Inference of Chromosomal Copy Number Variation from Short-Read Sequencing Data

Samuele Soraggi[*,1], **Oliver Tarrant**[†] **and Matteo Fumagalli**[†]

[*]Department of Mathematics, University of Copenhagen, Denmark, [†]Department of Life Sciences, Imperial College London, United Kingdom

ABSTRACT The inference of ploidy numbers from genetic data is an important yet challenging task for deciphering the evolutionary mechanisms underpinning genome evolution. High-throughput sequencing machines are now providing researchers with massive amount of genomic data. However, the data produced is typically affected by large sequencing errors and the assignment of individual genotypes is challenging when a low-depth strategy is employed.

Statistical methods that take genotype uncertainty into account have been introduced, allowing for an accurate estimation of nucleotide diversity even when little data is present. However, most of the available software and approaches are based on classic assumptions of random mating and diploidy. To solve this issue, we propose a novel statistical framework to estimate ploidy from sequencing data, taking into account base qualities and depth, through a Hidden Markov Model.

The method shows good performances in estimating trajectories of ploidy numbers even at low depth (2X) from simulated data. We also discuss how this method can be adopted to perform variant and genotype calling and estimation of summary statistics under an arbitrary number of ploidy directly from genotype likelihoods.

We finally demonstrate the utility of such method for estimating the chromosomal copy number variation in *Batrachochytrium dendrobatis* (Bd) from whole genome sequencing data. Bd is an amphibian fungus that is imposing a huge burden on its host. Genomes of Bd strains have been shown to be highly dynamic, with changes in ploidy observed even over short timescales. By analysing more than 200 samples from worldwide geographical locations, we aim to assess whether such rapid changes in chromosomal number copies are indeed associated to increased virulence. Unveiling how ploidy variation relates to fungal pathogenicity might hold the key for effective molecular monitoring of one of the most threatening epidemics for animal biodiversity.

KEYWORDS Ploidy; Genotipe Likelihoods; Poliploidy; Next Generation Sequencing; Genomics

Introduction

Ploidy number (or ploidy) is the number of sets of chromosomes in a cell. Humans are known to be diploid, but other species are often characterized by a different ploidy. When the ploidy of an organism is higher than two, it is usually referred to as poliploidy. The polyploidy state is often the consequence of hybridization or whole genome duplications, as often observed in plants. For instance, the genus of the perennial *Spartina* is characterized by triploid, hexaploid and dodecaploid species (Ainouche *et al.* 2003).

The changes in ploidy are considered to be playing an essential role in evolution of plants in natural populations (Adams and Wendel 2005) and is probably the most important factor concurring in speciation of plants (Otto and Whitton 2000). Moreover poliploidy can be an advantage for adapting to environmental factors when it causes alterations of the morphology and phenology of the organisms (Soltis and Soltis 2012). Those alterations can happen even as fast as one generation. For instance, poliploidy events have been detected in the ancestry of some types of crops and tomatoes (Schlueter *et al.* 2004), in lineages of the maize (Messing *et al.* 2004; Lai *et al.* 2004), in the common ancestry of cotton types (Rong *et al.* 2004; Blanc and Wolfe 2004) and

doi: 10.1534/genetics.XXX.XXXXXX
Manuscript compiled: Friday 30th March, 2018%

[1] Department of Mathematics, University of Copenhagen, Universitetsparken 5, 2100 Copenhagen, Denmark, samuele@math.ku.dk

soybeans (Schlueter *et al.* 2004), and in fungi (Todd *et al.* 2017; Wertheimer *et al.* 2016).

An experimental method to detect ploidy numbers in a genome is by using flow cytometry procedures (Kron *et al.* 2007). Flow cytometry is a high-throughput technique to obtain a quantification of optical properties, such as fluorescence, from particles floating in a special fluid. When flow cytometry is applied to a cell, it is possible to accurately determine the amount of genetic material in the nucleus, and estimate the ploidy number. Modern flow cytometry instruments are very sensible and reliable. However their cost is high (bennett and Leitch 2005; Greilhuber *et al.* 2007) and not justifiable when we are solely interested in the detection of ploidy numbers.

