O_{m,n}$ as the ploidy, genotype and sequencing data, respectively. Given $Y_{m,n}$, the genotype $G_{m,n}$ assumes values in $\{0, 1, ..., Y_{m,n}\}$, i.e. the number of alternate (or derived) alleles of the genotype. Any of the two alleles can be considered to define $G_{m,n}$.

The likelihood of the sequenced data, conditionally on the ploidy $Y_{m,n}$ and the population frequency F_n at locus n, is expressed by

$$p(O_{m,n}|Y_{m,n}, F_n) = \sum_{G_{m,n}\in\{0,\dots,Y_{m,n}\}} p(O_{m,n}|G_{m,n}, Y_{m,n})p(G_{m,n}|Y_{m,n}, F_n), \quad (1)$$

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 (32); the second factor is the probability of the genotype given the population frequency and the ploidy number. 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 (22; 47). 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 (9; 23; 24). However, we will leave the treatment of DHW for the inference of ploidy variation in 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 algorithm analysis 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 number of ploidy levels (i.e. cardinality of $\mathcal{Y}$) (Fig. 1C).

In the HMM structure, each k-th ploidy emits two observations. Those contain a dependency on which ploidy is assigned to that window. The observations consist of the sequenced reads $O_{m,n}^{(k)}$ and the average sequencing depth $C_{m,n}^{(k)}$ in the k-th window (Fig. 1B). The former is modelled by the probability in Equation 1; the latter by a Poisson-Gamma distribution (6; 8) (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.

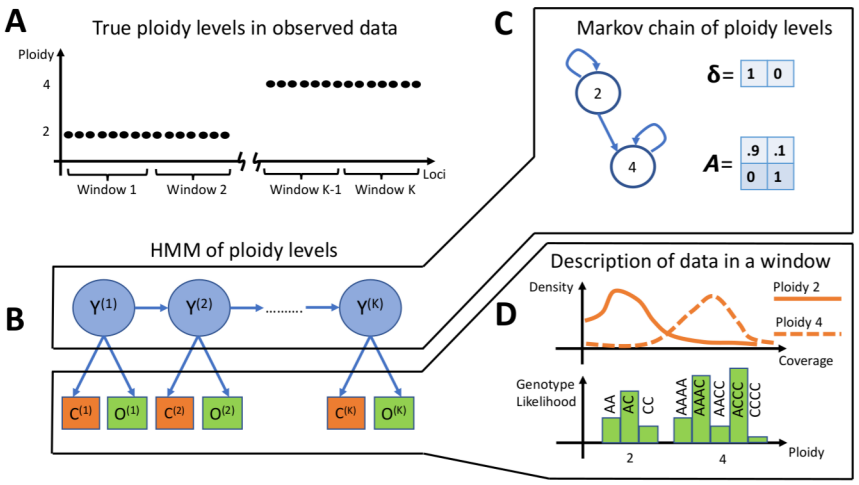


Fig. 1: **HMM for two ploidy levels.** (A) Consider a NGS dataset consisting of a sequence of two ploidy levels. (B) The HMM describing the data has a sequence of hidden states $Y^{(1)}, \ldots, Y^{(K)}$ - one for each window of loci - that can assume one of two values of the ploidies. Observations $C^{(1)}, \ldots, C^{(K)}$ and $O^{(1)}, \ldots, O^{(K)}$ describe respectively the coverage and observed reads in each window. 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 coverage distribution of the loci. 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.

This allows modelling 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 (7).

Simulated data

The assessment of memory, runtime and ploidy detection power of `HMMploidy` compared to the other methods is performed on a wide range of simulated scenarios. We simulated sequencing data under a wide range of scenarios using a previously proposed approach (19; 18). Specifically, each locus is treated as an independent observation, without modelling the effect of linkage disequilibrium. The number of reads is modeled with a Poisson distribution with parameter given by the input coverage multiplied by the ploidy level. At each locus, individual genotypes are randomly drawn according to a probability distribution defined by 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), 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 of runtimes and memory usage consist of 100 diploid genomes of length 10kb, 100kb, 1Mb, 10Mb. Each simulated genome has of the expected density 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 (40). Whole-genome sequencing data was performed on an Illumina machine following an established protocol for sample preparation (39) and data processing (40). Reads are mapped onto *C. neoformans* H99 reference genome (28), 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`) (48). 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` reaches maximum power with 20 individuals for 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 make it easier to compare the data to a simulated genome with low ploidy number 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 perform 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).While ploidy is not expected to vary within each chromosome, the distribution of local ploidy tracts as inferred by `HMMploidy` can provide statistical support to whole-chromosome estimates. Additionally, any detected within-chromosome ploidy variation can serve as a diagnostic tool to investigate possible mapping or assembly errors.

