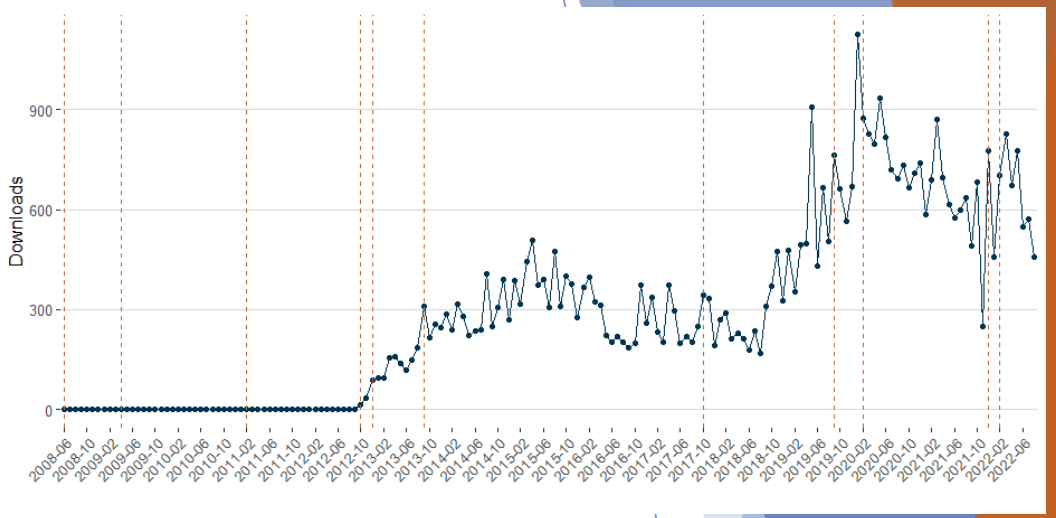
Building highly saturated genetic maps with OneMap 3.0

New approaches using workflows

Cris Taniguti

This project is funded by USDA NIFA Specialty Crop Research Initiative Award # 2020-51181-32156 (09/01/20 - 08/31/24)

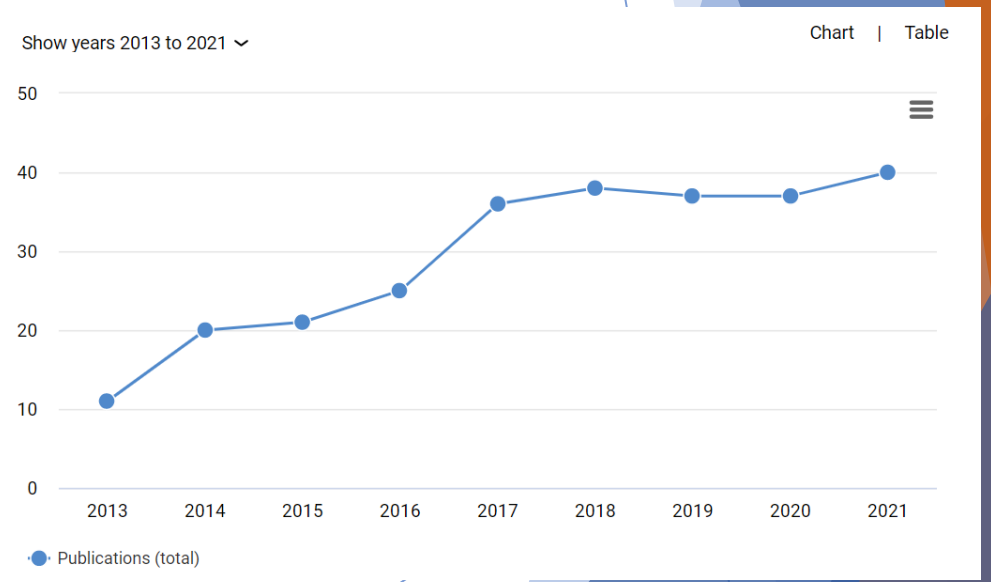
OneMap package



Margarido et al., 2007

OneMap: software for genetic mapping in outcrossing species

- R package
- Diploids
- Inbred and outcrossing populations
- Integrated genetic maps
- Tutorials
- Diagnostic graphics
- Updates by Statistical Genetics Lab members
- CRAN
- GitHub: Cristianetaniguti/onemap