The advances in high-throughput sequencing techniques of the recent years, such as Next Generation Sequencing (NGS) (Goodwin *et al.* 2016; Reuter *et al.* 2015), have rapidly resulted in a vast amount of cost-effective high-throughput data available for a wide range of genetic studies. The available NGS protocols (Goodwin *et al.* 2016; Reuter *et al.* 2015; Metzker 2010) essentially result into an output that consists of short reads whose length is in the order of hundreds of bases, that are further aligned to a reference genome or *de novo* assembled in scaffolds. Many studies based on NGS data rely on low-depth sequencing ($< 10X$) because of cost-efficiency and/or degradation of the samples. Additionally NGS data is affected by a higher sequencing error than the one typical of Sanger sequencing (Ratan *et al.* 2013; Lam *et al.* 2012). These conditions may result in potentially unreliable estimates of allele frequencies in the data, and consequently a poor frequency-based estimation of genotypes.

Many of the current methods for the estimation of ploidy numbers in NGS data are based on analysis of sequencing depth and allele frequencies. For instance, `conPade` (Margarido and Heckerman 2015) detects the ploidy of a given contig/scaffold using allele frequencies. The tool `ploidyNGS` (Augusto Corrêa dos Santos *et al.* 2017) estimates allele frequencies and provides a visualization tool through which ploidy can be assigned. The visual approach is very commonly used to empirically estimate the ploidy (Yoshida *et al.* 2013). `AbsCN-seq` (Bao *et al.* 2014) combines the information on depth and allele counts to estimate, amongst other parameters related to tumor-specific applications, the ploidy from NGS data. Analogous data is applied to cancer cells' data with a different approach in the package `sequenza` (Favero *et al.* 2015).

We propose a method, called `hiddenMarkovPloidy`, dedicated to infer ploidy numbers from NGS data. In our method we build a Hidden Markov Model (HMM) (Cappe *et al.* 2005; Rabiner 1989) with a double set of observations, that consists of sequencing depths and observed reads. The formers are used to detect changes in ploidy, while the latters are based on the genotype likelihoods (Nielsen *et al.* 2011), and contribute in assigning each hidden state to its corresponding ploidy number. Notably this method is able to output the optimal number of ploidy numbers given an arbitrary initial interval of ploidies.

Simulations at haploid depth 2X show good performances in estimating ploidy numbers as high as five. We believe that this implementation can be also applied to the detection of Copy Number variants (CNV). Tools such as `CNVnator` Abyzov *et al.* (2011), `HadoopCNV` Yang *et al.* (2017) and `CNVfinder` Mccallum and Wang (2013) detect CNVs using sequencing depth and eventually allele frequencies. Here, we aim at using sequencing depth to detect changes in ploidy, and guess the levels based only on depths. Further, we can use genotype likelihoods to compare the guess on ploidy numbers to the ones estimated from genotype likelihoods, and flag the loci where those two estimates are different.

Emerging infectious diseases caused by fungi are a serious threat to global biodiversity and food security. The chytrid fungus *Batrachochytrium dendrobatidis* (Bd) is responsible for the dramatic decline of amphibians worldwide, causing one of the largest losses of biodiversity in recent times Fisher *et al.* (2012). Despite much interest, the genetic mechanisms that underpin Bd's virulence are not yet known but appear to be driven by a highly dynamic genomic landscape with frequent events of gain/loss of chromosomal copies. The geographic origins and the timing of Bd's spread are yet to be fully unravelled, making this one of the most controversial problems in disease ecology (Fisher 2017). Understanding the genetic mechanisms underlying Bd's virulence through an accurate mapping of ploidy numbers at different lineages is a fundamental goal to plan molecular monitoring.

Materials and Methods

This section describes the statistical framework in which the data is modelled and the Hidden Markov Model is built. In what follows data is assumed to be diallelic, without loss of generality.