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 coverage or larger sample size is likely to be re-

quired for HMMploidy to detect higher levels of ploidy, as the power of the method lowers with increasing ploidy (Fig. S4).

Computational performance

The benchmark of HMMploidy shows 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 corresponding to 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 HECM algorithms has a slow convergence. However, training a separate HMM on each genome allows to overcome two main issues: sequencing data at different depths, and variations in ploidy among sampled genomes. On the former point, assume one could limit itself to define only one HMM for all genomes. To do this, sequencing depth should be standardized in each genome. However, this would make the estimation of the distributions of standardized counts very difficult, especially in samples with a lot of noise, errors, and/or low-depth sequencing. 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 parametersThe 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 tuntimes. Such options might be implemented in future versions of the software.

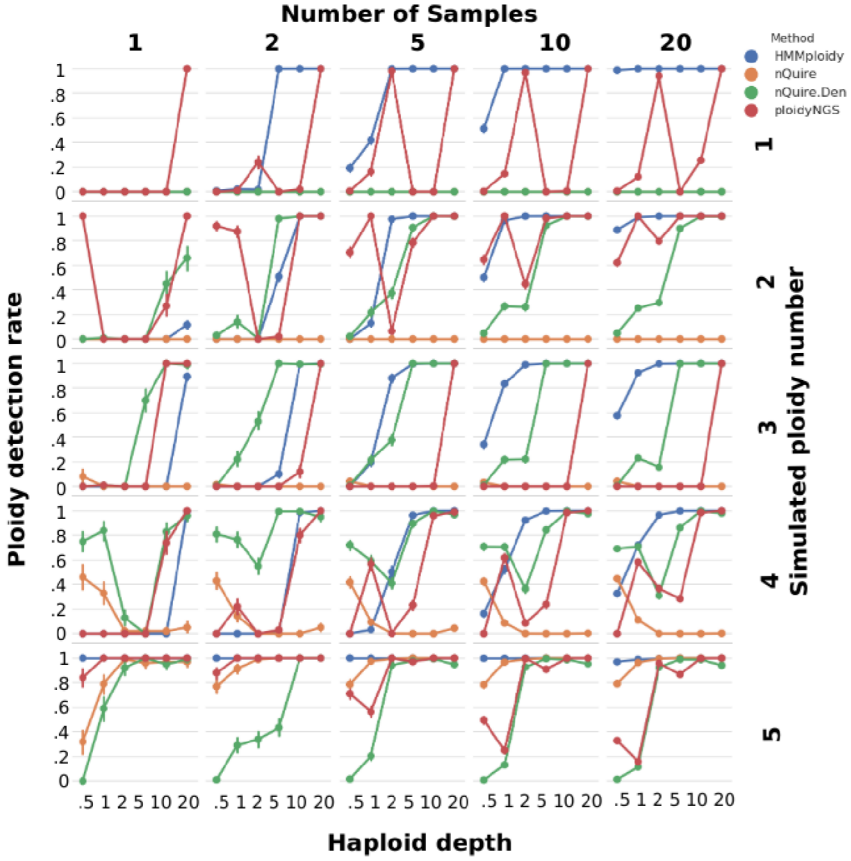


Fig. 2: **Comparison of ploidy detection rates for different methods at various experimental scenarios.** The rate of detecting the correct ploidy (y-axis) is shown against the haploid sequencing depth (x-axis) for different sample sizes (on columns) and simulated ploidy levels (on rows). For every simulated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

Application to real data

We used `HMMploidy` to infer ploidy variation in 23 isolates of *Cryptococcus neoformans* recovered from HIV-infected patients (40). In accordance with the original study (40), 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). This finding suggests the presence of a recent copy number variation. In fact, as no sufficient genetic variation has built up on the recently duplicated triploid chromosome yet, the data is modeled as a single chromosome by the genotype likelihoods. Sample CCTP50 has on average a higher depth at day 409, but chromosome 1 changes from diploid (day 1) to haploid (day 409). Chromosome 12 is 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 when artificially down-sampling 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 (31). 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 (29). Resistance to antifungal drugs is thought to be responsible for such poor outcomes and relapse episodes, but its molecular mechanisms are not yet understood (44). Resistance to oral fluconazole antifungal drugs in *Cryptococcus neoformans* was associated with aneuploidy (43). Recent genomic studies identified multiple occurrences of aneuploidy in resistant and relapse isolates (44). 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 provising novel insight into the pathogen's evolution (41) 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.