TOOLS FOR POLYPLOIDS

STATISTICAL GENETICS LAB
ESALQ / USP

OneMap users

TOOLS FOR POLYPLOIDS

STATISTICAL GENETICS LAB
ESALQ / USP

OneMap updates

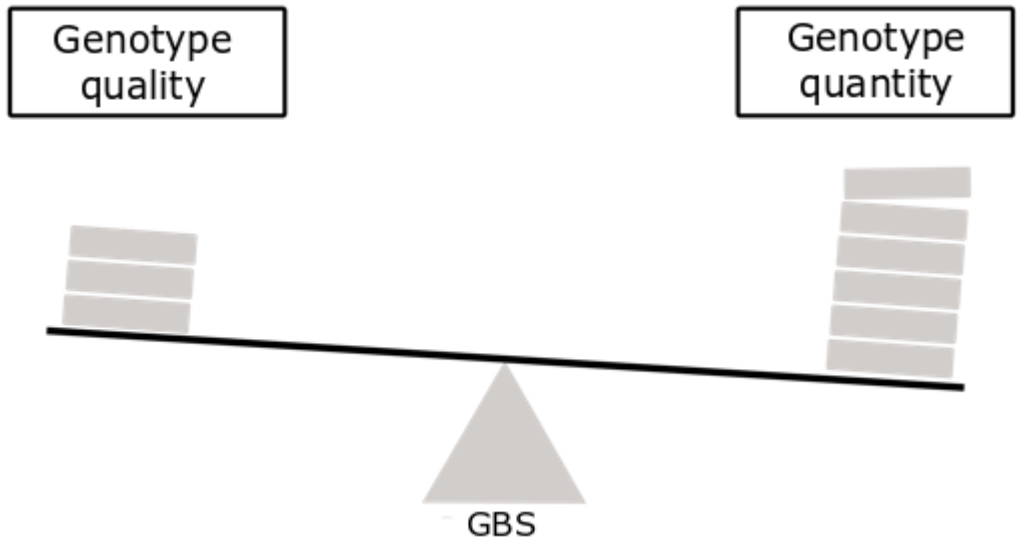
- **R language**: Didactic tutorials, courses and utilities functions
- **Maintenance**: Package “house keeping” and dependencies upgrades
- **VCF file conversion**: new features to deal with all VCF format versions
- **Group markers**: new features to consider physical position and UPGMA method
- **Markers ordering**: MDS method
- **Processing time and memory**: RAM optimization and process parallelization
- **Graphical visualization**: sequencing depth, genotypes, segregation, haplotypes and recombination breakpoints counts

TOOLS FOR POLYPLOIDS

STATISTICAL GENETICS LAB · ESALQ / USP

Motivation

- Sequencing based markers
- Characteristics:
  - Thousands of markers
  - Different types of libraries
  - Lower cost
  - Genotyping errors
  - Biallelic markers
- Consequences:
  - Requires bioinformatic skills
  - Requires computational resources
  - Wrong number of recombination events
  - Difficulties in ordering markers

