

METHODOLOGY ARTICLE

Open Access



nQuire: a statistical framework for ploidy estimation using next generation sequencing

Clemens L. Weiß¹, Marina Pais², Liliana M. Cano^{2,3}, Sophien Kamoun² and Hernán A. Burbano^{1*}

Abstract

Background: Intraspecific variation in ploidy occurs in a wide range of species including pathogenic and nonpathogenic eukaryotes such as yeasts and oomycetes. Ploidy can be inferred indirectly - without measuring DNA content - from experiments using next-generation sequencing (NGS). We present nQuire, a statistical framework that distinguishes between diploids, triploids and tetraploids using NGS. The command-line tool models the distribution of base frequencies at variable sites using a Gaussian Mixture Model, and uses maximum likelihood to select the most plausible ploidy model. nQuire handles large genomes at high coverage efficiently and uses standard input file formats.

Results: We demonstrate the utility of nQuire analyzing individual samples of the pathogenic oomycete *Phytophthora infestans* and the Baker's yeast *Saccharomyces cerevisiae*. Using these organisms we show the dependence between reliability of the ploidy assignment and sequencing depth. Additionally, we employ normalized maximized log-likelihoods generated by nQuire to ascertain ploidy level in a population of samples with ploidy heterogeneity. Using these normalized values we cluster samples in three dimensions using multivariate Gaussian mixtures. The cluster assignments retrieved from a *S. cerevisiae* population recovered the true ploidy level in over 96% of samples. Finally, we show that nQuire can be used regionally to identify chromosomal aneuploidies.

Conclusions: nQuire provides a statistical framework to study organisms with intraspecific variation in ploidy. nQuire is likely to be useful in epidemiological studies of pathogens, artificial selection experiments, and for historical or ancient samples where intact nuclei are not preserved. It is implemented as a stand-alone Linux command line tool in the C programming language and is available at <https://github.com/clwgg/nQuire> under the MIT license.

Keywords: Ploidy, Probabilistic modeling, Next generation sequencing

Background

Polyploidy, the presence of more than two complete sets of chromosomes, can under certain circumstances accelerate evolutionary adaptation by influencing the generation and maintenance of genetic diversity [1, 2]. In addition, polyploidy also poses short- and long-term challenges to organismal fitness, which are associated with increased nuclear and cellular volume, propensity to aneuploidy, and disruption of gene expression regulation [3]. Interspecific comparisons between eukaryotic genomes can identify ancient polyploidization events. In contrast, more recent polyploidization events result in intraspecific variation of

ploidy and, in some cases, aneuploidy. The presence of individuals of different ploidy in a population can hinder mating. Therefore, intraspecific ploidy variation tends to occur - although not exclusively - in organisms that have the capacity to reproduce asexually [4–6], are self-compatible or perennial [7].

Although ploidy traditionally has been investigated by measuring DNA content using flow cytometry [8], it can also be inferred from next generation sequencing (NGS) data either by examining k-mer distributions, or by assessing the distribution of allele frequencies at biallelic single nucleotide polymorphisms (SNPs) [4]. This methodology has been used to estimate ploidy in newly assembled genomes in order to identify the number of likely collapsed haplotypes on a per-contig basis [9], as well as to detect intraspecific variation of ploidy in the oomycete *Phytophthora infestans* [4, 6] and in the Baker's yeast *Saccharomyces cerevisiae* [5]. It also was successfully used for

*Correspondence: hernan.burbano@tuebingen.mpg.de

¹Research Group for Ancient Genomics and Evolution, Department of Molecular Biology, Max Planck Institute for Developmental Biology, Tuebingen, Germany

Full list of author information is available at the end of the article

ploidy estimation in *P. infestans* historic herbaria samples that are not suitable for flow cytometry, since they do not contain intact nuclei [4]. The method assumes that alleles present at biallelic SNPs occur at different ratios for different ploidy levels, that is, 0.5/0.5 in diploids, 0.33/0.67 in triploids, and a mixture of 0.25/0.75 and 0.5/0.5 in tetraploids. These ratios hold true for recent autopolyploids, as well as recent allopolyploids from heterozygous source genomes. To determine the ploidy level, the distribution of biallelic SNPs can be inspected visually [10], and/or qualitatively compared with simulated data [4]. However, this methodology does not provide summary statistics that permit quantifying how well the data fit the expected distributions, which is especially critical when dealing with noisy distributions typical for highly-repetitive genomes. An additional disadvantage of this approach is that it is preceded by the identification of variable sites (“SNP calling”), which is carried out using methodologies that benefit from a previously known ploidy level [11]. In a further development Gompert et al. [12] used biallelic SNPs in a Bayesian statistical approach to distinguish between ploidy levels from genotyping-by-sequencing data. This method was primarily developed for resequencing studies, where typically multiple individuals from populations with ploidy variation are genotyped, as it benefits from preexisting knowledge about the ploidy levels that may be observed. Based on the posterior probabilities emitted by the Bayesian model, this approach separates samples into ploidy clusters, using dimensionality reduction methods such as PCA. Since it allows the inclusion of training data of known ploidy, test samples can be assigned a ploidy level if they belong to a cluster that includes samples of known ploidy.