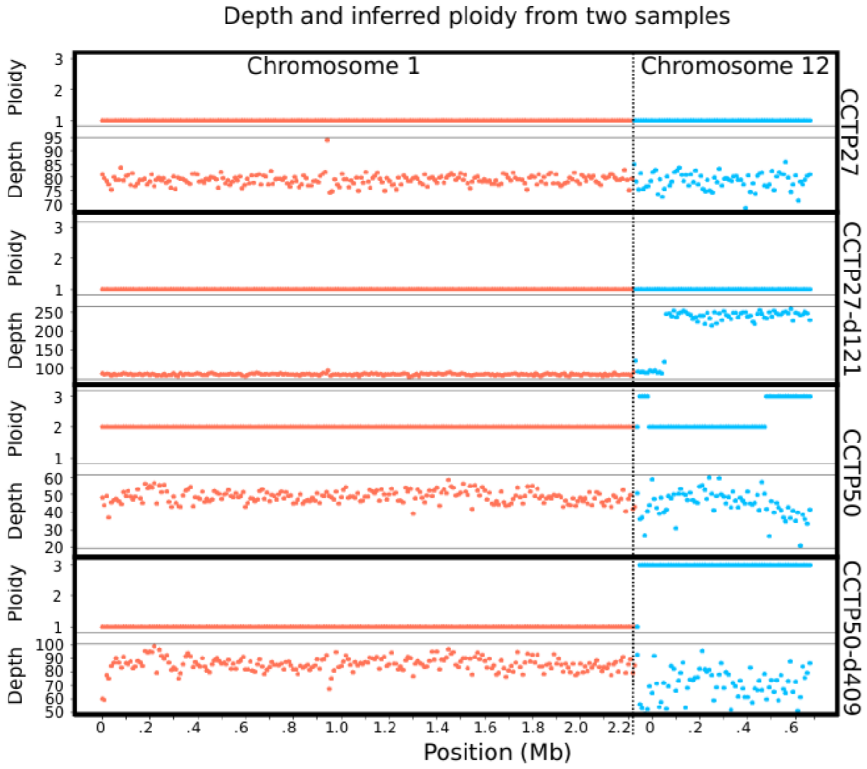


Fig. 3: **Inference of ploidy levels on two samples of *Cryptococcus neoformans* at different time points using `HMMploidy`.** Inferred ploidy and corresponding sequencing depth are shown in genomic windows for two samples at day 1 (CCTP27 and CCTP50) and day 121 (CCTP27-d121) and 409 (CCTP50-d409) on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

Acknowledgements

We are grateful to Alan Rogers, Benjamin Peter, and Nicolas Galtier for reviewing 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.

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

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

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

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

[10] 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

[11] 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

[12] 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

[13] 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

[14] 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

[15] 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

[16] 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

[17] 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)

[18] 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

[19] 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

[20] 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

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

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

[23] 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)

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

[25] 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

[26] 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

[27] 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

[28] 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

[29] 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

[30] 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

[31] 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

[32] 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

[33] 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

[34] 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

[35] 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)

[36] 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

[37] 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

[38] 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/

[39] 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

[40] 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

[41] 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 Crypto-

coccus 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

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

[43] 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

[44] 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)

[45] 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

[46] Viterbi, A., A.: Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Transactions on Information Theory **13**(2), 260–269 (apr 1967). https://doi.org/10.1109/TIT.1967.1054010, http://ieeexplore.ieee.org/document/1054010/