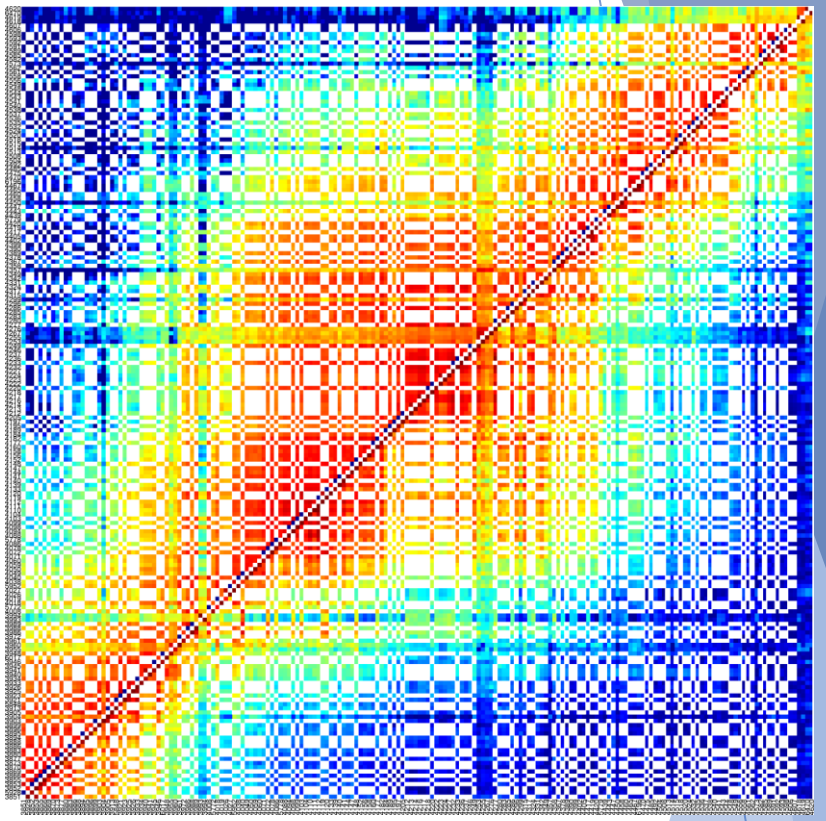


TOOLS FOR POLYPLOIDS

STATISTICAL GENETICS LAB · ESALQ / USP

OneMap updates - HMM

- Genotyping errors
- Hidden Markov Model (HMM)
- create_probs(global_error = 0.05)
- create_probs(genotype_prob =
  prob_data.frame)
- create_probs(genotype_error =
  prob_matrix)



833 cM

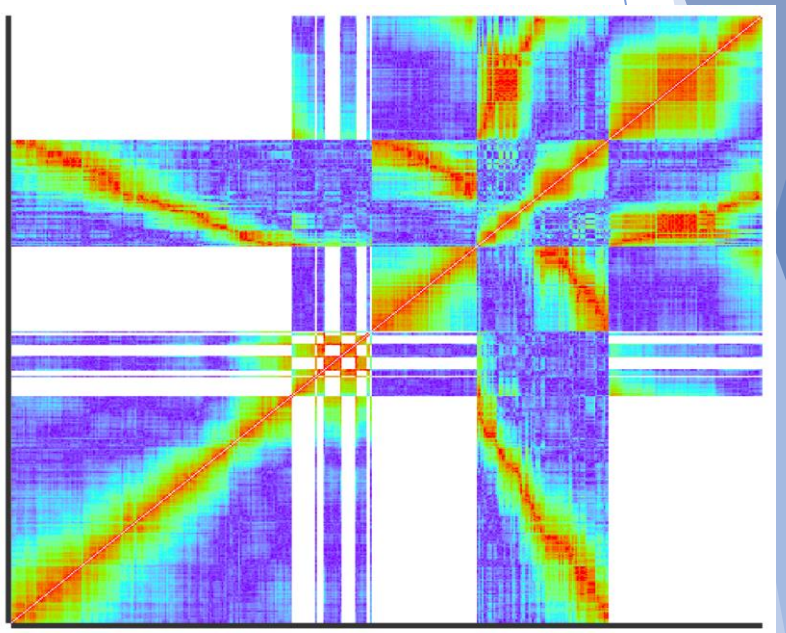
TOOLS FOR POLYPLOIDS

STATISTICAL GENETICS LAB · ESALQ / USP

OneMap updates - Multiallelics

- Difficulties in ordering
- Biallelics marker types

  ab x ab
  aa x ab
  ab x aa



- Haplotype-based markers
- onemap_read_vcfR(only_biallelics = FALSE)

