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

Clemens L. Weiß[1,*], Marina Pais[2], Liliana M. Cano[2,3], Sophien Kamoun[2], Hernán A. Burbano[1,*]

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

*** To whom correspondence should be addressed.**

Abstract

Summary: nQuire is a statistical framework that distinguishes between diploids, triploids and tetraploids using next generation sequencing. 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.
Availability and Implementation: The model is implemented as a stand-alone Linux command line tool in the C programming language and is available at github under the MIT licence. Please also refer to github.com/clwgg/nQuire for usage instructions.
Contact: clemens.weiss@tuebingen.mpg.de or hernan.burbano@tuebingen.mpg.de

Introduction

Polyploidy, the presence of more than two complete sets of chromosomes, can both fuel and hinder adaptation [1, 8, 10]. Polyploidization can lead to aneuploidy, which burdens mating due to the presence of individuals of different ploidy. Therefore, intraspecific variation in ploidy tends to occur - although not exclusively - in organisms that have the capacity to reproduce asexually [5, 11, 12], are self-compatible or are perennial [7].

Ploidy can be inferred from Next Generation Sequencing (NGS) data, for instance, by assessing the distribution of allele frequencies at biallelic Single Nucleotide Polymorphisms (SNPs) [11, 12]. This 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. To determine the ploidy level, the distribution of biallelic SNPs can be inspected visually and/or qualitatively compared with simulated data [11]. However, this methodology is not quantitative and relies on the identification of variable sites ("SNP calling"), which is performed using approaches that benefit from a previously known ploidy level [9]. A further development based on biallelic SNPs uses a Bayesian statistical approach to estimate allelic ratios followed by a clustering procedure that helps discriminating between ploidy levels from Genotyping-By-Sequencing data [3].

Here we present a new 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 at different coverages using published genomes of *Saccharomyces cerevisiae* [12] and high-coverage genomes of *Phytophthora infestans* produced for this study.

Methods

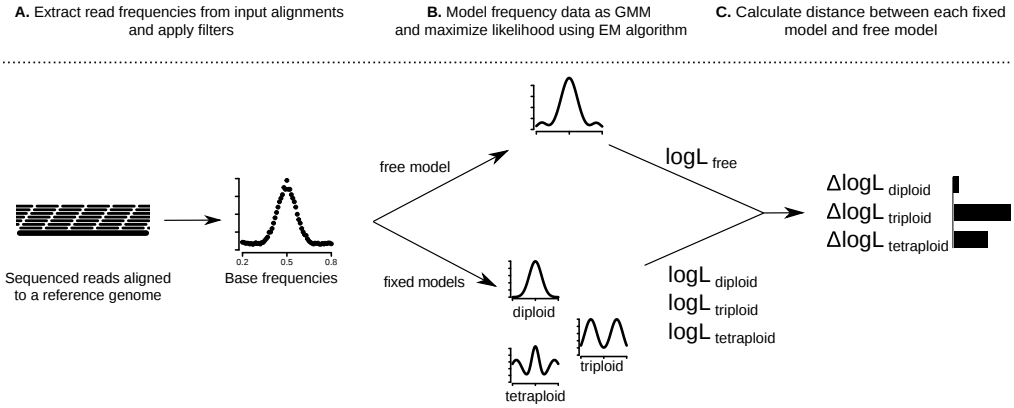

Figure 1. Overview of the Gaussian Mixture Model (GMM) based method used by nQuire to estimate ploidy. We illustrate the workflow using a diploid individual as example. A. After sequenced reads are mapped to a reference genome, base frequencies are calculated at variable sites where only two bases are segregating. B. The base frequencies are modeled using a GMM and the likelihood is maximized using an Expectation-Maximization (EM) algorithm for both the free and the three fixed models (diploid, triploid and tetraploid). The maximized log-likelihoods ($logL$) are extracted for subsequent model comparison. The curves show a possible final state of the GMM under the assumptions of each of the four models. C. The $\Delta logL$ is calculated between the free model and each of the three fixed models (here represented as barplots). The fixed model with the smallest $\Delta logL$ is chosen as the ploidy level (diploid in this example).