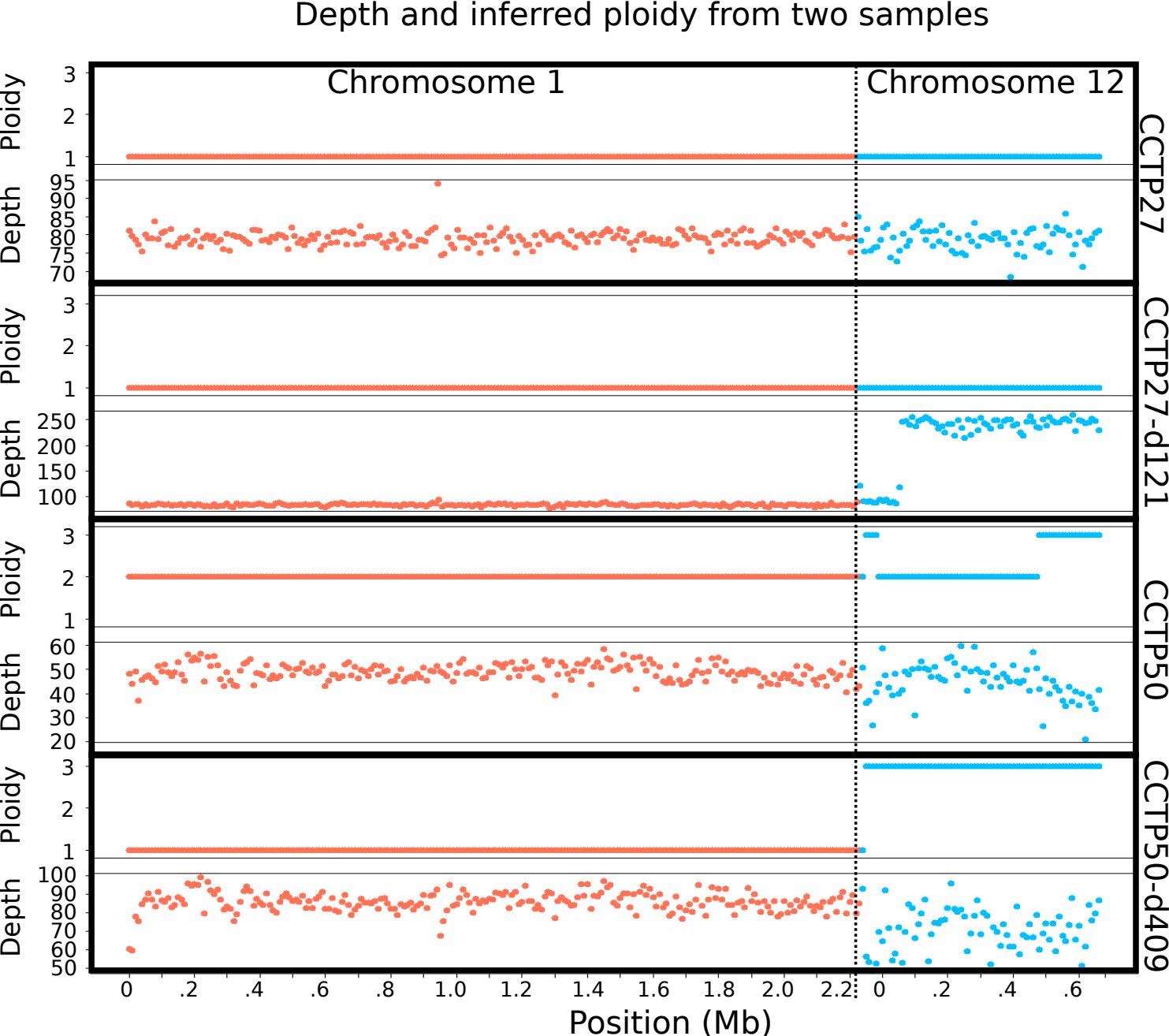
[47] Weinberg, W.: Über den Nachweis der Vererbung beim Menschen. Jahresh. Ver. Vaterl. Naturkd. Württemb. **64**, 369–382 (1908)