Consider N sequenced individuals with M sequenced bases. Only the loci that are covered by at least one of the genomes are considered. For $i \in 1, \ldots, M$, and $j \in 1, \ldots, N$, let $Y_{j,i}$ be the ploidy number and $G_{j,i}$ be the genotype of individuals j at locus i. Denote with S_y the set of possible genotypes with ploidy $Y_{j,i} = y$, expressed as

$$S_{j,i}^{y} = \{0, 1, ..., y\},$$

where $\{0, 1, ..., y\}$ is the number of alternate (or derived) alleles per genotype.

Probability of Sequenced Data

Denote by O the sequenced data, and consider it independent between loci and individuals. Let $R_{j,i}$ the number of sequenced reads at locus i for individual j and $O_{j,i,r}$ be the r-th sequenced read for individual j at locus i, for $j = 1, \ldots, N$, $i = 1, \ldots, M$ and $r = 1, \ldots, R_{j,i}$. Denote with $O_{j,i,*}$ all the sequenced reads of individual j at locus i. The probability of $O_{j,i,*}$ conditionally on the ploidy number $Y_{j,i} = y_{j,i}$, the alternate allele frequency x_i at locus i and the inbreeding coefficient I_j of individual j is expressed by

$$p(O_{j,i,*}|y_{j,i}, x_i, I_j) = \sum_{g_{j,i} \in S_{j,i}^{y}} p(O_{j,i,*}|g_{j,i}, y_{j,i}) p(g_{j,i}|y_{j,i}, x_i, I_j), \quad (1)$$

where the left-hand side of the equation has been marginalized 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 (Nielsen *et al.* 2011); the second factor is the probability of the genotype given the frequency, the ploidy and the inbreeding coefficient. Throughout the analysis carried out in this paper, we assume absence of inbreeding and model such a probability with a binomial distribution.

Genotype Likelihood for Arbitrary ploidy number

The genotype likelihood is the probability of observing genotype $g_{j,i}$ for individual j at locus i, for $j = 1, \dots, N$, and $i = 1, \dots, M$, given the observed data. In its simplest formulation the genotype likelihood is determined considering the individual's base qualities as probabilities of incorrect sequenced bases, and assuming independence of the bases through the reads.

Let $R_{j,i}$ be the number of sequencing reads at a locus i for individual j, $O_{j,i,*}$ the individuals's observed data at that locus, $o_{j,i,r}$ and $q_{j,i,r}$ the observed nucleotide and the Phred base quality for the individual's read r at locus i, respectively. The i-th base of genotype g is denoted by g_i, $i \in 1, \dots, y$. The genotype likelihood of $g_{j,i}$ for ploidy number $y_{j,i}$ is expressed as

$$\ln p(O_{j,i,*}|g_{j,i}, y_{j,i}) = \sum_{r=1}^{R} \ln \Big(\sum_{i=1}^{y_{j,i}} \frac{1}{y_{j,i}} p(o_{j,i,r}|g_{j,i}, q_{j,i,r}, y_{j,i}) \Big)$$

where

$$p(o_{j,i,r}|g_{j,i}, q_{j,i,r}, y_{j,i}) = \begin{cases} 1 - \epsilon_{j,i,r}, & \textit{if } o_{j,i,r} = g_{j,i} \\ \frac{\epsilon_{j,i,r}}{3} & \textit{otherwise} \end{cases}$$

and $\epsilon_{j,i,r}$ is the Phred probability related to the score $q_{j,i,r}$. The probabilities of observing incorrect nucleotides are considered homogeneous through the possible nucleotides.

Consider $L_1, \dots, L_W$ a set of $W > 0$ non-overlapping windows of adjacent loci. We write $i \in L_w$, with abuse of notation, when locus i is in the w-th window, for $i = 1, \dots, M$ and $w = 1, \dots, W$. In each window only loci that are covered by at least one individual are considered. Under the hypothesis that loci are independent and the samples have the same ploidy number in each window, define

$$p_{j,L_w} = \prod_{i \in L_w} \prod_{j=1}^{N} p(O_{j,i,*}|y_{j,i}, x_i, I_j) \tag{2}$$

as the probability of the sequenced data in the w-th window for the j-th samples.

