

SNP and Dosage calling

Genetic data analysis in polyploids: From allelic dosage to QTL mapping

Cristiane Taniguti
Gabriel Gesteira
Jeekin Lau
Zhao-Bang Zeng
David Byrne
Oscar Riera-Lizarazu
Marcelo Mollinari

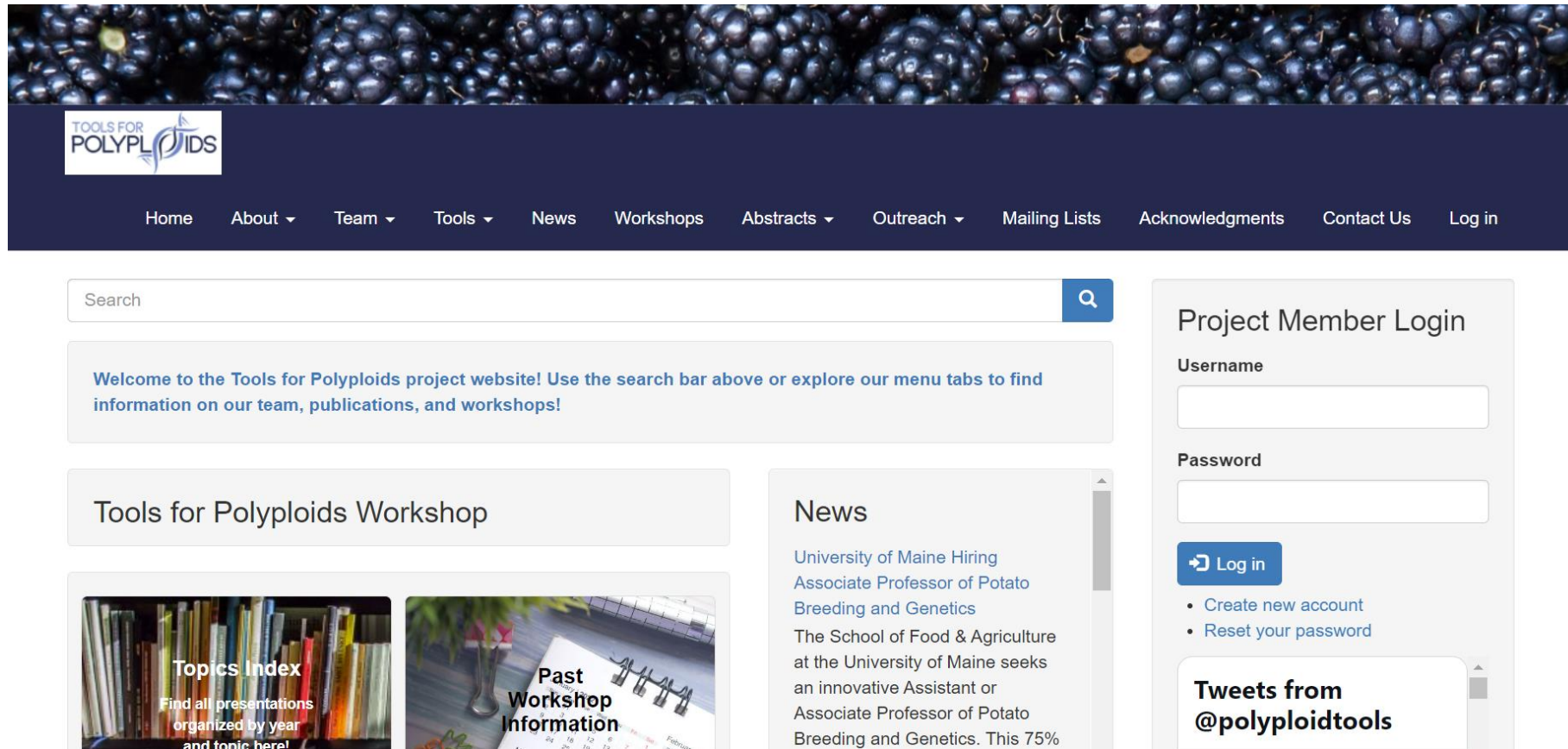


NC STATE
UNIVERSITY



polyploids.org

► Tools for Polyploids Workshop 2023 (January 12-13)



The screenshot shows the homepage of the Tools for Polyploids project website. At the top is a banner image of dark grapes. Below it is the project logo and a navigation menu with links: Home, About, Team, Tools, News, Workshops, Abstracts, Outreach, Mailing Lists, Acknowledgments, Contact Us, and Log in. A search bar is located below the navigation. A welcome message reads: "Welcome to the Tools for Polyploids project website! Use the search bar above or explore our menu tabs to find information on our team, publications, and workshops!". The main content area is divided into two columns. The left column features a "Tools for Polyploids Workshop" section with two sub-sections: "Topics Index" (with a bookshelf image) and "Past Workshop Information" (with a calendar image). The right column features a "News" section with a headline: "University of Maine Hiring Associate Professor of Potato Breeding and Genetics". The text below the headline states: "The School of Food & Agriculture at the University of Maine seeks an innovative Assistant or Associate Professor of Potato Breeding and Genetics. This 75%". To the right of the main content is a "Project Member Login" form with fields for Username and Password, a "Log in" button, and links for "Create new account" and "Reset your password". Below the login form is a "Tweets from @polyploidtools" section.

Outline

Genome variations

Sequencing libraries types

Sequencing experiment planning

Genotyping-by-Sequencing

SNP calling

Errors sources

Dosage calling

Which is the best pipeline?

Tutorial

Polyploid species



- ▶ Organisms that have multiple copies of the complete set of chromosomes
- ▶ Genome variations - applications
 - ▶ Quantitative traits mapping
 - ▶ Genome Wide Association studies
 - ▶ Phenotypic predictions - Genome Selection
 - ▶ Evolution and diversity studies
 - ▶ Gene expression studies

Genome variations

- ▶ Short sequences (SNPs, indels)
- ▶ Structural variants (number of copies, inversions, translocations)