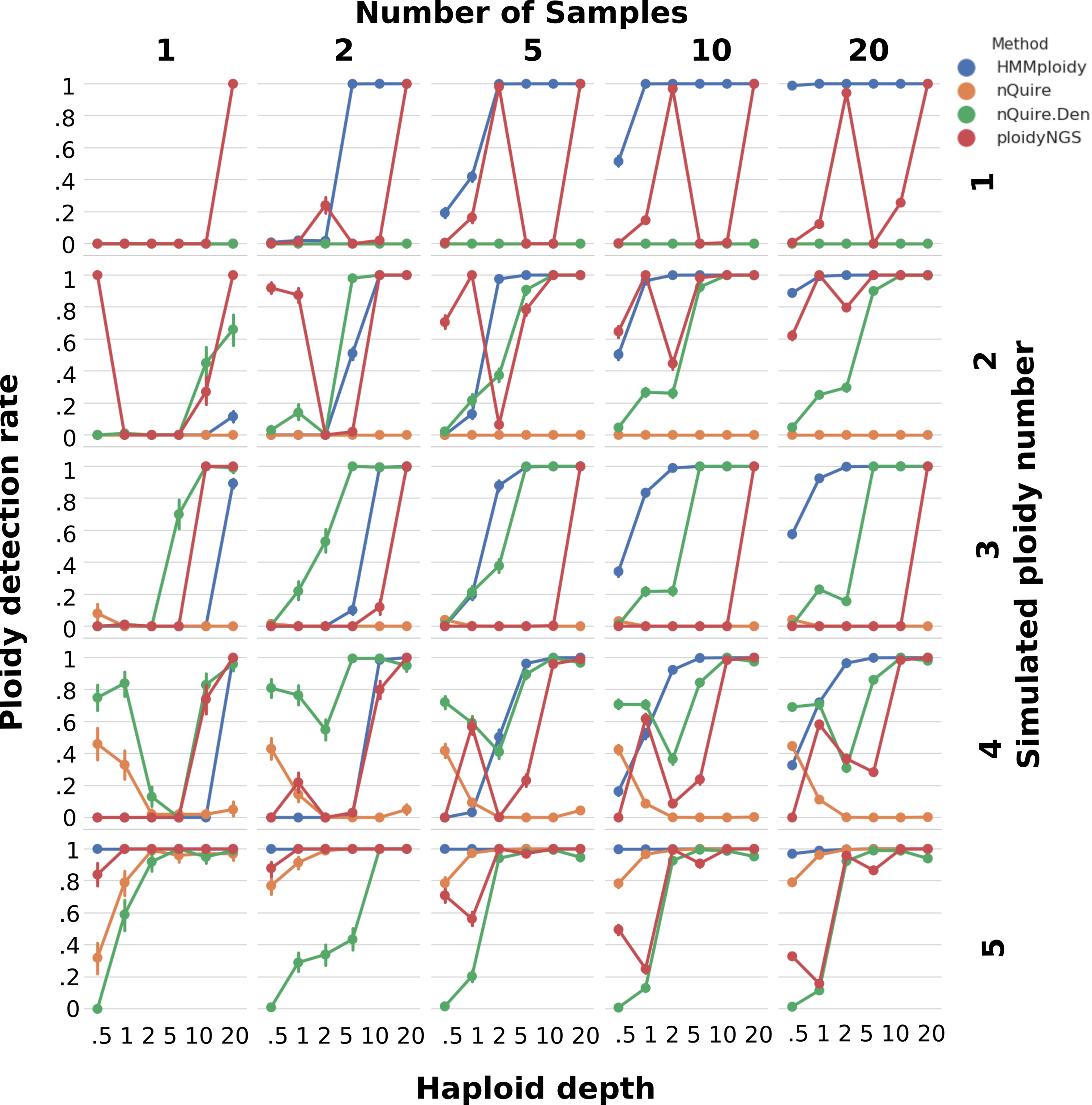
[48] Weiß, C.L., Pais, M., Cano, L.M., Kamoun, S., Burbano, H.A.: nQuire: a statistical framework for ploidy estimation using next generation sequencing (2018). https://doi.org/10.1186/s12859-018-2128-z, https://doi.org/10.1186/s12859-018-2128-z