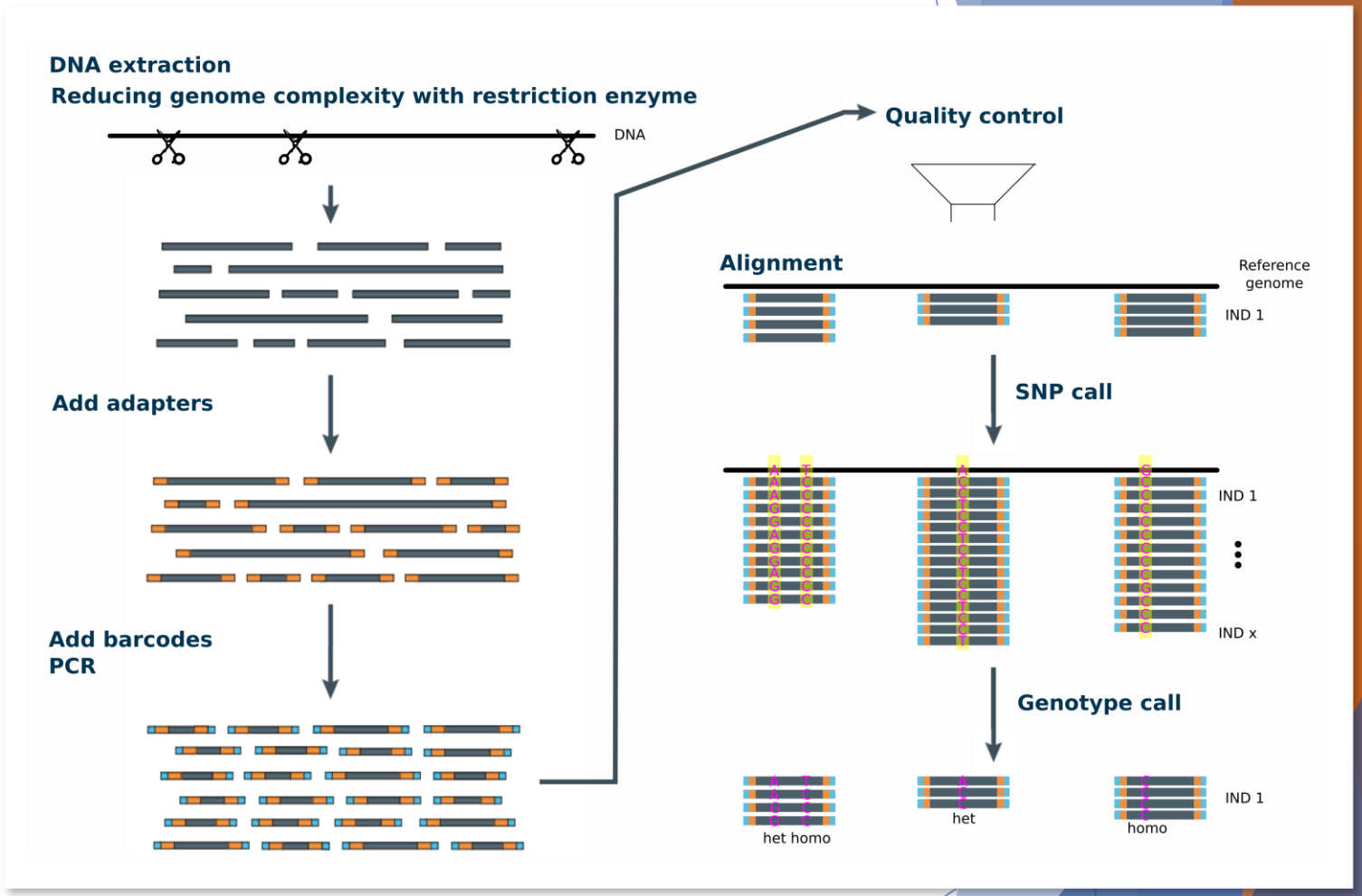
  ab x cd

TOOLS FOR POLYPLOIDS

STATISTICAL GENETICS LAB
ESALQ / USP

Motivation

- Genotyping-by-Sequencing (GBS) markers
- Genetic map quality is related to upstream processes