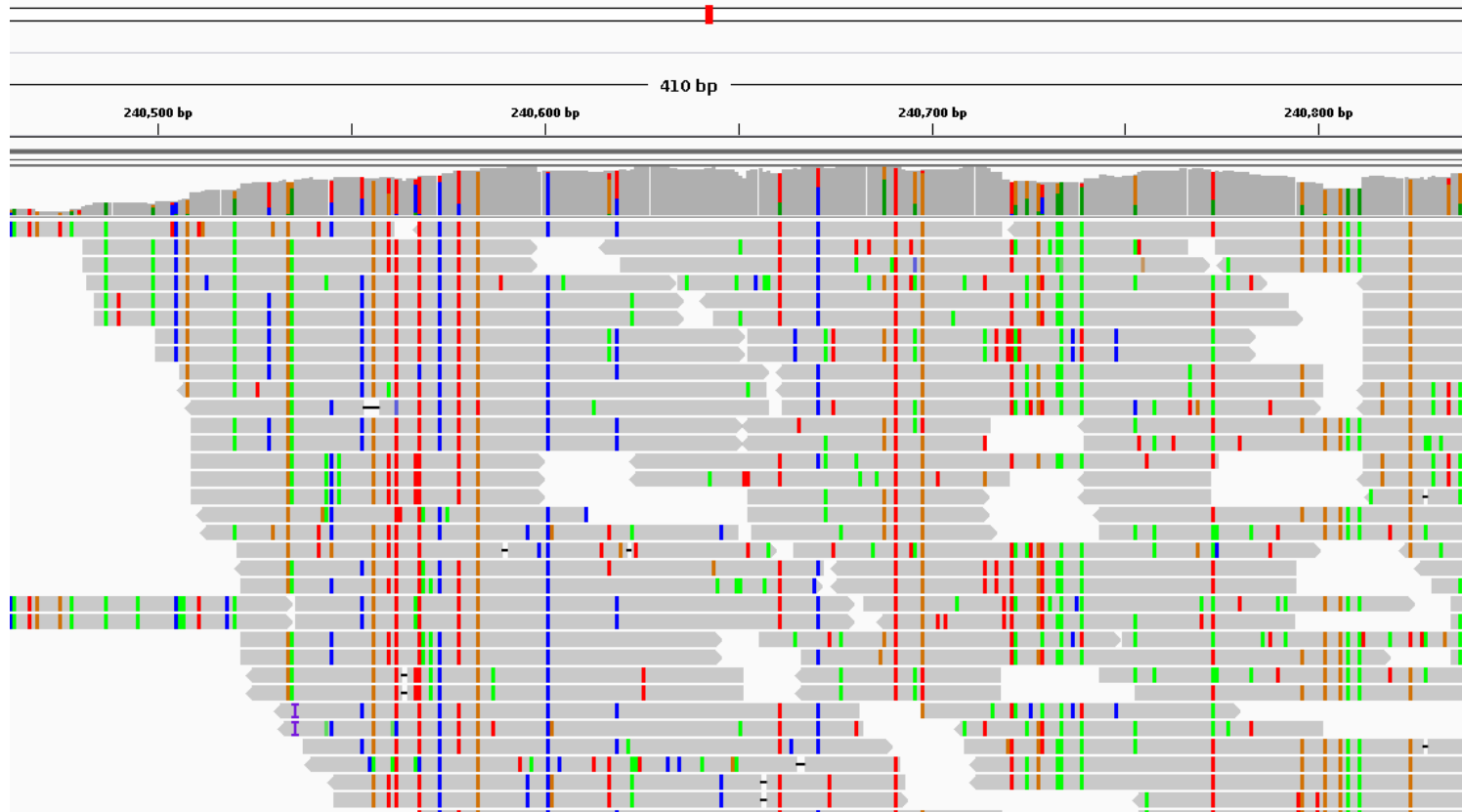
Molecular markers

- ▶ RFLP, RAPD, AFLP, and SSR
- ▶ Arrays (For Roses: \$\$\$\$\$)
- ▶ Sequencing (For Roses: \$)

Sequencing libraries

- ▶ Whole Genome Sequencing (WGS)

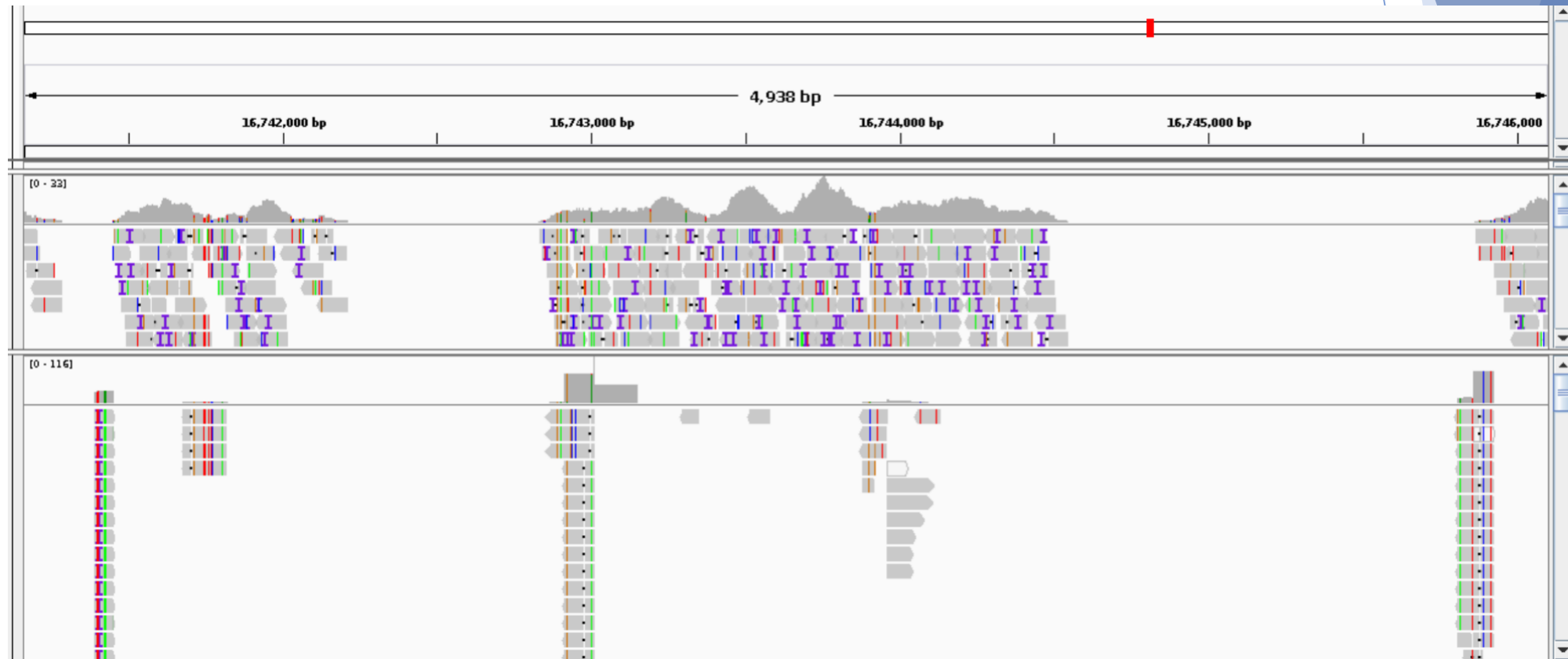
Image: IGV



Sequencing libraries

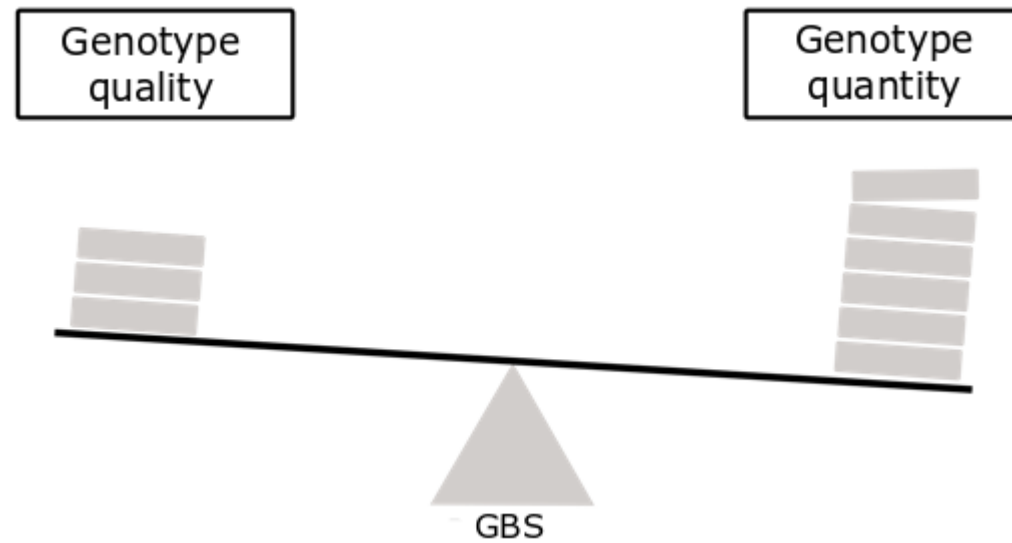
- ▶ Exome sequencing (top) and Genotyping-by-Sequencing (bottom)