Estimation of population frequencies

If multiple samples are available, the population frequency x_i at each locus $i = 1, \dots, M$ is estimated assuming infinite ploidy. Consider, for each individual $j = 1, \dots, N$, the estimator $\hat{x}_{j,i}$ given by the relative frequency of the A allele. In each individual, the sequenced reads are a sample with replacement from the true genotype.

By assuming infinite ploidy, and therefore an infinitely long genotype for each individual, each sample can be considered as drawn from a different position of the genotype. Hence the reads are considered independent, and the amount of information contained in the estimator $\hat{x}_{j,i}$ is proportional to the number of reads at locus i for individual j. We define the population frequency estimator for x_i, say $\hat{x}_i$, as the weighted sum

$$\hat{x}_i = \sum_{j=1}^{N} \frac{R_{j,i}}{R_{*,i}} \hat{x}_{j,i},$$

where $R_{j,i}$ is the number of reads at locus i for individual j, and $R_{*,i} = \sum_{j=1}^{N} R_{j,i}$.

In case the sample size is limited, or even one single sample is analysed, $\hat{x}_i$ is not a valuable estimator of the population size and therefore (1) might be biased. In fact, in the case of a single sample the derived allele frequency provides the genotype, and therefore does not contain additional information. In this case it is thus assumed that the frequency is the same at each locus, in order to approximate the expected population allele frequency over all loci. Under this scenario, we further assume that one of the two alleles can be assigned to an ancestral (e.g. wild-type) state, while the other to a derived (e.g. mutant) state.

Under the standard coalescent model with infinite sites mutations (Tavaré 2004; Ewens 2004), the probability mass function of the population derived allele frequencies x in a sample of N individuals is (Kingman 1982):

$$f_X(x) = \frac{1/x^k}{\sum_{j=1}^{-1} \frac{1}{j^k}}, \tag{3}$$

with X the random variable describing the allele frequency and $k \in (0, \infty)$ being a positive real number, that determines whether the population is deviating from a model of constant population size. For instance, $k = 1$ is equal to the distribution of x under constant population size, while $k > 1$ models a population shrinkage and $k < 1$ population growth. Given the probability distribution (3), the expected derived allele frequency in a population of size n is:

$$E(X) = \sum_{j=1}^{n-1} \frac{x^{-k}}{\sum_{j=1}^{n-1} \frac{1}{w^k}}. \tag{4}$$

Using the expected value of the frequency it is then possible to calculate quantities that involve the allele frequencies when only few samples are available.

Unknown or Uncertain Ancestral Allelic State. One of our main assumptions for the single-sample case is that we know which allele can be assigned to an ancestral state, and which one to a derived state. However, in many practical cases, such assignment is either not possible or associated with a certain level of uncertainty due to, for instance, ancestral polymorphisms or genome from a closely related species not being available. Under these circumstances, we extend our formulation by adding a parameter underlying the probability that the assigned ancestral state is incorrectly identified.

Let us define v as the ancestral state. This can take value in V, the set of the two most likely alleles from $\{A, C, G, T\}$. Assume that the true ancestral state is contained in $\mathcal{V}$.

The log-probability of the data for a single sample is

$$\ln p(O|y) = \sum_{i=1}^{N} \ln \Big(\sum_{v \in V} \sum_{g \in S_y} p(O|g_i, y_i) p(g_i|y_i, v) p(v) \Big) \tag{5}$$

where $p(v)$ denotes the probability that allele v is the ancestral state and is invariant across sites. Note that $\sum_{v \in V} p(v) = 1$. If $p(v) = 0.5$ for each $v \in V$, then (5) refers to the scenario of folded allele frequencies, where each allele is equally probable to be the ancestral state.