Implementation

To distinguish between diploids, triploids and tetraploids based on their distributions of base frequencies at variable sites with only two bases segregating (Figure 1A), we implemented a GMM that models the base frequency profiles as a mixture of three Gaussian distributions (Figure 1B), which are scaled relative to each other as:

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

Here, n describes the numbers of data points, x_i describes the value of each data point (i.e. the base frequency) and 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. The only constraint here is that $\sum_{j=1}^{3} \alpha_j = 1$.

This model allows 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 (Figure 1B), which is specific to the GMM but can be extended to similar models. The algorithm estimates all parameters at once and computes a likelihood ("free model"). Alternatively, a likelihood can be calculated when parameters are held constant ("fixed models") to the expected values under diploidy (one Gaussian with mean 0.5), triploidy (two Gaussian with means 0.33 and 0.67) and tetraploidy (three Gaussian with means 0.25, 0.5 and 0.75). Since all fixed models are nested within the free model, it is possible to directly compute the log-likelihood ratios, following:

$$\Delta logL_{diploid} = logL_{free} - logL_{diploid}$$
$$\Delta logL_{triploid} = logL_{free} - logL_{triploid}$$
$$\Delta logL_{tetraploid} = logL_{free} - logL_{tetraploid}$$

The $\Delta logL$s describe the distance between each fixed model and the best fit under the assumptions of the GMM. A substantially lower $\Delta logL$ of one fixed model over the others supports the ploidy level described by this fixed model (Figure 1C). Therefore, we used $\Delta logL$ 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:

$$logL = \sum_{i=1}^{n} log \left[\alpha_4 \times U(x_i) + \sum_{j=1}^{3} \alpha_j \times N(x_i|\mu_j, \sigma_j) \right]$$

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. This is important when the Gaussian peaks are observable but embedded in a basal noise, which could be caused by highly repetitive genomes or low coverage.

Phytophthora infestans libraries

The two benchmarking libraries from *P. infestans* were generated according to the protocol by Meyer and Kircher [6] from DNA extracted from cultures [2]. The libraries were sequenced to high coverage on an Illumina HiSeq 3000 machine in paired end 150bp mode. This read data is available at the European Nucleotide Archive (ENA) under study number PRJEB20998.

Results

Method evaluation

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 $\Delta logL$ of each of the fixed models to the free model at full coverage is shown in Table 1. At those coverages, the $\Delta logL$ of the best model is more than two times closer to the free model than the second best. Also, it coincides in all samples with the ploidy level inferred by visually inspecting the empirical distributions of base frequencies at

Table 1. Samples of *Saccharomyces cerevisiae* and *Phytophthora infestans* used to evaluate and benchmark nQuire. The smallest $\Delta logL$ for each sample is highlighted in bold.

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

$^{+}$ Average per-base coverage

full coverage (Figure S1A-C and S2A-B). To investigate the impact of coverage on the performance of the GMM, we downsampled mapped reads from all *S. cerevisiae* (Figure S1G-I) and *P. infestans* (Figure S2E-F) strains to different coverage levels ranging from almost full to 1x coverage. These analyses showed that while the $\Delta logL$s between the free model and the true fixed model start to plateau at low coverage, the $\Delta logL$ between the free model and the two incorrect fixed models keeps increasing at higher coverage (Figure S1G-I and S2E-F).

Method performance

nQuire directly processes BAM files [4] and is designed to be efficient in memory usage and runtime. To process a 1GB *S. cerevisiae* BAM file (100x coverage), nQuire needs 70 seconds to build appropriate data structures, 6 seconds 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 seconds, 100 seconds and 60 Mb of RAM, respectively.