Image: IGV

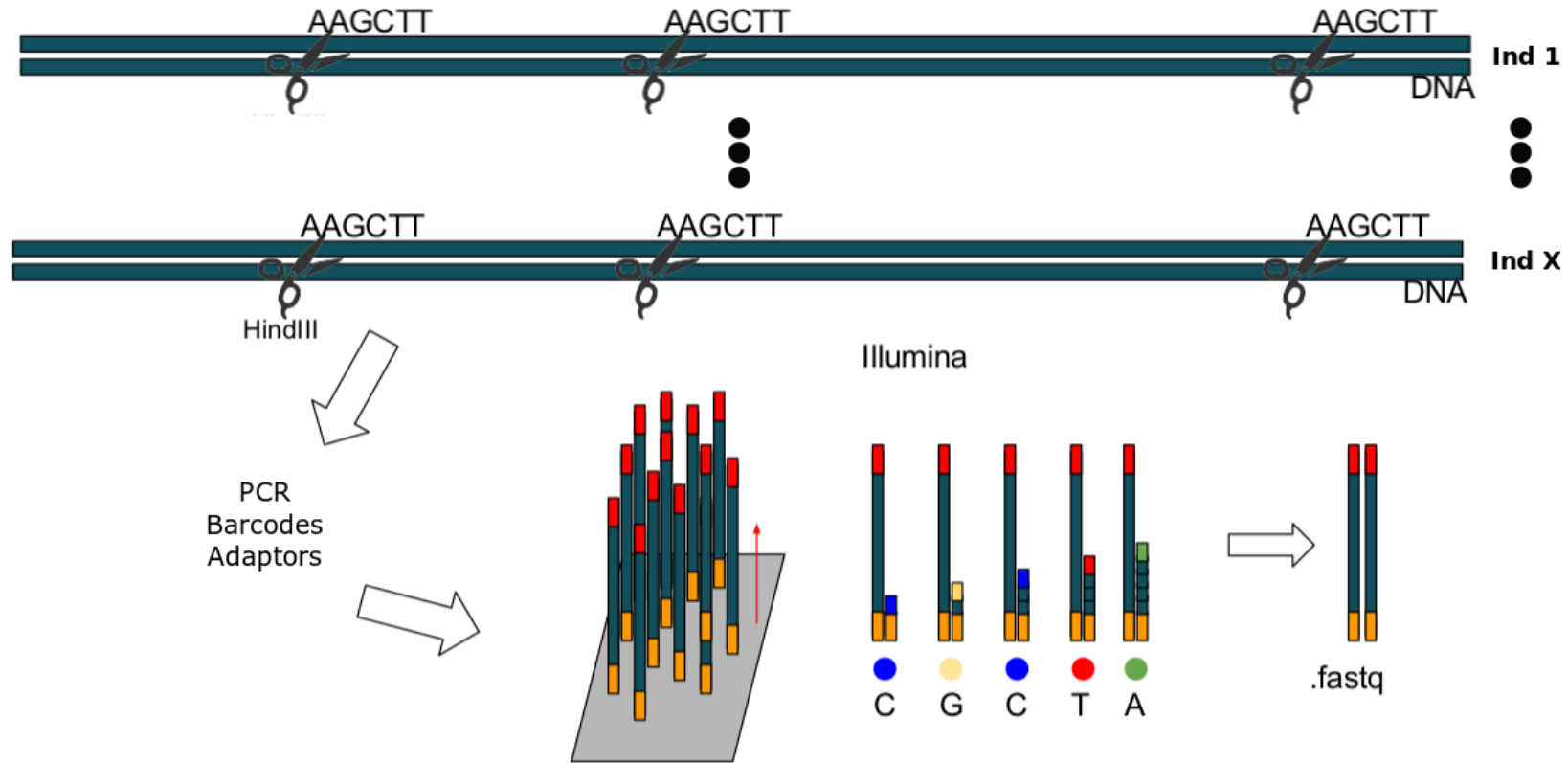


Sequencing experiment design

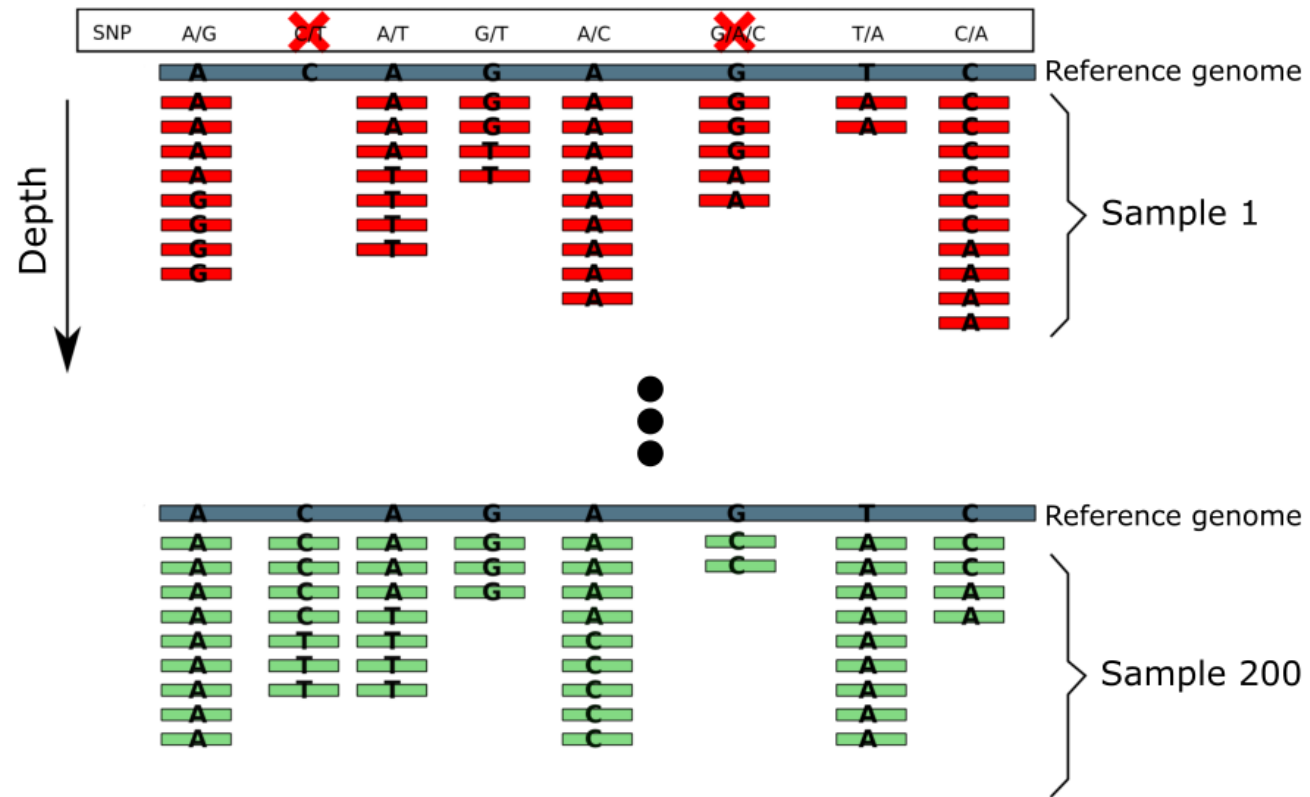
- ▶ Study goal
- ▶ Sequencer capacity
- ▶ Number of individuals per lane
- ▶ Number of sequenced loci



GBS methods



SNP calling



SNP calling

- ▶ STACKS (Catchen et al., 2013)
 - ▶ Focus on diploid RADseq data
 - ▶ No need for a reference genome
 - ▶ Requires previous efficient sequences filtering
- ▶ TASSEL (Glaubitz et al., 2014)
 - ▶ Focus on diploid RADseq data
 - ▶ No need for a reference genome
 - ▶ Adaptations for polyploids (Pereira et al., 2018)

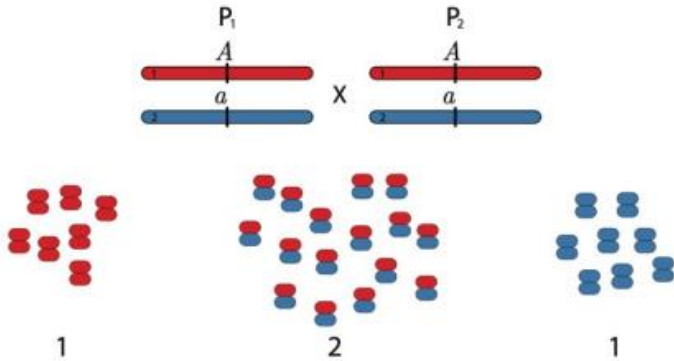
SNP calling

- ▶ Freebayes (Garrison and Marth, 2012)
 - ▶ Any library type
 - ▶ Diploids and polyploids