Hidden Markov Model for Ploidy Inference

Under the assumption that in each window of loci the ploidy is constant, we infer the ploidy numbers using a hidden markov model (HMM) with double emissions. Let an HMM for ploidy inference be defined by a discrete process

$$\{Y_{j,L_w}, C_{j,L_w}, O_{j,L_w,*}\}_{w=1}^{W},$$

where W is the number of windows of adjacent loci considered. The unobservable chain Y_{j,L_w} represents the unknown ploidy numbers, C_{j,L_w} the observed depth and $O_{j,L_w,*}$ the observed sequenced data for the j-th individual in the w-th window. The transition probabilities of the unknown markov chain are denoted by $\boldsymbol{A} = \{a_{ij}\}_{i,j=1}^{T}$, and the stationary probability of the chain by the vector $\boldsymbol{\pi}$ of length T, where T is the number of ploidy numbers considered in the model.

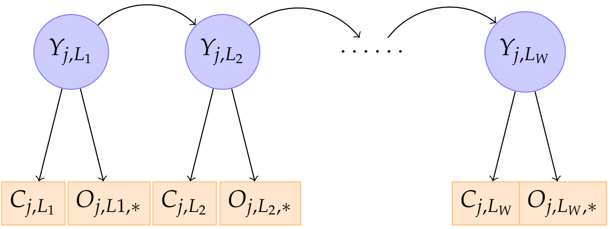

Figure 1 Hidden markov model for the detection of the unknown ploidy numbers Y_{j,L_w} of an individual j in adjacent windows of loci L_w, for $j = 1, \dots, N$ and $w = 1, \dots, W$. The ploidy-dependent emissions consist of the average coverage C_{j,L_w} and the sequenced data $O_{j,L_w,*}$.

Using the HMM defined above implies that some probabilistic relationships are assumed, amongst which:

- conditionally on the sequence of ploidy numbers, the average depth and the data in a window both depend on the ploidy at that window,
- the average depth and the data in a window are conditionally independent, given the ploidy number.

At each window, the average coverage given the ploidy number is modelled by a negative binomial distribution to capture the behaviour of overdispersed values. The observed data given the ploidy number at a certain window is described by the probability in equation (2).

The estimation of the parameters $\boldsymbol{A}, \boldsymbol{\pi}, \boldsymbol{\theta}$, where $\boldsymbol{\theta}$ characterizes the ploidy-dependent distributions of the depth, is performed through the EM-algorithm (Cappe *et al.* 2005; Rabiner 1989). The EM-algorithm is modified using the AIC criterion (Bishop 2006) to find the optimal number of ploidy numbers by following an approach similar to the one (Li and Biswas 1999). Here the EM algorithm is started with the maximum number of hidden states T of the Markov chain. When the convergence criteria of the EM procedure is met, one of the states is removed and the EM algorithm restarted. If the AIC criteria suggests that removing a state is not necessary, then the optimal number of states is found.

The genotype likelihoods solve the problem of the identifiability of the states (given T hidden states of the chain, there are $T!$ relabeled HMMs that provides the same result with the EM algorithm) (Rabiner 1989; Bishop 2006). The optimal sequence of ploidy numbers is inferred using the Viterbi algorithm, that detects the most probable sequence of ploidies once the parameters of the model have been optimized (Rabiner 1989; Bishop 2006; Viterbi and A. 1967; Forney 1973).

Simulations

To assess the accuracy of estimating ploidy from sequencing data, mapped reads in *mpileup* format are simulated for different scenarios of haploid depth and changes in ploidy numbers. Each site i, for $i = 1, \dots, M$, is treated as an independent observation, without modelling the effect of linkage disequilibrium. The number of reads is distributed as a Poisson(cy_i), where c is the haploid depth and y_i the ploidy at locus i.

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. nucleotides' bases) are sampled with replacement with a certain probability. Such a probability is given by the quality scores.

All simulated configurations involve 20 individuals, known ancestral allele and absence of inbreeding. In the simulated scenario, 10^4 loci are simulated in two situations, with haploid depth $0.5X$ and $2X$. Here the ploidy changes every 1000 loci increasing from 1 to 5, and decreasing from 5 to 1.

Real Data

We applied our method to detect ploidy numbers to whole-genome sequencing data of Bd strains (Farrer *et al.* 2013). The assembled genome is 20Mbp long comprising more than 20 supercontigs. We first investigated changes in ploidy for a sample previously discovered to be highly variables in chromosomal copies. We will then aim at analysing more than 200 samples of Bd for different geographical locations, comprising the suggestive source of the panzootic (South Africa, North America, South America, Japan and East Asia).