Here we present a statistical model that aims to distinguish between the distribution of base frequencies at variable sites for diploids, triploids and tetraploids, directly from read mappings to a reference genome. It models base frequencies as a Gaussian Mixture Model (GMM), and uses maximum likelihood to assess empirical data under the assumptions of diploidy, triploidy and tetraploidy. We evaluated the performance of our method for different sequencing coverages using published genomes of *S. cerevisiae* [5], and high-coverage genomes of *P. infestans* produced for this study.

Methods

Model and Implementation

We used base frequencies at variable sites with only two bases segregating to distinguish between diploids, triploids and tetraploids (Fig. 1a). For that, we implemented a GMM that models the base frequency profiles as a mixture of three Gaussian distributions (Fig. 1b), which are scaled relative to each other. A log-likelihood can be calculated following:

$$\log L = \sum_{i=1}^n \log \sum_{j=1}^3 \alpha_j N(x_i; \mu_j, \sigma_j)$$

Here, n describes the numbers of data points and x_i describes the value of each data point (i.e. the base frequencies). At sites with two bases segregating, we use the frequencies of both bases to achieve symmetry in the frequency distribution for triploids and tetraploids. μ_j and σ_j are the parameters of the j^{th} of three Gaussian distributions N_j that are scaled relative to each other through the parameter α_j , with $\sum_{j=1}^3 \alpha_j = 1$.

This model allows for estimating the parameters of the Gaussian mixture components, as well as their mixture proportions, by maximizing the log-likelihood, either with or without constraints on the possible parameter space.

The likelihood maximization of the GMM is implemented through an Expectation-Maximization (EM) algorithm (Fig. 1b).

During the EM, we make use of the latent variables Z_i . They represent the assignment of a data point to one of the mixture components. In the E-step, we use the current estimates for μ_j , σ_j and α_j to calculate:

$$P(Z_i = j | x_i) = \frac{\alpha_j N(x_i; \mu_j, \sigma_j)}{\sum_{j=1}^3 \alpha_j N(x_i; \mu_j, \sigma_j)} = \gamma_{Z_i}(j)$$

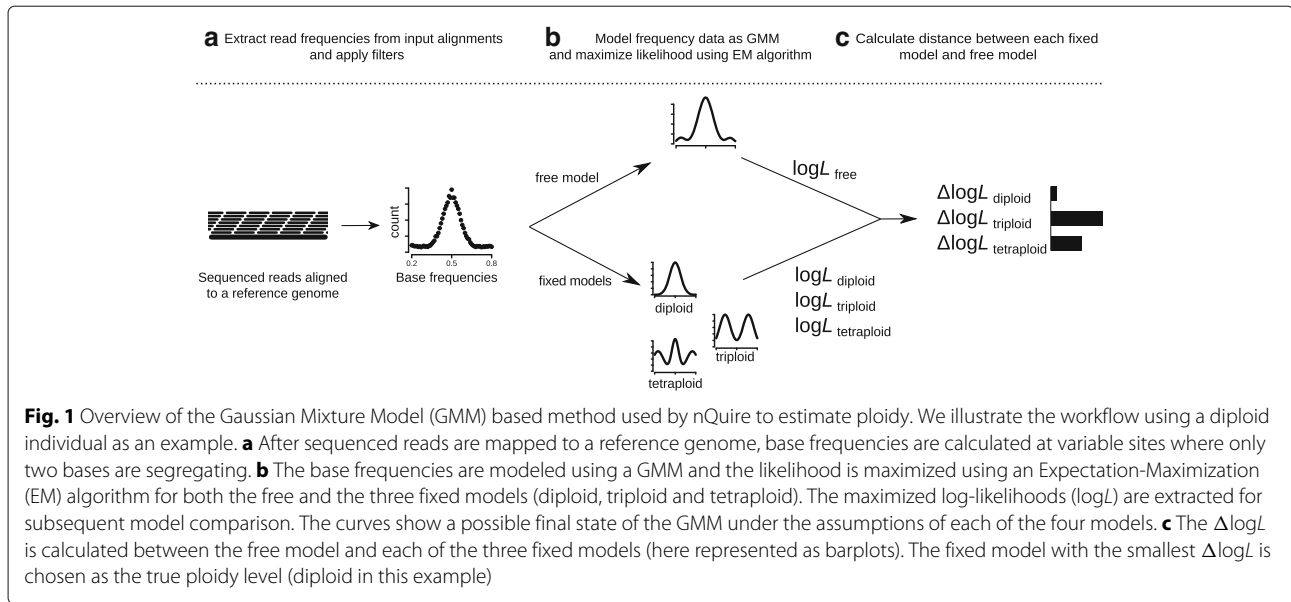
This probability is calculated for all x_i and $j \in \{1, 2, 3\}$ to form a $n \times 3$ matrix, the columns of which represent the probability of each data point belonging to either of the three mixture components. From this matrix we can calculate the column sum $S_j = \sum_{i=1}^n \gamma_{Z_i}(j)$, which represents the size of each mixture component. In the following M-step, we update our estimates of μ_j , σ_j and α_j :

$$\begin{aligned} \hat{\mu}_j &= \frac{1}{S_j} \sum_{i=1}^n \gamma_{Z_i}(j) x_i \\ \hat{\sigma}_j^2 &= \frac{1}{S_j} \sum_{i=1}^n \gamma_{Z_i}(j) (x_i - \mu_j)^2 \\ \hat{\alpha}_j &= \frac{S_j}{n} \end{aligned}$$

The log-likelihood is calculated after the M-step, and the next E-step is initiated unless the log-likelihood has changed by less than $\epsilon = 0.01$ from the previous M-step.