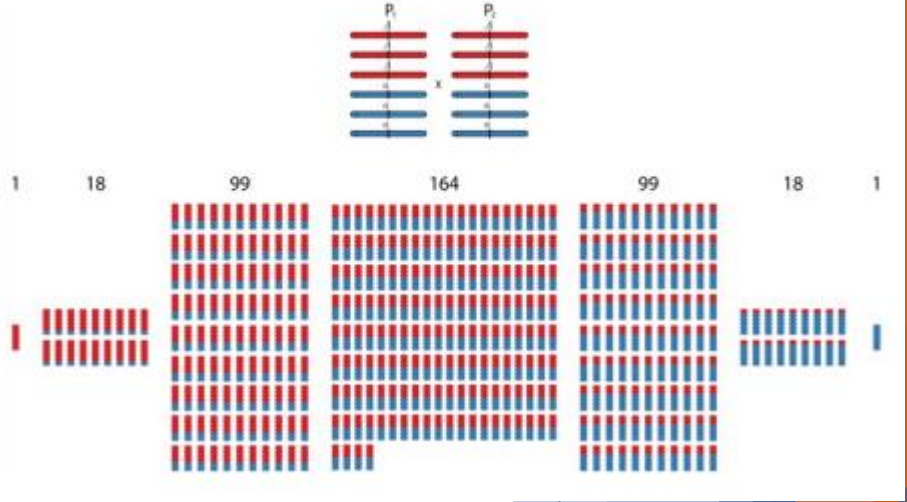
- ▶ GATK (McKenna et al., 2012)
 - ▶ Focus on WGS or target enrichment libraries
 - ▶ Diploids and polyploids
 - ▶ Implemented in GBSapp (Wadl et al., 2018)

