Detection of copy number variation for chromosomal sliding windows using high throughput sequencing data in the R environment

Brian J. Knaus and Niklaus J. Grünwald
USDA, ARS, Horticultural Crops Research Unit

Rationale

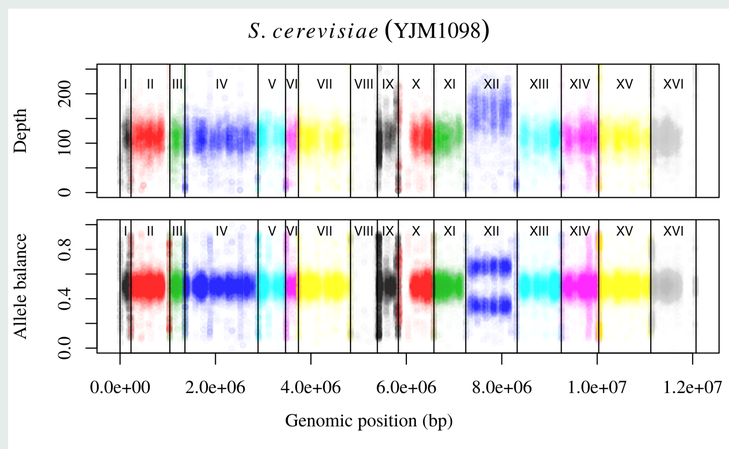
Inference of copy number variation presents a technical challenge because variant callers typically require the copy number of a genome or genomic region to be known a priori. Our project required us to address this question, so we designed and implemented a method with the following:

- No a priori known base ploidy is required
- Allows for genomic windows to be analyzed
- Flexibility to use with non-model organisms
- Works with VCF format data
- Implemented in R

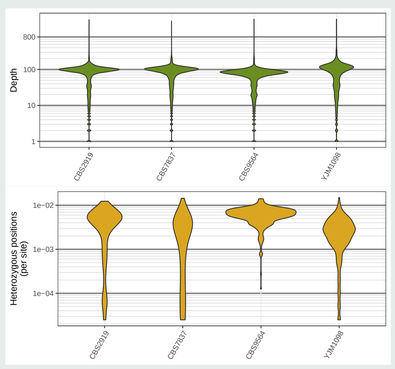
We validated these approaches with the model system of *Saccharomyces cerevisiae*, an organism known to vary in ploidy and copy number. This method has been implemented in the R package vcfR.

Chromosomal perspectives

Sequence depth can be used to characterize base ploidy and deviations from base ploidy. However, if a research question includes 'what is base ploidy' we need another perspective. Allele balance can help provide inferences on whether base ploidy is diploid, triploid, or tetraploid. This is a reproduction of Figure 7 from Zhou et al. (2016) created in vcfR and highlights chromosome XII as having three copies while base ploidy is two copies.



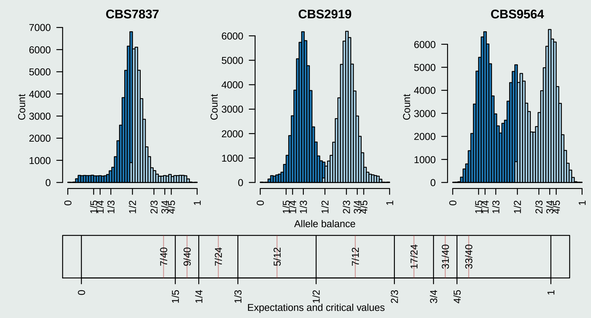
Depth and heterozygosity



Here we summarize the per window sequencing depth and rate of heterozygosity throughhout *S. cerevisiae* genomes. This can be used to identify regions of the genome that are anomalous. For example, we observe long tails of low heterozygosity that may be regions of lost heterozygosity.

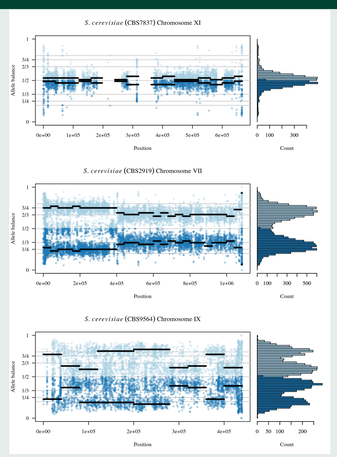
Allele balance

Allele balance is the frequency that the most abundant and second most abundant allele were sequenced at. For diploids we expect half of the sequences to be from each allele. For triploids we expect the alleles to be sequenced at frequencies of one thirds (1/3 and 2/3). For tetraploids we expect the alleles to be sequenced at frequencies of quarters (1/4, 1/2, and 3/4). This information can be used to assign a copy number to a genome or a genomic fraction.



Windowing

Allele balance can also be used to assign copy number to genomic windows of variants. Windows of user specified lengths can be made and allele balance is estimated for each window. A distance from expectation and the number of heterozygous position for each window are reported to help determine the quality of estimates. Estimates determined to be of low quality may be censored.



Acknowledgments

This research is supported in part by U.S. Department of Agriculture (USDA) Agricultural Research Service Grant 5358-22000-039-00D and USDA National Institue of Food and Agriculture Grant 2011-68004-30154.

Saccharomyces cerevisiae data from: 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.

vcfR can be found at: https://CRAN.R-project.org/package=vcfR