TOOLS FOR POLYPLOIDS

STATISTICAL GENETICS LAB · ESALQ / USP

Motivation - Polyploid species

- Organisms that have multiple copies of the complete set of chromosomes

MAPpoly

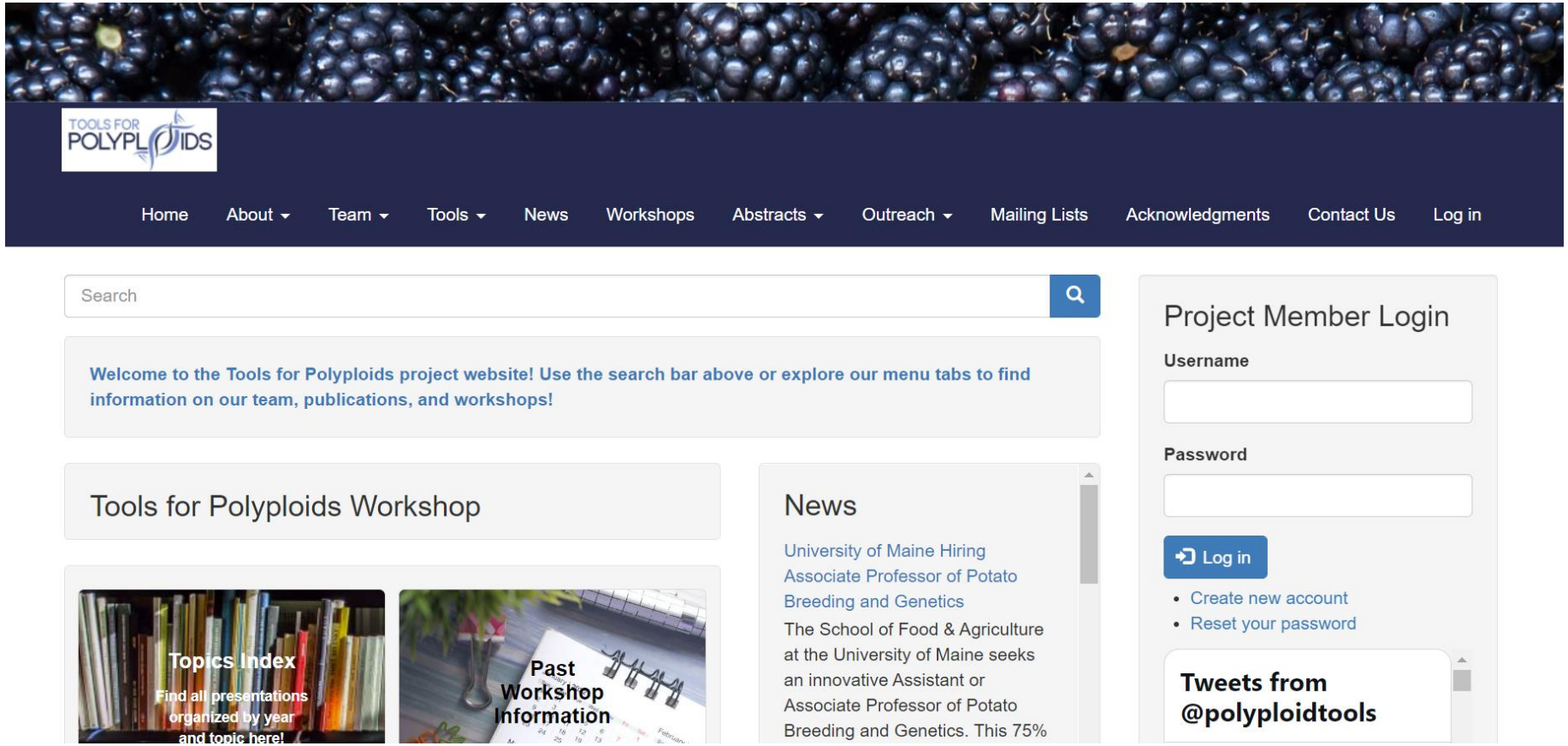
Mollinari and Garcia, 2019
Linkage Analysis and Haplotype Phasing in Experimental Autopolyploid Populations with High Ploidy Level Using Hidden Markov Models

TOOLS FOR POLYPLOIDS

STATISTICAL GENETICS LAB · ESALQ / USP

polyploids.org

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

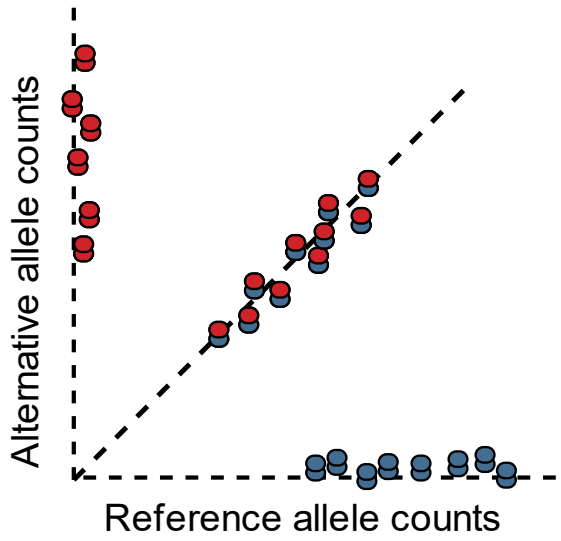


This project is funded by USDA NIFA Specialty Crop Research Initiative Award # 2020-51181-32156 (09/01/20 - 08/31/24)