Dosage calling

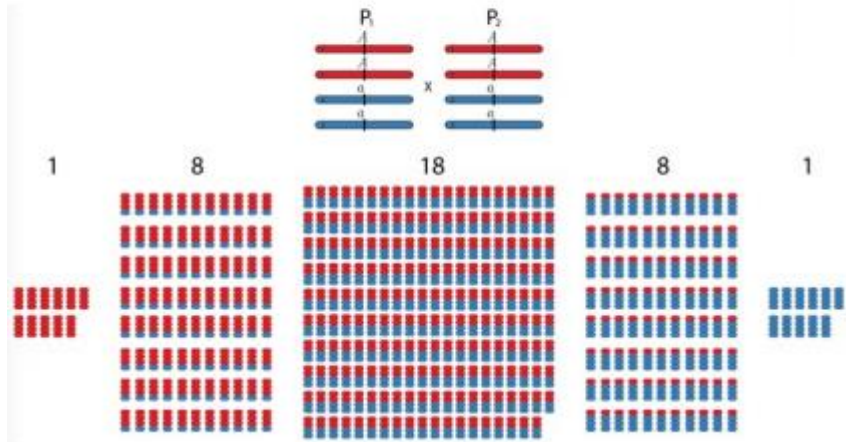
Diploid



Hexaploid



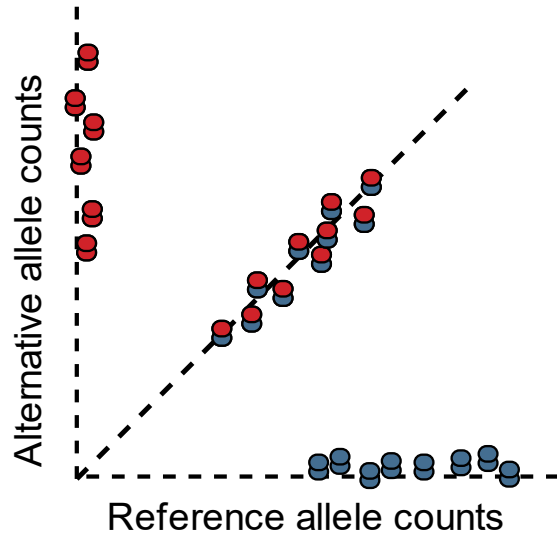
Tetraploid



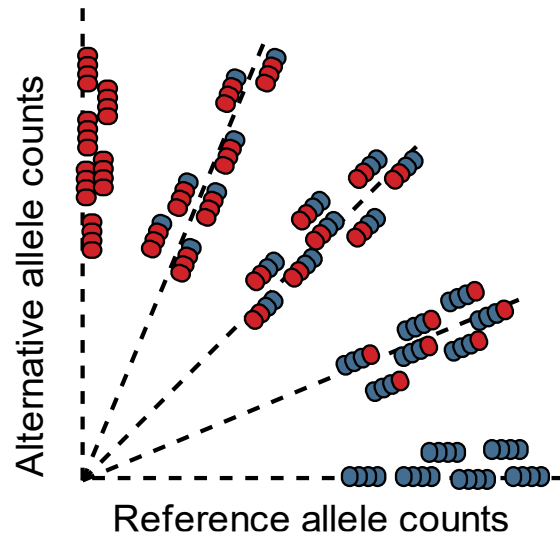
Dosage calling

- ▶ The theory

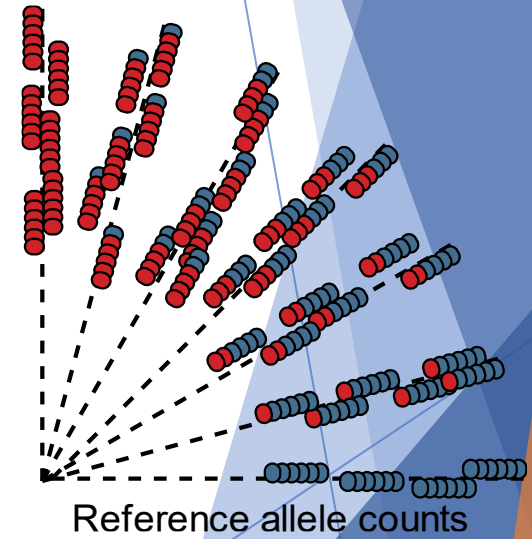
Diploid



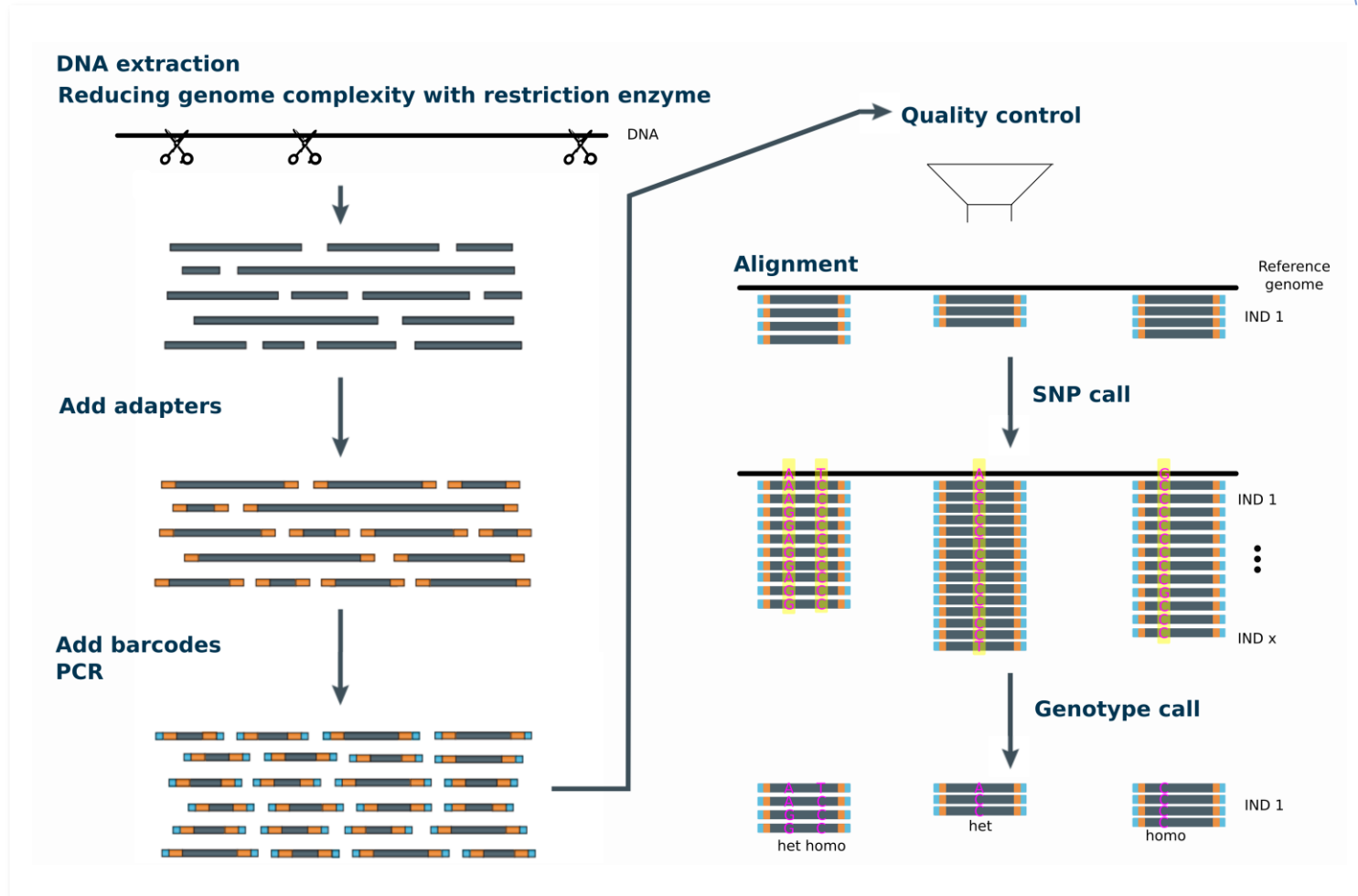
Tetraploid



Hexaploid



Sources of errors



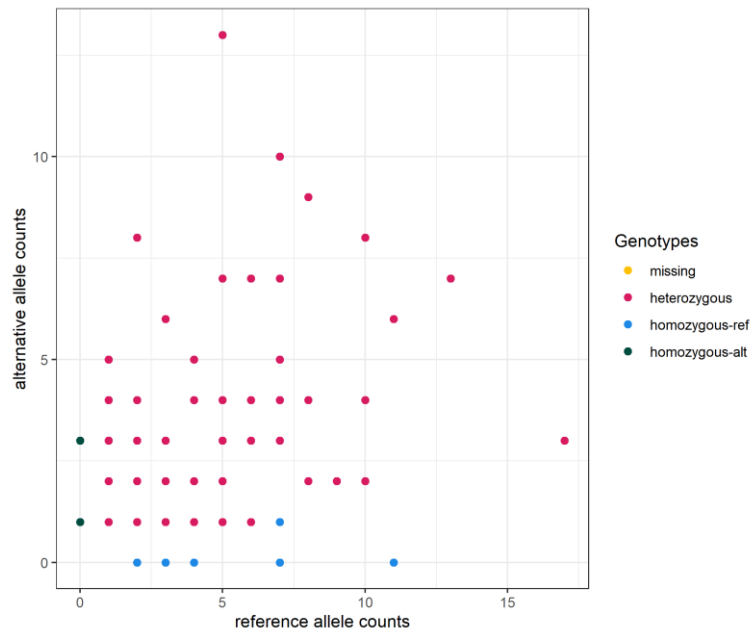
Source of errors

► The reality

Diploid (mean depth 6)

N = 200

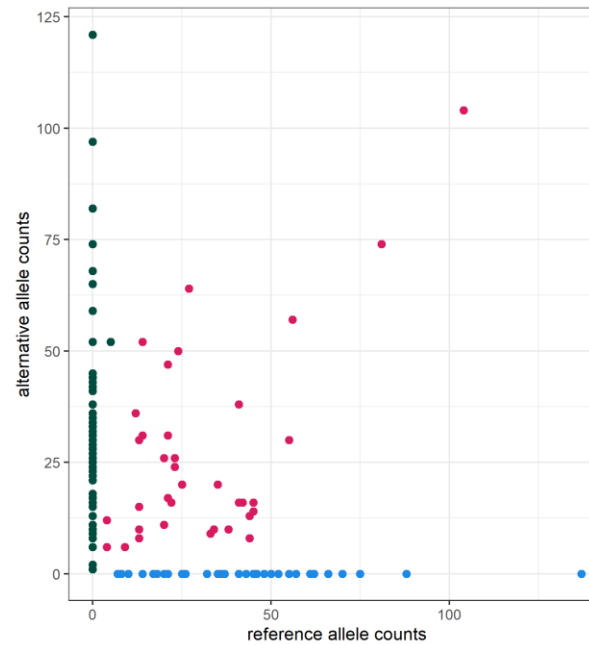
Aa x Aa



Diploid (mean depth 96)

N = 138

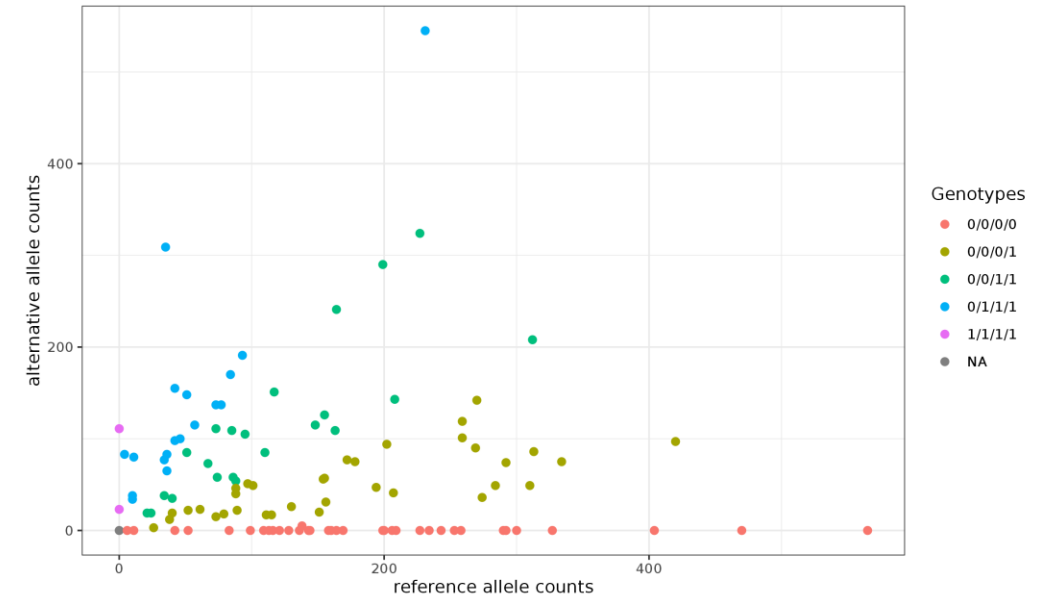
Aa x Aa



Tetraploid (mean depth 83)