As shown above, the algorithm allows the estimation of μ_j , σ_j and α_j simultaneously. Henceforth, we call this setup the “free model”. The log-likelihood of the free model upon convergence represents the optimum under the assumptions of the model.

We can also set certain parameters to fixed values and forgo their update in the M-step. We use this to maximize the log-likelihood under the expectation of diploidy (one Gaussian with mean 0.5), triploidy (two Gaussians with means 0.33 and 0.67) and tetraploidy (three Gaussians



with means 0.25, 0.5 and 0.75), and call these the three “fixed models”:

$$\log L_{\text{diploid}} = \sum_{i=1}^n \log N(x_i; 0.5, \sigma)$$

$$\log L_{\text{triploid}} = \sum_{i=1}^n \log \sum_{j=1}^2 0.5 \cdot N(x_i; \mu_j, \sigma_j),$$

$$\mu_j \in \{0.33, 0.67\}$$

$$\log L_{\text{tetraploid}} = \sum_{i=1}^n \log \sum_{j=1}^3 0.33 \cdot N(x_i; \mu_j, \sigma_j),$$

$$\mu_j \in \{0.25, 0.5, 0.75\}$$

In these three models, we only estimate σ_j , while μ_j and α_j are fixed as shown above. Since all fixed models are nested within the free model, it is possible to directly compute the log-likelihood ratios, following:

$$\Delta \log L_{\text{diploid}} = \log L_{\text{free}} - \log L_{\text{diploid}}$$

$$\Delta \log L_{\text{triploid}} = \log L_{\text{free}} - \log L_{\text{triploid}}$$

$$\Delta \log L_{\text{tetraploid}} = \log L_{\text{free}} - \log L_{\text{tetraploid}}$$

The $\Delta \log L$ s describe the distance between each fixed model and the best fit under the assumptions of the GMM. A substantially lower $\Delta \log L$ of one fixed model over the others supports the ploidy level described by this fixed model (Fig. 1c). Therefore, we used $\Delta \log L$ as summary statistics where the minimum value supports a given ploidy level.

Additionally, the GMM can be extended to a Gaussian Mixture Model with Uniform noise component (GMMU), by adding a uniform mixture component:

$$\log L = \sum_{i=1}^n \log \left(\alpha_4 U(x_i) + \sum_{j=1}^3 \alpha_j N(x_i; \mu_j, \sigma_j) \right)$$

The constraint on the mixture proportions then becomes $\sum_{j=1}^4 \alpha_j = 1$.

The uniform noise component is used in our implementation to allow base-line noise removal (Additional file 1: Figure S1). This can be useful when the Gaussian peaks are observable but embedded in a basal noise, which could be caused by highly repetitive genomes or low coverage.

Multivariate Gaussian clustering

To cluster samples in three dimensions based on the normalized maximized log-likelihoods of the three fixed models, we used the mclust5 package [13] of the R programming language [14]. This package utilizes mixtures of multivariate Gaussian distributions to detect clusters in an arbitrary number of dimensions. mclust5 allows to set constraints on the volume, shape and orientation of each mixture component, by varying features of their covariance matrix either within each sample, or for all samples at once. For the analysis displayed in Fig. 4, we used clusters of equal volume, but varying shape and orientation. This configuration represented the data the best, as assessed by the recovery of ploidy levels from cluster assignments.

Phytophthora infestans libraries

The two benchmarking libraries from *P. infestans* were generated according to the protocol by Meyer and Kircher [15] from DNA extracted from lab cultures [16]. These libraries were sequenced to high coverage on an Illumina HiSeq 3000 machine in paired end 150 bp mode. This

Table 1 Samples of *Saccharomyces cerevisiae* and *Phytophthora infestans* used to evaluate and benchmark nQuire

Sample	Ploidy	Species	Cov. ^a	$\Delta\log L_{2n}$	$\Delta\log L_{3n}$	$\Delta\log L_{4n}$
CBS7837	2n	<i>S. cerev.</i>	116	6319	35721	23033
CBS2919	3n	<i>S. cerev.</i>	111	33614	920	29347
CBS9564	4n	<i>S. cerev.</i>	101	37003	22971	6429
99189	2n	<i>P. infes.</i>	210	119218	369682	293194
88069	3n	<i>P. infes.</i>	368	599933	42717	390002

^a Average per-base coverage. The smallest $\Delta\log L$ for each sample is highlighted in bold

sequencing data is available at the European Nucleotide Archive (ENA) under study number PRJEB20998.