Results and Discussion

In both simulations and real data scenarios, non overlapping windows with size of ten loci are used. In those, only the loci where the allele frequencies estimated with ANGSD fall in the interval $[0.1, 0.9]$ are selected.

In the simulated scenario of Figure 2, the Hidden Markov Model is able to recognize the simulated ploidy numbers from 1 to 4 with few errors at depth $0.5X$. However it does not identify ploidy number 5. This is likely due to two causes. The first is a poor estimation of allele frequencies from low-depth samples, causing the probability of observed data to be maximum for a lower ploidy number. Indeed, the higher is the ploidy number, the easier is that the bias on allele frequencies confound the selection of the correct ploidy value. The second cause is the lack of reads, and therefore the difficulty in inferring some of the genotypes. In fact, in some loci the number of reads available from the 20 individuals might be lower than five. However, the case of depth $0.5X$ is extreme and real data is in general at higher depth. Only minor issues are observed in case of haploid depth $2X$, where ploidy is estimated correctly except in few windows for levels 4 and 5.

Figure 3 shows the performances on 15 contigs of one strain of Bd. Each window of loci has a size of 50Kb. In the graph representing the depth, red lines represent the mean of the depth distribution. The bottom plot shows the minor allele frequencies estimated with `ANGSD` as an additional sanity check. Here the inferred ploidies are compatible with the ones that can be deduced by visual analyse at the sequencing depth variation. Minor errors observable are caused by oscillations in the sequencing depth.

Note that the frequency estimation needs a high number available samples, especially at low depth. In case of few samples, or even only one, are available, the use of the expect frequency over all loci calculated in (4) is an alternative to estimate the frequencies used in the Hidden Markov Model framework.

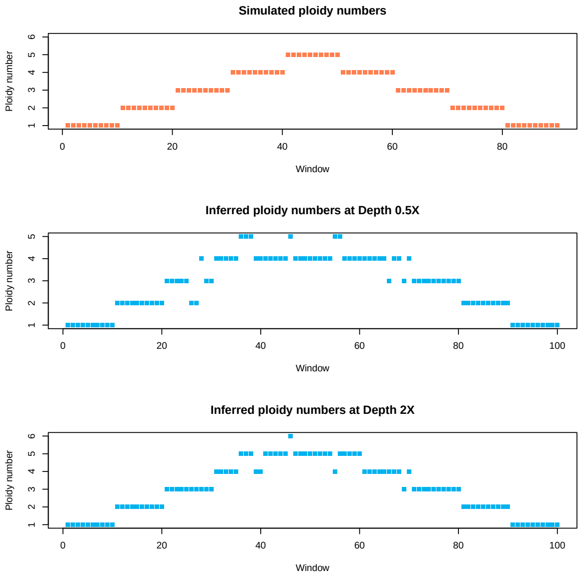

Figure 2 Ploidy inference from simulated data. Inference of simulated ploidy numbers (red), where the ploidy changes from 1 to 5 and is constant in each window of loci. In all plots the window size is 10 loci. The results are shown in blue dots for depth 0.5X and 2X.

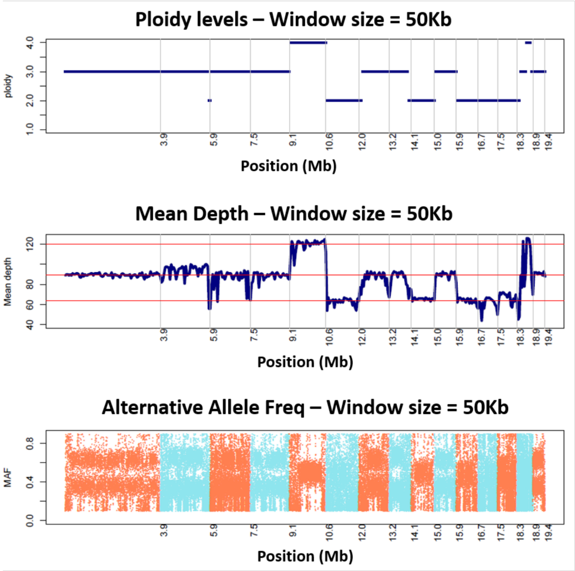

Figure 3 Ploidy inference from a strain of the Bd fungi. Inference of ploidy numbers from a strain of the Bd fungi. For each window of loci of size 50Kb, the plot shows the inferred ploidies, the average sequencing depth and the estimated minor allele frequencies.