N = 114

AAaa x AAaa

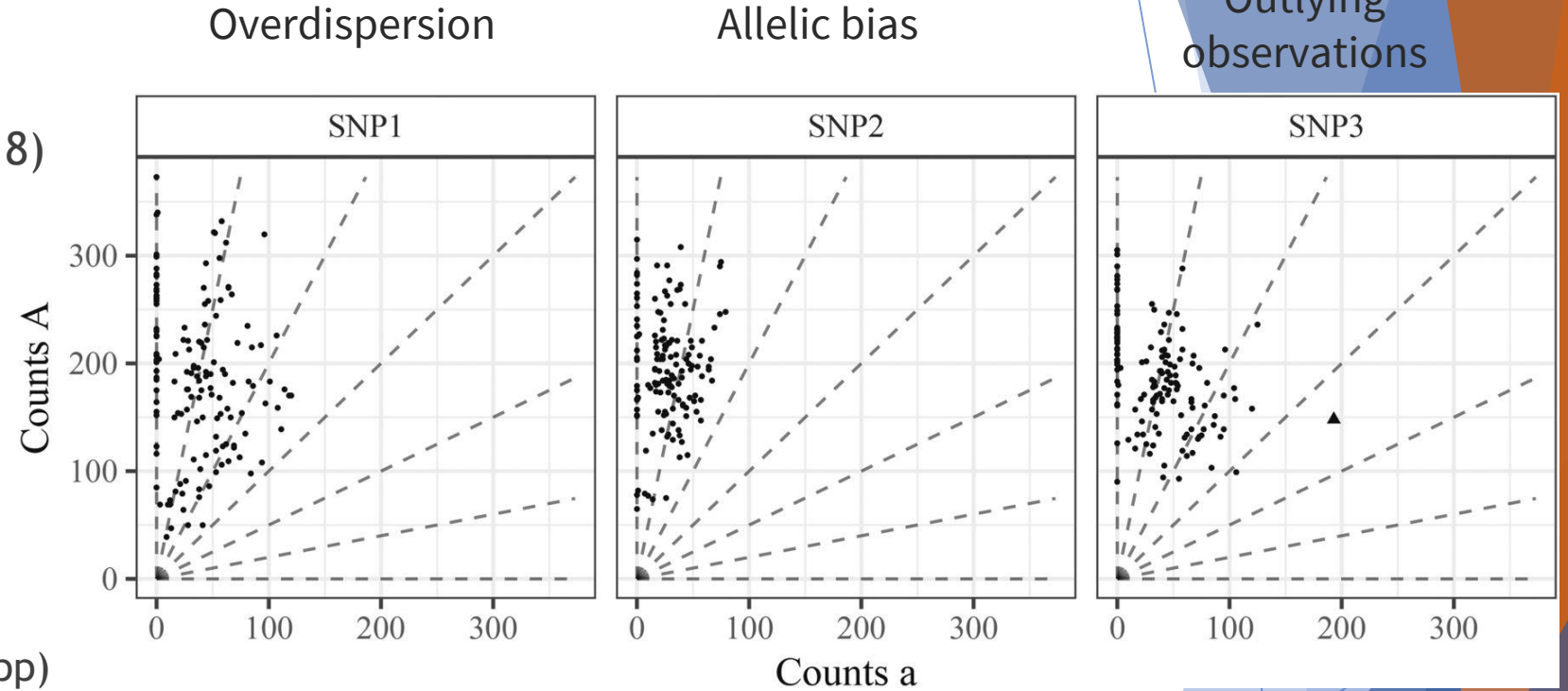


Dosage calling

- ▶ Freebayes (Garrison and Marth, 2012)
 - ▶ Alignment quality
 - ▶ Base call quality around indels
 - ▶ Depth
- ▶ GATK (McKenna et al., 2010)
 - ▶ Alignment quality
 - ▶ Base call quality of SNPs and indels
 - ▶ Depth
 - ▶ Hard filtering

Dosage calling

- ▶ updog (Gerard et al., 2018)
 - ▶ Any ploidy
 - ▶ Allelic bias
 - ▶ Overdispersion
 - ▶ Sequencing errors
 - ▶ Outliers
 - ▶ Population structure
- (F1, S1, HW, F1pp, S1pp)



Gerard et al., 2018

Dosage calling

- ▶ SuperMASSA (Serang et al., 2012)
 - ▶ Any ploidy and variable ploidy
 - ▶ Overdispersion
 - ▶ Population structure (F1 and HW)
- ▶ polyRAD (Clark et al., 2019)
 - ▶ Any ploidy
 - ▶ Sequencing errors
 - ▶ Population structure (F1, S1 and HW)

Which is the best pipeline?

▶ Challenges:

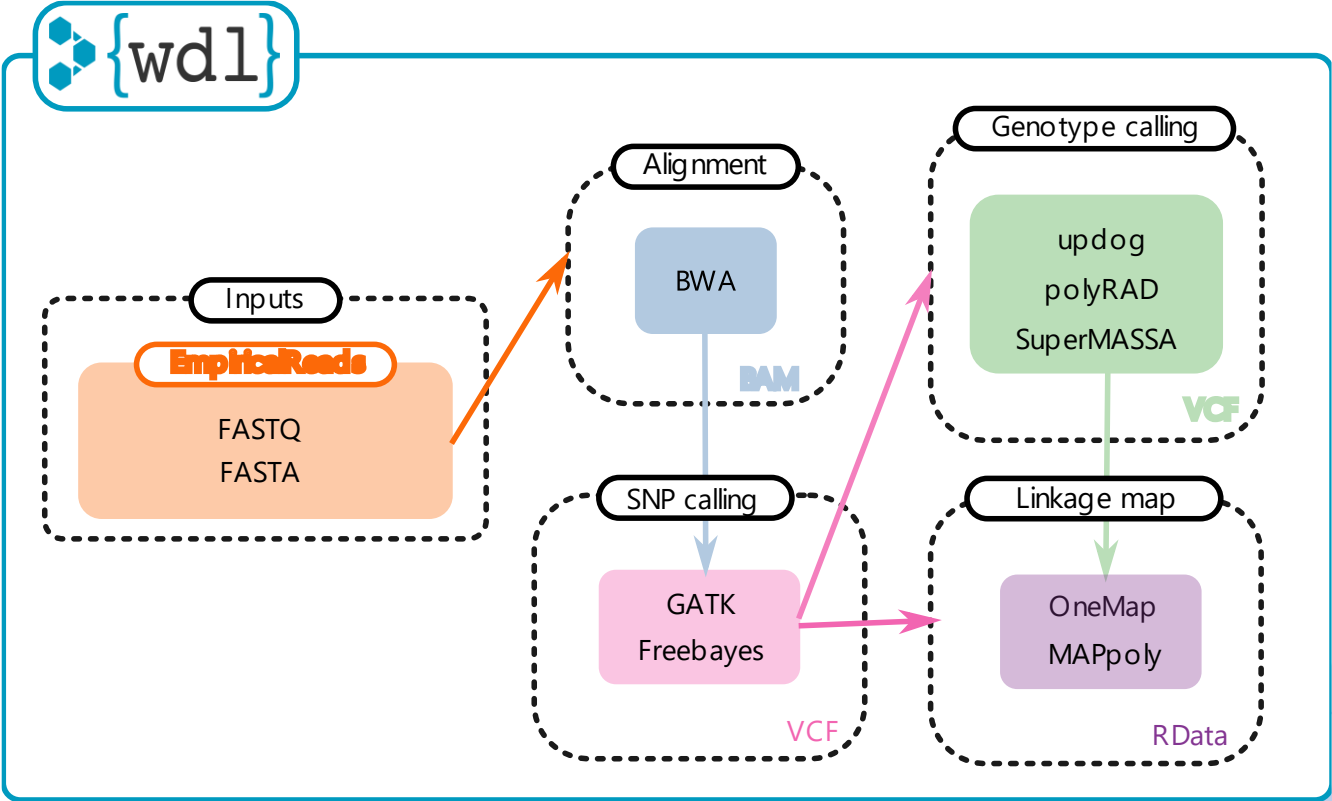
- ▶ Many software, many dependencies
- ▶ Different input and output formats
- ▶ Collaborative work
- ▶ Computational resources
- ▶ Quality criteria
- ▶ Explore and visualize results
- ▶ Reproducibility
- ▶ Adapt to software updates

▶ Useful tools:

- ▶ Containers (Docker and singularity)
- ▶ Workflow Description Language (WDL)
- ▶ GitHub
- ▶ HPC and Google Cloud
- ▶ Linkage map
- ▶ Shiny

Reads2Map

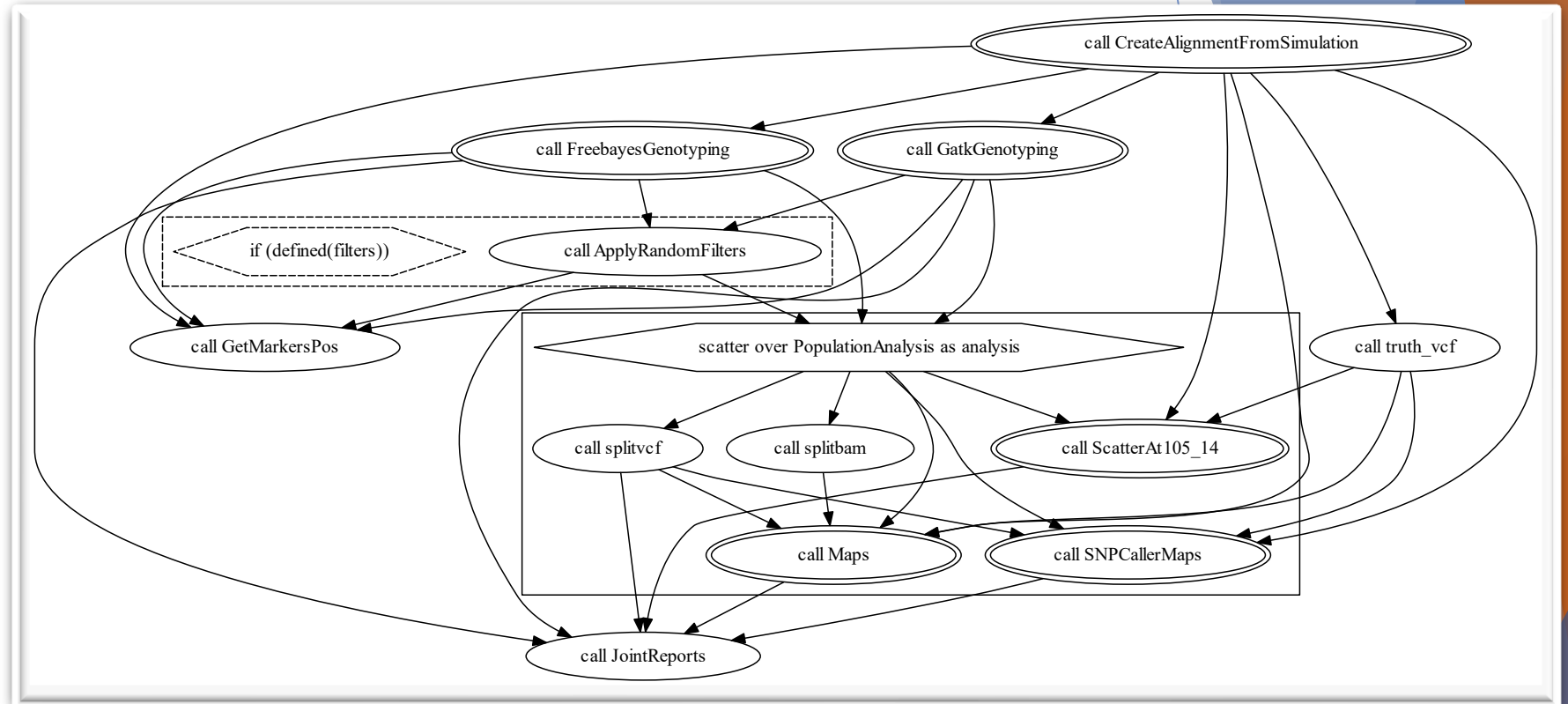
- ▶ Join several bioinformatics and statistical analyses
- ▶ Best practices