Results

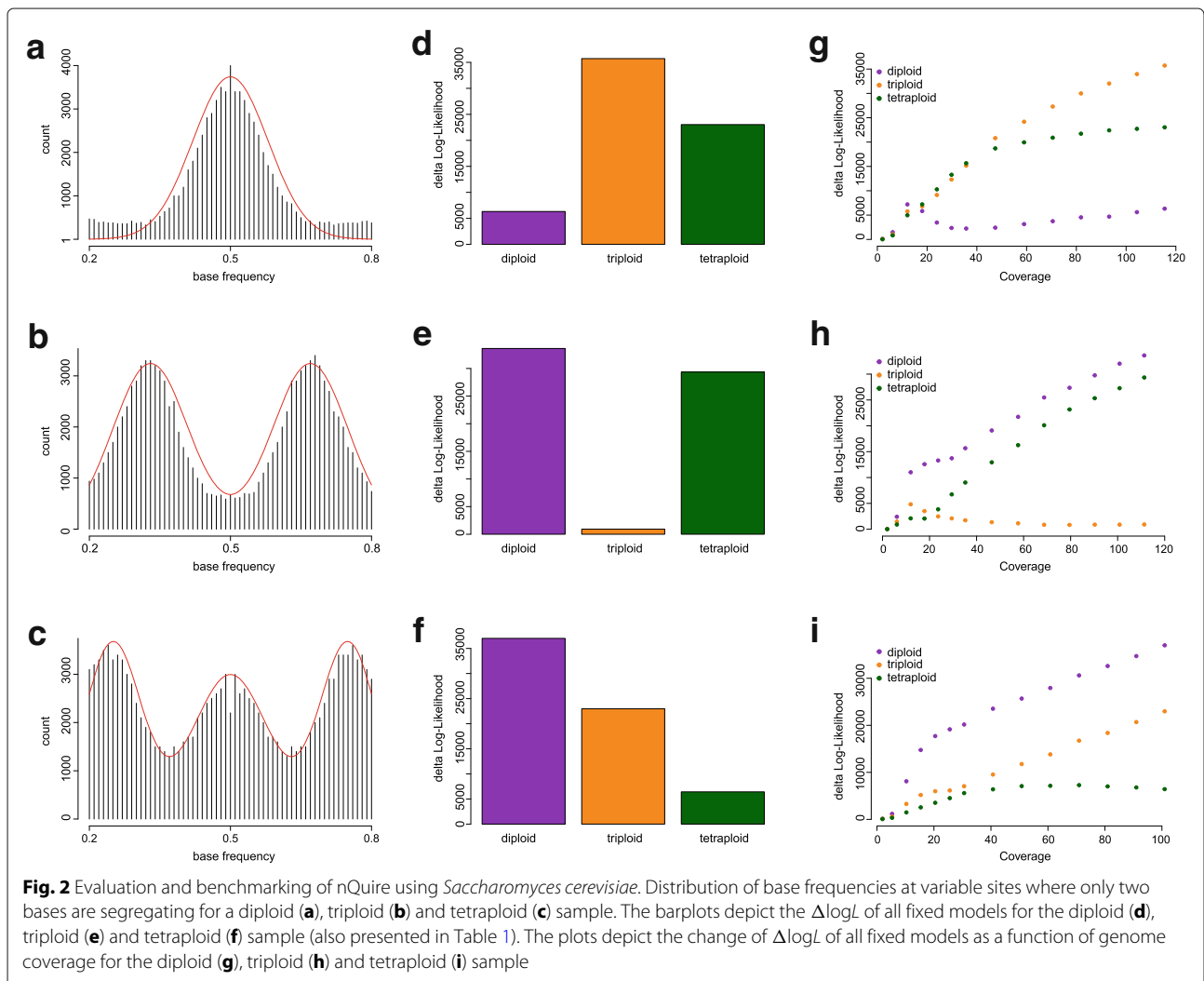
Performance

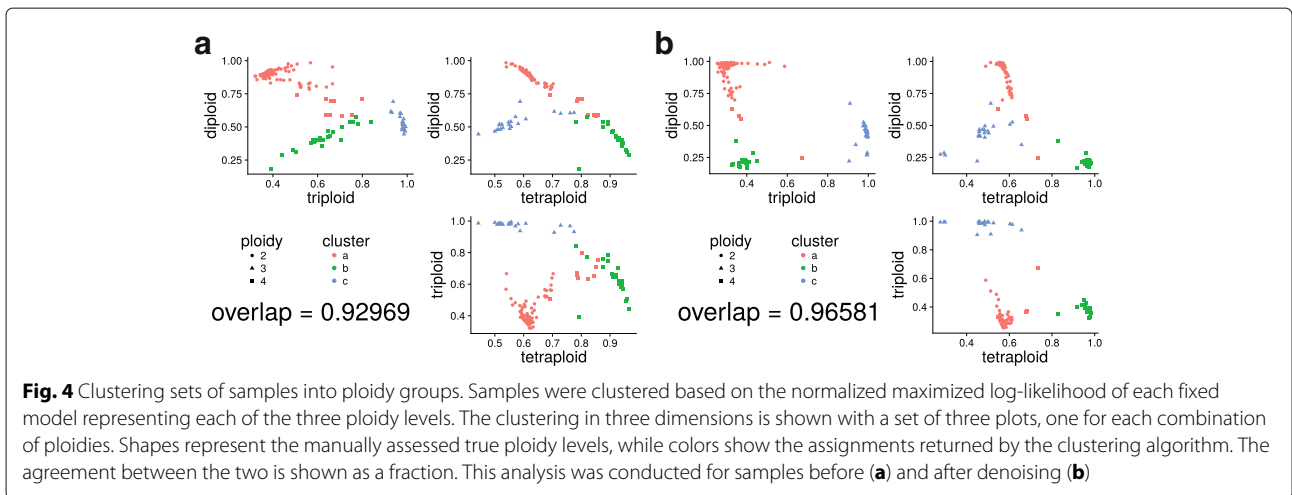
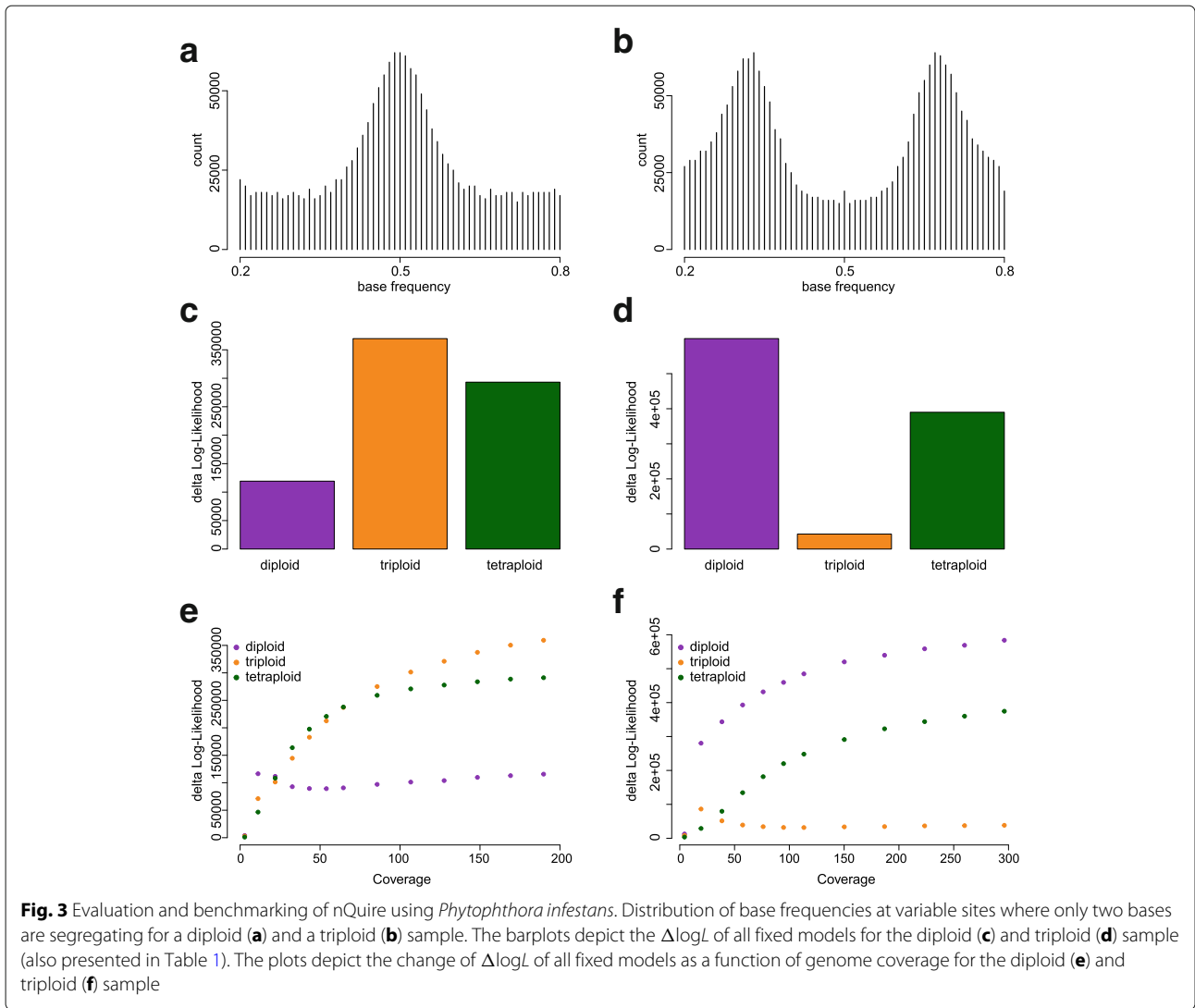
nQuire directly processes BAM files [17] and is designed to be efficient in memory usage and runtime. To process a

1GB *S. cerevisiae* BAM file (100x coverage), nQuire needs 70 s to build appropriate data structures, 6 s to run the models and calculate the maximum likelihood estimates, and uses a maximum of 8 Mb of RAM, whereas for processing a 10GB *P. infestans* BAM file (100x coverage) it needs 760 s, 100 s and 60 Mb of RAM, respectively. These benchmarks were performed on a single core of a Intel® Core™ i5-4670 CPU on a system with 16Gb of DDR3-1600 RAM and an SSD.