Conclusion

We present nQuire, a statistical approach to distinguish diploids, triploids and tetraploids from NGS data. In comparison to a previous quantitative approach [3], nQuire requires neither SNP calls nor reference individuals with known ploidy. The higher level of noise resulting from omitting SNP calling is accounted for by using Gaussian distributions to approximate a binomial process, since such distributions are impacted less by the effects of outliers. nQuire will be useful to assess intraspecific variation in ploidy from both historic and modern samples, as well as in experimental evolution experiments.

Acknowledgements

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; the AGE group and Michael Dannemann for comments on the manuscript; and the Presidential Innovation Fund of the Max Planck Society for financial support.

References

1. L. Comai. The advantages and disadvantages of being polyploid. *Nat. Rev. Genet.*, 6(11):836–846, Nov. 2005.

2. D. E. L. Cooke, L. M. Cano, S. Raffaele, R. A. Bain, L. R. Cooke, G. J. Etherington, K. L. Deahl, R. A. Farrer, E. M. Gilroy, E. M. Goss, N. J. Grünwald, I. Hein, D. MacLean, J. W. McNicol, E. Randall, R. F. Oliva, M. A. Pel, D. S. Shaw, J. N. Squires, M. C. Taylor, V. G. A. A. Vleeshouwers, P. R. J. Birch, A. K. Lees, and S. Kamoun. Genome analyses of an aggressive and invasive lineage of the irish potato famine pathogen. *PLoS Pathog.*, 8(10):e1002940, 4 Oct. 2012.

3. Z. Gompert and K. E. Mock. Detection of individual ploidy levels with genotyping-by-sequencing (GBS) analysis. *Mol. Ecol. Resour.*, 1 Feb. 2017.

4. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 15 Aug. 2009.

5. Y. Li, H. Shen, Q. Zhou, K. Qian, T. van der Lee, and S. Huang. Changing ploidy as a strategy: The irish potato famine pathogen shifts ploidy in relation to its sexuality. *Mol. Plant. Microbe. Interact.*, 30(1):45–52, Jan. 2017.

6. M. Meyer and M. Kircher. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.*, 2010(6):db.prot5448, June 2010.

7. S. P. Otto and J. Whitton. Polyploid incidence and evolution. *Annu. Rev. Genet.*, 34:401–437, 2000.

8. A. M. Selmecki, Y. E. Maruvka, P. A. Richmond, M. Guillet, N. Shoresh, A. L. Sorenson, S. De, R. Kishony, F. Michor, R. Dowell, and D. Pellman. Polyploidy can drive rapid adaptation in yeast. *Nature*, 519(7543):349–352, 19 Mar. 2015.

9. G. A. Van der Auwera, M. O. Carneiro, C. Hartl, R. Poplin, G. Del Angel, A. Levy-Moonshine, T. Jordan, K. Shakir, D. Roazen, J. Thibault, E. Banks, K. V. Garimella, D. Altshuler, S. Gabriel, and M. A. DePristo. From FastQ data to high confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinformatics*, 43:11.10.1–33, 2013.

10. S. Venkataram, B. Dunn, Y. Li, A. Agarwala, J. Chang, E. R. Ebel, K. Geiler-Samerotte, L. Hérissant, J. R. Blundell, S. F. Levy, D. S. Fisher, G. Sherlock, and D. A. Petrov. Development of a comprehensive Genotype-to-Fitness map of Adaptation-Driving mutations in yeast. *Cell*, 166(6):1585–1596.e22, 8 Sept. 2016.

11. K. Yoshida, 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. The rise and fall of the phytophthora infestans lineage that triggered the irish potato famine. *Elife*, 2:e00731, 28 May 2013.

12. Y. O. Zhu, G. Sherlock, and D. A. Petrov. Whole genome analysis of 132 clinical saccharomyces cerevisiae strains reveals extensive ploidy variation. *G3*, 6(8):2421–2434, 9 Aug. 2016.