Literature Cited

Abyzov, A., A. E. Urban, M. Snyder, and M. Gerstein, 2011 CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Research **21**: 974–984.

Adams, K. L. and J. F. Wendel, 2005 Polyploidy and genome evolution in plants. Current Opinion in Plant Biology **8**: 135–141.

Ainouche, M. L., A. Baumel, A. Salmon, and G. Yannic, 2003 Hybridization, polyploidy and speciation in Spartina (Poaceae). New Phytologist **161**: 165–172.

Augusto Corrêa dos Santos, R., G. H. Goldman, and D. M. Riaño-Pachón, 2017 ploidyNGS: visually exploring ploidy with Next Generation Sequencing data. Bioinformatics **33**: 2575–2576.

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.

bennett, M. D. and I. J. Leitch, 2005 Nuclear DNA Amounts in Angiosperms: Progress, Problems and Prospects. Annals of Botany **95**: 45–90.

Bishop, C. M., 2006 *Pattern recognition and machine learning*. Springer.

Blanc, G. and K. H. Wolfe, 2004 Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. The Plant cell **16**: 1667–78.

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

Ewens, W. J., 2004 *Mathematical population genetics : 1. Theoretical introduction*. Springer.

Farrer, R. A., D. A. Henk, T. W. J. Garner, F. Balloux, D. C. Woodhams, and M. C. Fisher, 2013 Chromosomal Copy Number Variation, Selection and Uneven Rates of Recombination Reveal Cryptic Genome Diversity Linked to Pathogenicity. PLoS Genetics **9**: e1003703.

Favero, F., T. Joshi, A. M. Marquard, N. J. Birkbak, M. Krzystanek, Q. Li, Z. Szallasi, and A. C. Eklund, 2015 Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. Annals of Oncology **26**: 64–70.

Fisher, M. C., 2017 Ecology: In peril from a perfect pathogen. Nature **544**: 300–301.

Fisher, M. C., D. A. Henk, C. J. Briggs, J. S. Brownstein, L. C. Madoff, S. L. McCraw, and S. J. Gurr, 2012 Emerging fungal threats to animal, plant and ecosystem health. Nature **484**: 186–194.

Forney, G., 1973 The viterbi algorithm. Proceedings of the IEEE **61**: 268–278.

Goodwin, S., J. D. McPherson, and W. Richard McCombie, 2016 Coming of age: ten years of next-generation sequencing technologies .

Greilhuber, J., E. M. Temsch, and J. C. M. Loureiro, 2007 Nuclear DNA Content Measurement. In *Flow Cytometry with Plant Cells*, pp. 67–101, Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany.

Kingman, J., 1982 The coalescent. Stochastic Processes and their Applications **13**: 235–248.

Kron, P., J. Suda, and B. C. Husband, 2007 Applications of Flow Cytometry to Evolutionary and Population Biology. Annu. Rev. Ecol. Evol. Syst **38**: 847–76.

Lai, J., J. Ma, Z. Swigonová, W. Ramakrishna, E. Linton, V. Llaca, B. Tanyolac, Y.-J. Park, O.-Y. Jeong, J. L. Bennetzen, and J. Messing, 2004 Gene Loss and Movement in the Maize Genome.

Genome Research **14**: 1924–1931.

Lam, H. Y. K., M. J. Clark, R. Chen, R. Chen, G. Natsoulis, M. O'Huallachain, F. E. Dewey, L. Habegger, E. A. Ashley, M. B. Gerstein, A. J. Butte, H. P. Ji, and M. Snyder, 2012 Performance comparison of whole-genome sequencing platforms. Nature biotechnology **30**: 78.