Analysis of individual samples

We evaluate nQuire’s performance using three *S. cerevisiae* samples at 100x coverage, which represent each of the three ploidy levels evaluated by the model, as well as two *P. infestans* samples, one diploid and one triploid, at 210x and 368x coverage, respectively. The $\Delta\log L$ of each of the fixed models to the free model at full coverage is shown in Table 1. At those coverages, the $\Delta\log L$ of the best model is more than two times closer to the free model than the second best.





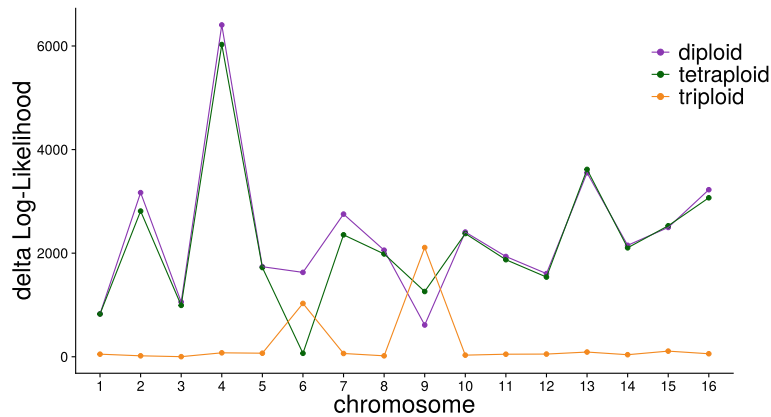


Fig. 5 Detection of chromosome-wide aneuploidies. The ploidy estimation can be run for each chromosome independently, which enables detection of aneuploidies. The sample displayed here shows genome wide triploidy, but splitting the analysis by chromosome shows tetraploidy for chromosome 6 and diploidy for chromosome 9, as detected by $\Delta\log L$

Additionally, it coincides in all samples with the ploidy level inferred by visually inspecting the empirical distributions of base frequencies at full coverage (Figs. 2a-c and 3a-b).

Coverage dependence

To investigate the impact of coverage on the performance of the GMM, we downsampled mapped reads from the *S. cerevisiae* (Fig. 2g-i) and *P. infestans* (Fig. 3e-f) strains shown in Table 1 to different coverage levels. In all cases the $\Delta\log L$ s of the two improper models increases with increasing coverage. In contrast, the $\Delta\log L$ of the best model stabilizes at low coverage and doesn't increase further as coverage increases. The coverage at which the $\Delta\log L$ of the best model is stable will be different for genomes of different size and complexity, as shown

in the difference between the two organisms used for benchmarking.

Analysis of population samples

In cases where multiple samples are sequenced simultaneously, it might be impractical to assess ploidy in each sample individually. In these cases, we propose to use maximized log-likelihoods of the three fixed models, normalized by that of the free model, to cluster samples into ploidy groups. The rationale is that within one species, the relative likelihoods of the fixed models will be similar within each ploidy level. As a proof of concept, we applied this to all di-, tri- and tetraploids from the *S. cerevisiae* test set [5], and clustered the samples into three groups in three dimensions using multivariate Gaussian clustering (see "Methods" section). The sample set was

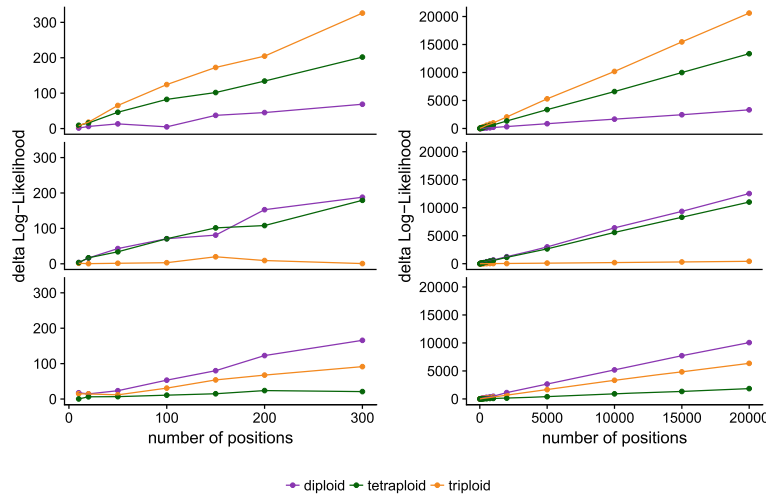


Fig. 6 Number of positions required for reliable ploidy assignment. The diploid (top), triploid (middle) and tetraploid (bottom) *S. cerevisiae* datasets from Table 1 were used to randomly sample 10 to 300 positions (left). The subsampling was extended to 20,000 positions (right)

manually scored for ploidy, and the overlap between clusters and manually assessed ploidy level was calculated (Fig. 4). Running nQuire on raw data showed high recovery of ploidy level (93%, Fig. 4a), which was further improved through our denoising implementation utilizing the GMMU (96%, Fig. 4b).