Implementation



- ▶ Workflows
 - ▶ Sub-workflows
 - ▶ Tasks

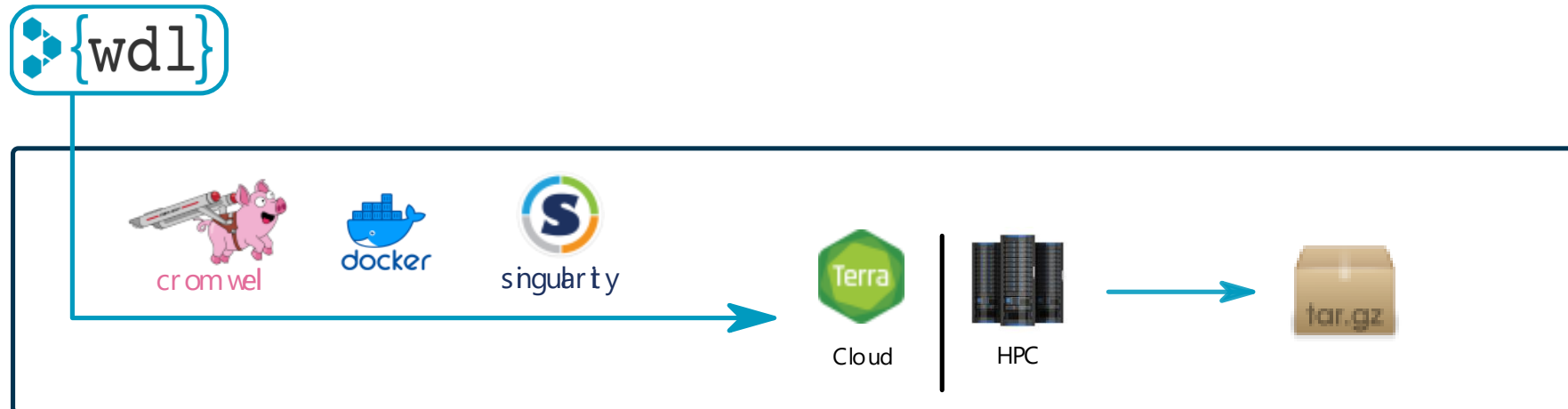


```
$ java -jar /path/to/womtool.jar graph tasks/SimulatedSingleFamily.wdl > SimulatedSingleFamily.dot  
$ dot -Tsvg SimulatedSingleFamily.dot -o SimulatedSingleFamily.svg
```


Implementation

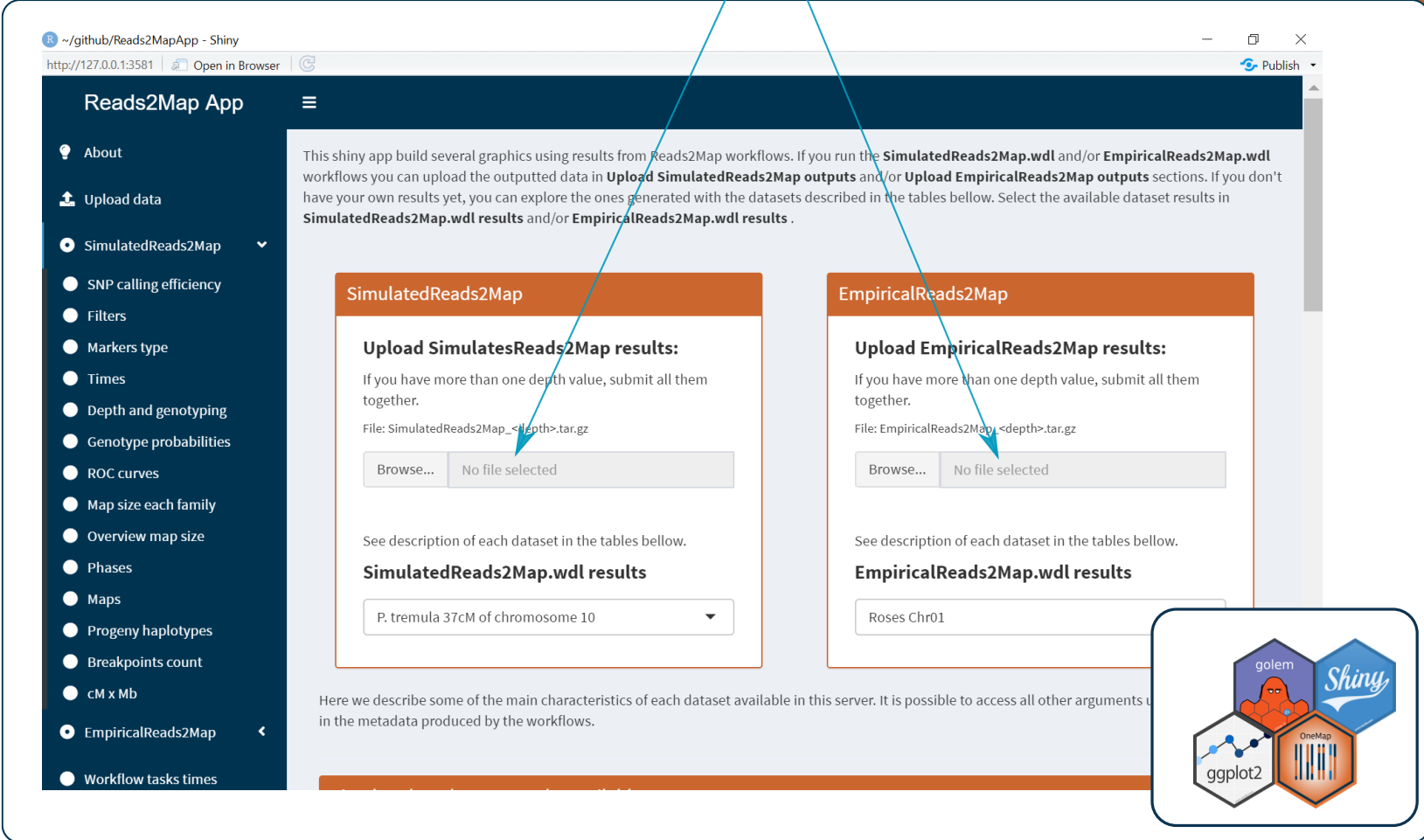
- ▶ Containers
- ▶ High Performance Computing (HPC) or Cloud environments (terra.bio)

```
$ java -jar /path/to/cromwell.jar run -i inputs/EmpiricalSNPCalling.inputs.json EmpiricalSNPCalling.wdl
```



Implementation

► Visualization and exploration



The screenshot displays the Reads2Map App interface. At the top, a brown box labeled "tar.gz" is connected by blue lines to two upload sections: "SimulatedReads2Map" and "EmpiricalReads2Map".