Li, C. and G. Biswas, 1999 Temporal Pattern Generation Using Hidden Markov Model Based Unsupervised Classification. In *IDA 1999: Advances in Intelligent Data Analysis*, pp. 245–256, Springer, Berlin, Heidelberg.

Margarido, G. R. A. and D. Heckerman, 2015 ConPADE: Genome Assembly Ploidy Estimation from Next-Generation Sequencing Data. PLOS Computational Biology **11**: e1004229.

Mccallum, K. J. and J.-P. Wang, 2013 Quantifying copy number variations using a hidden Markov model with inhomogeneous emission distributions. Biostatistics **14**: 600–611.

Messing, J., A. K. Bharti, W. M. Karlowski, H. Gundlach, H. R. Kim, Y. Yu, F. Wei, G. Fuks, C. A. Soderlund, K. F. X. Mayer, and R. A. Wing, 2004 Sequence composition and genome organization of maize. Proceedings of the National Academy of Sciences of the United States of America **101**: 14349–54.

Metzker, M. L., 2010 Sequencing technologies — the next generation. Nature Reviews Genetics **11**: 31–46.

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.

Otto, S. P. and J. Whitton, 2000 Polyploid Incidence and Evolution. Annual Review of Genetics **34**: 401–437.

Rabiner, L., 1989 A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE **77**: 257–286.

Ratan, A., W. Miller, J. Guillory, J. Stinson, S. Seshagiri, and S. C. Schuster, 2013 Comparison of sequencing platforms for single nucleotide variant calls in a human sample. PloS one **8**: e55089.

Reuter, J., D. V. Spacek, and M. Snyder, 2015 High-Throughput Sequencing Technologies. Molecular Cell **58**: 586–597.

Rong, J., C. Abbey, J. E. Bowers, C. L. Brubaker, C. Chang, P. W. Chee, T. A. Delmonte, X. Ding, J. J. Garza, B. S. Marler, C.-h. Park, G. J. Pierce, K. M. Rainey, V. K. Rastogi, S. R. Schulze, N. L. Trolinder, J. F. Wendel, T. A. Wilkins, T. D. Williams-Coplin, R. A. Wing, R. J. Wright, X. Zhao, L. Zhu, and A. H. Paterson, 2004 A 3347-locus genetic recombination map of sequence-tagged sites reveals features of genome organization, transmission and evolution of cotton (Gossypium). Genetics **166**: 389–417.

Schlueter, J. A., P. Dixon, C. Granger, D. Grant, L. Clark, J. J. Doyle, and R. C. Shoemaker, 2004 Mining EST databases to resolve evolutionary events in major crop species. Genome **47**: 868–876.

Soltis, P. S. and D. E. Soltis, 2012 *Polyploidy and genome evolution*. Springer.

Tavaré, S., 2004 *Ancestral Inference in Population Genetics*. Springer, Berlin, Heidelberg.

Todd, R. T., A. Forche, and A. Selmecki, 2017 Ploidy Variation in Fungi: Polyploidy, Aneuploidy, and Genome Evolution. Microbiology spectrum **5**.

Viterbi, A. and A., 1967 Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Transactions on Information Theory **13**: 260–269.

Wertheimer, N. B., N. Stone, and J. Berman, 2016 Ploidy dynamics and evolvability in fungi. Philosophical transactions of the Royal Society of London. Series B, Biological sciences **371**.

Yang, H., G. Chen, L. Lima, H. Fang, L. Jimenez, M. Li, G. J. Lyon, M. He, and K. Wang, 2017 HadoopCNV: A Dynamic Programming Imputation Algorithm To Detect Copy Number Variants From Sequencing Data. bioRxiv p. 124339.

Yoshida, K., V. J. Schuenemann, L. M. Cano, M. Pais, B. Mishra, R. Sharma, C. Lanz, F. N. Martin, S. Kamoun, J. Krause, M. Thines, D. Weigel, and H. A. Burbano, 2013 The rise and fall of the Phytophthora infestans lineage that triggered the Irish potato famine. eLife **2**: e00731.