Detection of Aneuploidies

Recent polyploidization is often associated with aneuploidies. To be able to detect those, nQuire allows to split the analysis of a sample by regions defined in BED format. We used this to reanalyze the sample YJM466 from the *S. cerevisiae* test set [5]. This sample had been shown to be triploid on whole genome level, but tetraploid for chromosome 6 and diploid for chromosome 9. The $\Delta\log L$ s for the three fixed models individually calculated for each of the 16 chromosomes of *S. cerevisiae* confirmed this observation (Fig. 5).

A natural extension of analyzing the genome by chromosomes is to use sliding windows to detect possible transitions between ploidies in aneuploid individuals. We used the three *S. cerevisiae* datasets shown in Table 1 at their full coverage to benchmark the number of randomly sampled positions needed to accurately assign ploidy (Fig. 6). For these test samples, 100-200 random sites at 100x coverage are enough to correctly assign ploidy based on the $\Delta\log L$. However, this will vary for regions of the genome with different complexity and repetitiveness.

Discussion

In addition to nucleotide and structural variation, certain organisms can also vary intraspecifically in their ploidy level, which constitutes another source of variation that selection might act upon. Using NGS data our method permits assessment of ploidy variation from data that is usually generated for variant detection. In contrast to previous methods that visually analyze the distributions of SNPs at biallelic heterozygous sites [4, 6, 10], we quantitatively distinguish between different ploidy levels based on the distribution of base frequencies at variable sites, using relative differences in likelihoods. In comparison to the approach proposed by Gompert and Mock [12], nQuire avoids the requirement of high quality SNP calls. The higher level of noise in the data resulting from this is accounted for by using Gaussian distributions. They approximate a binomial process, but are impacted less by the effects of high coverage outliers, which arise for example from misalignments of paralogous sequences (Additional file 1: Figure S2). Additionally, nQuire is a Linux command line tool that uses standard file formats as input and handles large genomes at high coverage efficiently.

In all test cases, triploids were the easiest to distinguish, most likely caused by the lack of probability density

around 0.5 compared to the other two models. While diploids and tetraploids are more difficult to tease apart, our results on coverage dependence show that at sufficient coverage, the data fits the true model much better than either of the two alternatives (Fig. 2g-i). For cases where the Gaussian peaks were largely overlapped by uniform noise, we extended our free model to include a uniform component, whose mixture proportion can be used - after likelihood maximization - for base-line removal. We show that this procedure improves the recovery of the true ploidy level when samples are clustered based on maximized likelihoods under the assumptions of the fixed models (Fig. 4b). We also show that few high quality positions are enough to estimate the correct ploidy level (Fig. 6). Such high quality positions can be selected by using stringent filtering criteria. Several filters are implemented in nQuire directly, thus no pre-processing of the BAM file (after duplicate removal) is necessary. They include minimum and maximum coverage, as well as mapping quality and base frequency filters. The default values of these filters have been configured to fit most applications. The exact coverage and number of positions needed for a reliable estimation of ploidy will however depend on the complexity and repetitiveness of the genome. Additionally, it is possible to obtain high quality positions by using BED files to define regions of low repetitiveness, where base frequencies can be more confidently assessed.

Conclusion

We present nQuire, a statistical approach to distinguish diploids, triploids and tetraploids of recent evolutionary origin based on the distribution of base frequencies at variable sites. The method facilitates analysis of ploidy in single samples, and we demonstrate how to apply it to population scale data, when available. nQuire can also interact with BED files, to limit the analysis to certain sequence features, or divide it by regions of the genome, for example to detect aneuploidies. Our approach will be useful to assess intraspecific variation in ploidy from both historic and modern samples, as well as in experimental evolution experiments.

Additional file

Additional file 1: Supplementary figures addressing the effect of denoising, as well as the effect of high coverage outliers on the likelihood of Gaussian and binomial mixtures. (PDF 144 kb)

Abbreviations

BAM: Binary alignment map; BED: Browser extensible data; CPU: Central processing unit; EM: Expectation maximization; ENA: European nucleotide archive; GMM: Gaussian mixture model; GMMU: Gaussian mixture model with uniform noise component; NGS: Next generation sequencing; RAM: Random access memory; SNP: Single nucleotide polymorphism

Acknowledgments

We thank Michael Dannemann, Richard Neher, Thomas Mailund, Oliver Kohlbacher, Moises Exposito-Alonso, Kay Pruefer, and members of the Research Group for Ancient Genomics and Evolution (AGE) for useful discussions and input on model implementation; Patricia Lang, the AGE group and Michael Dannemann for comments on the manuscript.

Funding

This work was supported by the Presidential Innovation Fund of the Max Planck Society.

Availability of data and materials

The *Phytophthora infestans* sequencing data generated for this study is available at the European Nucleotide Archive (ENA) under study number PRJEB20998. The nQuire source code is available at <https://github.com/clwgg/nQuire> under the MIT license.