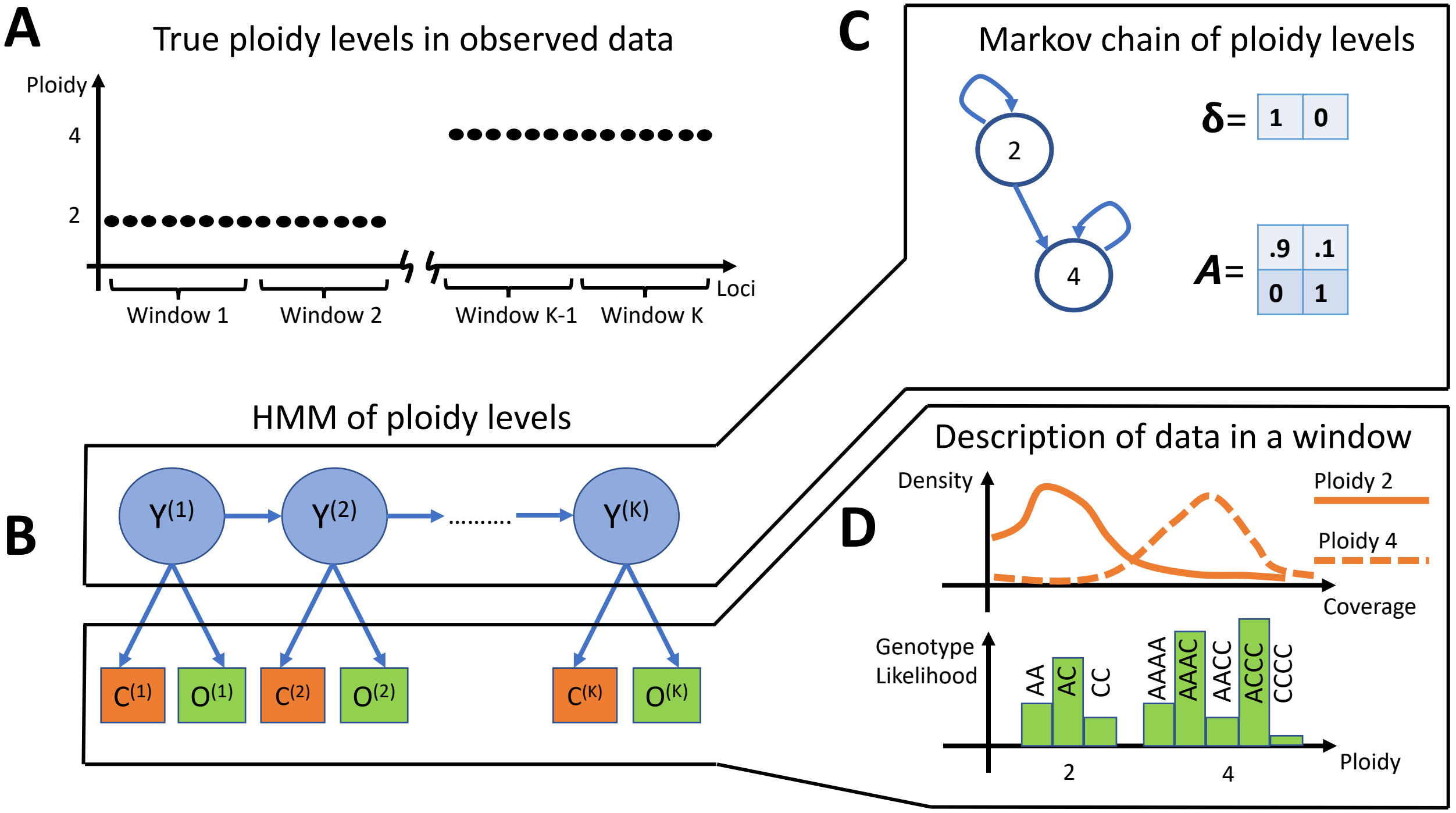
[49] Wood, T.E., Takebayashi, N., Barker, M.S., Mayrose, I., Greenspoon, P.B., H, R.L.: The frequency of polyploid speciation in vascular plants. Proc Natl Acad Sci USA (106), 13875–13879 (2009)

[50] 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

[51] 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


Depth and inferred ploidy from two samples
Chromosome 1
Chromosome 12
Ploidy
Depth
CCTP27
CCTP27-d121
CCTP50
CCTP50-d409
Position (Mb)


Number of Samples
1
2
5
10
20
Method
HMMploidy
nQuire
nQuire.Den
ploidyNGS
Ploidy detection rate
Simulated ploidy number
Haploid depth
0
.2
.4
.6
.8
1
.5 1 2 5 10 20


A
True ploidy levels in observed data
Ploidy
4
2
Loci
Window 1
Window 2
Window K-1
Window K
C
Markov chain of ploidy levels
2
4
δ=
1
0
A=
.9
.1
0
1
B
HMM of ploidy levels
Y(1)
Y(2)
.....
Y(K)
C(1)
O(1)
C(2)
O(2)
C(K)
O(K)
D
Description of data in a window
Density
Ploidy 2
Ploidy 4
Coverage
Genotype Likelihood
AA
AC
CC
AAAA
AAAC
AACC
ACCC
CCCC
2
4
Ploidy