SimulatedReads2Map

Upload SimulatedReads2Map results:
If you have more than one depth value, submit all them together.
File: SimulatedReads2Map_<depth>.tar.gz
Browse... No file selected

See description of each dataset in the tables below.

SimulatedReads2Map.wdl results
P. tremula 37cM of chromosome 10

EmpiricalReads2Map

Upload EmpiricalReads2Map results:
If you have more than one depth value, submit all them together.
File: EmpiricalReads2Map_<depth>.tar.gz
Browse... No file selected

See description of each dataset in the tables below.

EmpiricalReads2Map.wdl results
Roses Chr01

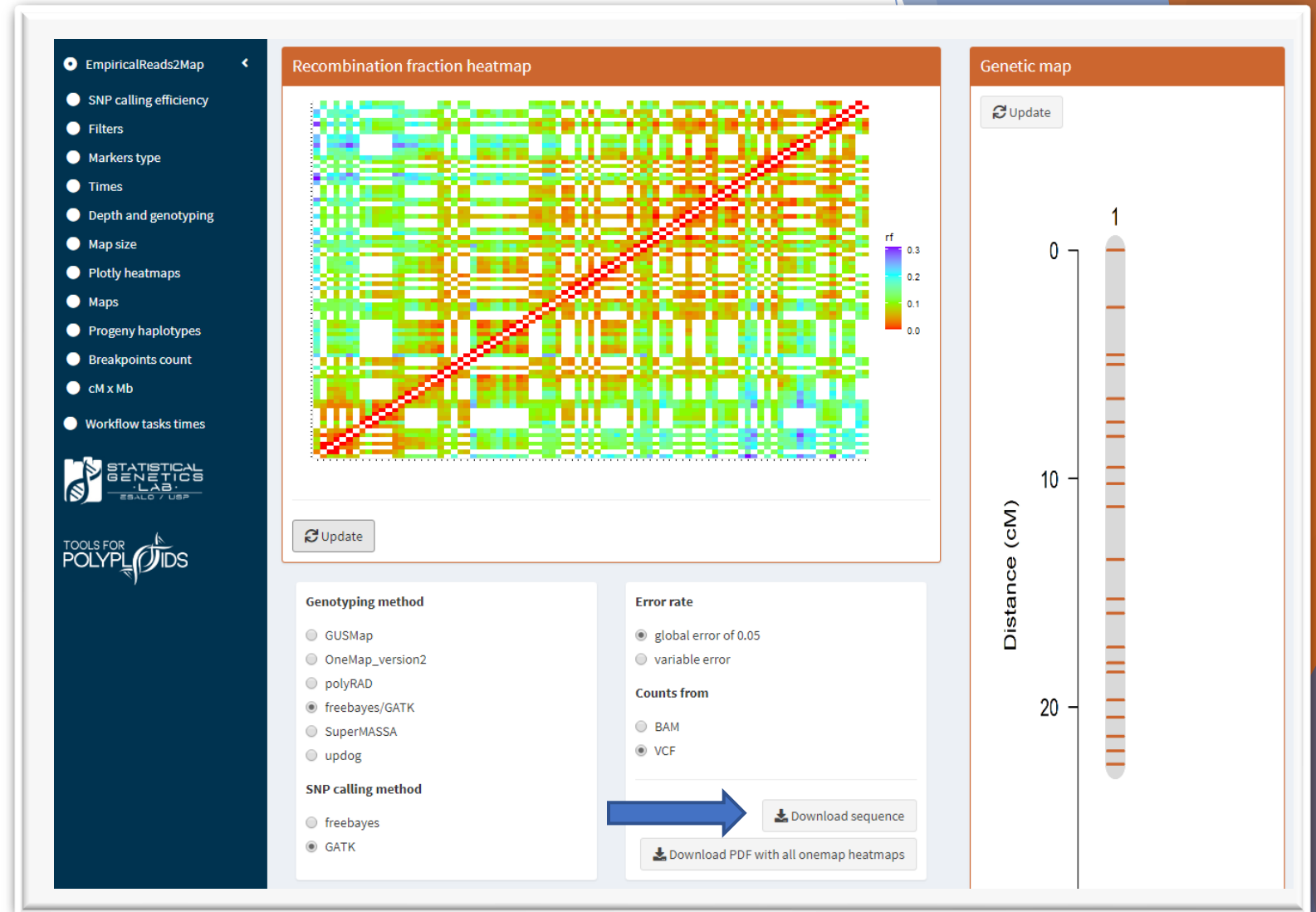
Here we describe some of the main characteristics of each dataset available in this server. It is possible to access all other arguments in the metadata produced by the workflows.

The interface includes a sidebar with navigation options: About, Upload data, SimulatedReads2Map (expanded), SNP calling efficiency, Filters, Markers type, Times, Depth and genotyping, Genotype probabilities, ROC curves, Map size each family, Overview map size, Phases, Maps, Progeny haplotypes, Breakpoints count, cM x Mb, EmpiricalReads2Map, and Workflow tasks times.

Logos for golem, Shiny, ggplot2, and OneMap are visible in the bottom right corner.

Example results -Diploids

- ▶ Outputted maps:
 - ▶ Empirical: 34
 - ▶ Simulated: 68
- ▶ Test only a subset of one group and repeat the pipeline to others



Tutorial

- ▶ Step-by-step of SNP and dosage calling using BWA and GATK
- ▶ https://bit.ly/GVENCKpoly_GATK

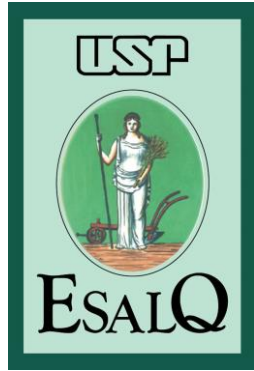
References

- ▶ Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A.; Cresko, W. A. (2013). Stacks: an analysis tool set for population genomics. *Molecular Ecology*, 22(11), 3124-3140. <https://doi.org/10.1111/mec.12354>
- ▶ Glaubitz, J. C., Casstevens, T. M., Lu, F., Harriman, J., Elshire, R. J., Sun, Q.; Buckler, E. S. (2014). TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. *PLoS ONE*, 9(2), 1-11. <https://doi.org/10.1371/journal.pone.0090346>
- ▶ Garrison, E.; Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *ArXiv E-Prints*, 9. <https://doi.org/1207.3907>
- ▶ McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M.; DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9), 1297-1303. <https://doi.org/10.1101/gr.107524.110>

References

- ▶ Gerard, D., Ferrão, L. F. V., Garcia, A. A. F., & Stephens, M. (2018). Genotyping Polyploids from Messy Sequencing Data. *Genetics*, 210(3), 789-807. doi: 10.1534/genetics.118.301468.
- ▶ Wadl, P. A., Olukolu, B. A., Branham, S. E., Jarret, R. L., Yencho, G. C.; Jackson, D. M. (2018). Genetic Diversity and Population Structure of the USDA Sweetpotato (*Ipomoea batatas*) Germplasm Collections Using GBSpoly. *Frontiers in Plant Science*, 9, 1166. <https://doi.org/10.3389/fpls.2018.01166>
- ▶ Serang, O., Mollinari, M.; Garcia, A. A. F. (2012). Efficient exact maximum a posteriori computation for bayesian SNP genotyping in polyploids. *PLoS ONE*, 7(2), 1-13. <https://doi.org/10.1371/journal.pone.0030906>
- ▶ Clark, L. v., Lipka, A. E.; Sacks, E. J. (2019). polyRAD: Genotype Calling with Uncertainty from Sequencing Data in Polyploids and Diploids. *G3: Genes | Genomes | Genetics*, 9(March), g3.200913.2018. <https://doi.org/10.1534/g3.118.200913>

Project Members



Other funding agencies



Other Project Members



Cornell University



WISCONSIN
UNIVERSITY OF WISCONSIN-MADISON



PennState



WAGENINGEN
UNIVERSITY & RESEARCH

WASHINGTON STATE
 **UNIVERSITY**



UNIVERSITY OF MINNESOTA
Driven to Discover®



UNIVERSITY OF
ARKANSAS

1865 THE UNIVERSITY OF
 **MAINE**



Oregon State
University

Other Collaborators



Neuhouse Farms



Woolf Roses L.L.C.