Authors' contributions

CLW, SK and HAB conceived the study. MP, LMC and CLW generated sequencing data. CLW developed the software application and analyzed the data. CLW and HAB wrote the the manuscript with input from all authors. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

¹Research Group for Ancient Genomics and Evolution, Department of Molecular Biology, Max Planck Institute for Developmental Biology, Tuebingen, Germany. ²The Sainsbury Laboratory, Norwich, UK. ³Department of Plant Pathology, Indian River Research and Education Center, University of Florida, Fort Pierce, USA.

Received: 14 December 2017 Accepted: 22 March 2018

Published online: 04 April 2018

References

- Selmecki AM, Maruyka YE, Richmond PA, Guillet M, Shores N, Sorenson AL, De S, Kishony R, Michor F, Dowell R, Pellman D. Polyploidy can drive rapid adaptation in yeast. *Nature*. 2015;519(7543):349–52. <https://doi.org/10.1038/nature14187>.
- Venkataram S, Dunn B, Li Y, Agarwala A, Chang J, Ebel E R, Geiler-Samerotte K, Hérisant L, Blundell JR, Levy SF, Fisher DS, Sherlock G, Petrov DA. Development of a comprehensive Genotype-to-Fitness map of Adaptation-Driving mutations in yeast. *Cell*. 2016;166(6):1585–9622. <https://doi.org/10.1016/j.cell.2016.08.002>.
- Comai L. The advantages and disadvantages of being polyploid. *Nat Rev Genet*. 2005;6(11):836–46. <https://doi.org/10.1038/nrg1711>.
- Yoshida K, Schuenemann VJ, Cano LM, Pais M, Mishra B, Sharma R, Lanz C, Martin FN, Kamoun S, Krause J, Thines M, Weigel D, Burbano HA. The rise and fall of the phytophthora infestans lineage that triggered the irish potato famine. *Elife*. 2013;2:00731. <https://doi.org/10.7554/eLife.00731>.
- Zhu YO, Sherlock G, Petrov DA. Whole genome analysis of 132 clinical *saccharomyces cerevisiae* strains reveals extensive ploidy variation. *G3*. 2016;6(8):2421–34. <https://doi.org/10.1534/g3.116.029397>.
- Li Y, Shen H, Zhou Q, Qian K, van der Lee T, Huang S. Changing ploidy as a strategy: The irish potato famine pathogen shifts ploidy in relation to its sexuality. 2017;30(1):45–52. <https://doi.org/10.1094/MPMI-08-16-0156-R>.
- Otto SP, Whitton J. Polyploid incidence and evolution. *Annu Rev Genet*. 2000;34:401–37. <https://doi.org/10.1146/annurev.genet.34.1.401>.

- Dirihan S, Terho P, Helander M, Saikkonen K. Efficient analysis of ploidy levels in plant evolutionary ecology. *Caryologia*. 2013;66(3):251–6. <https://doi.org/10.1080/00087114.2013.849414>.
- Margarido GRA, Heckerman D. ConPADE: genome assembly ploidy estimation from next-generation sequencing data. *PLoS Comput Biol*. 2015;11(4):1004229. <https://doi.org/10.1371/journal.pcbi.1004229>.
- Augusto Corrêa Dos Santos R, Goldman GH, Riaño-Pachón DM. ploidyNGS: visually exploring ploidy with next generation sequencing data. *Bioinformatics*. 2017;33(16):2575–6. <https://doi.org/10.1093/bioinformatics/btx204>.
- Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella KV, Altshuler D, Gabriel S, DePristo MA. From FastQ data to high confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr Protoc Bioinformatics*. 2013;43:11–10133. <https://doi.org/10.1002/0471250953.bi1110s43>.
- Gompert Z, Mock KE. Detection of individual ploidy levels with genotyping-by-sequencing (GBS) analysis. *Mol Ecol Resour*. 2017. <https://doi.org/10.1111/1755-0998.12657>.
- Scrucca L, Fop M, Murphy TB, Raftery AE. mclust 5: Clustering, classification and density estimation using gaussian finite mixture models. *R J*. 2016;8(1):289–317.
- R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. 2008. <http://www.R-project.org>.
- Meyer M, Kircher M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb Protoc*. 2010;2010(6):5448. <https://doi.org/10.1101/pdb.prot5448>.
- Cooke DEL, Cano LM, Raffaele S, Bain RA, Cooke LR, Etherington GJ, Deahl KL, Farrer RA, Gilroy EM, Goss EM, Grünwald NJ, Hein I, MacLean D, McNicol JW, Randall E, Oliva RF, Pel MA, Shaw DS, Squires JN, Taylor MC, Vleeshouwers VGAA, Birch PRJ, Lees AK, Kamoun S. Genome analyses of an aggressive and invasive lineage of the irish potato famine pathogen. *PLoS Pathog*. 2012;8(10):1002940. <https://doi.org/10.1371/journal.ppat.1002940>.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. The sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078–9. <https://doi.org/10.1093/bioinformatics/btp352>.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit