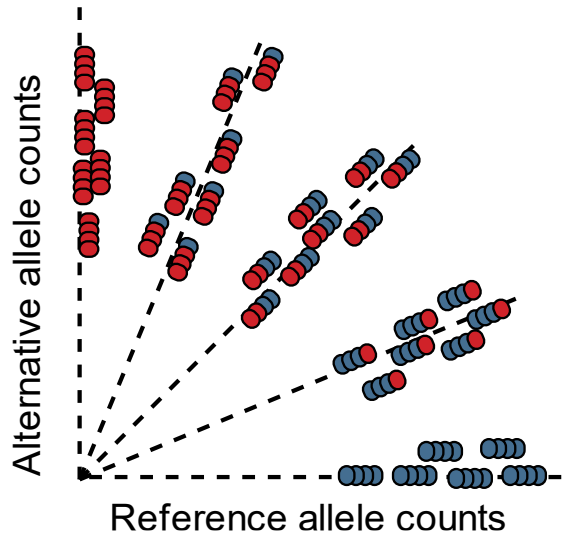
Dosage calling

- The theory

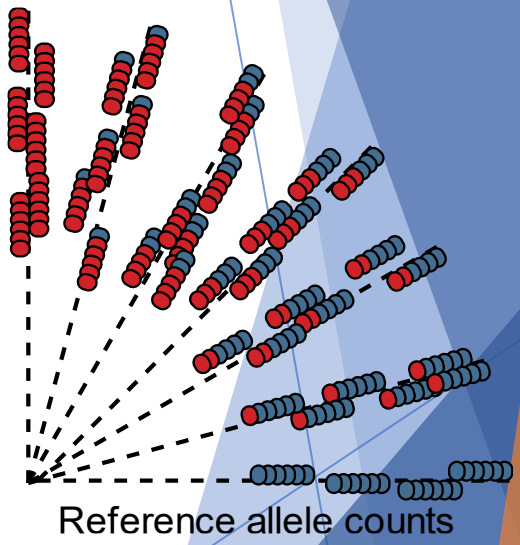
Diploid



Tetraploid



Hexaploid



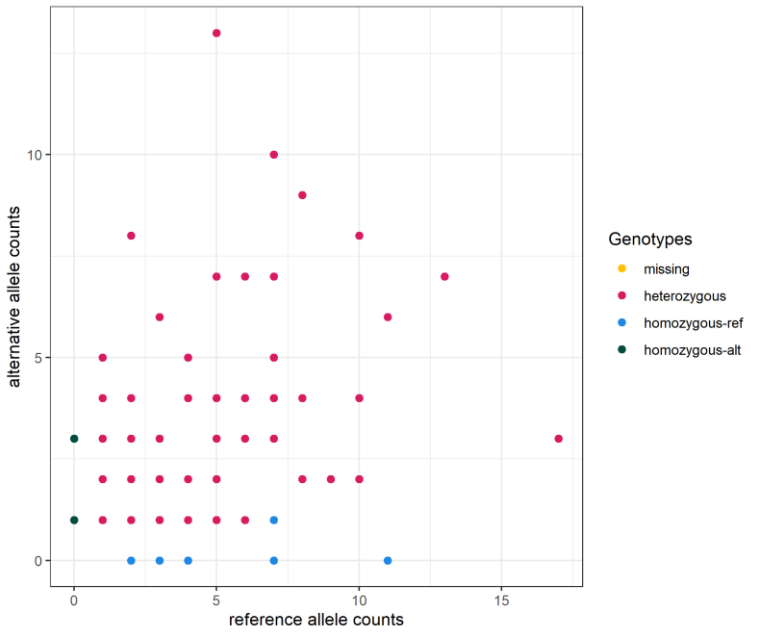
TOOLS FOR POLYPLOIDS

STATISTICAL GENETICS LAB · ESALQ / USP

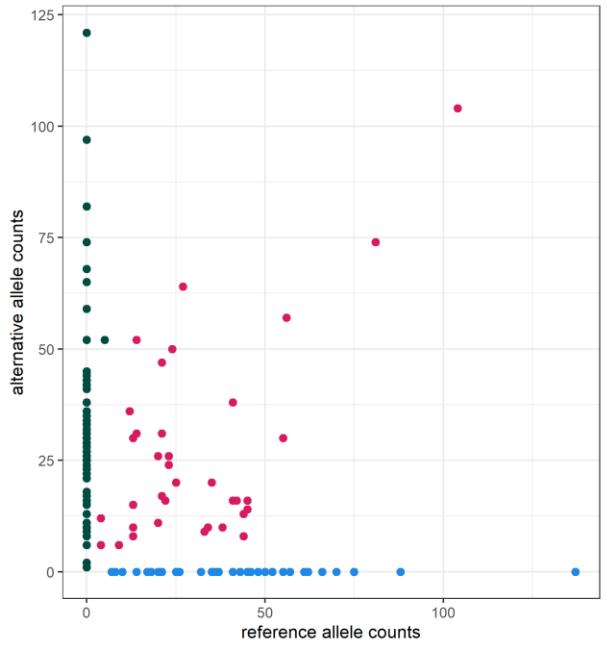
Source of errors

- The reality

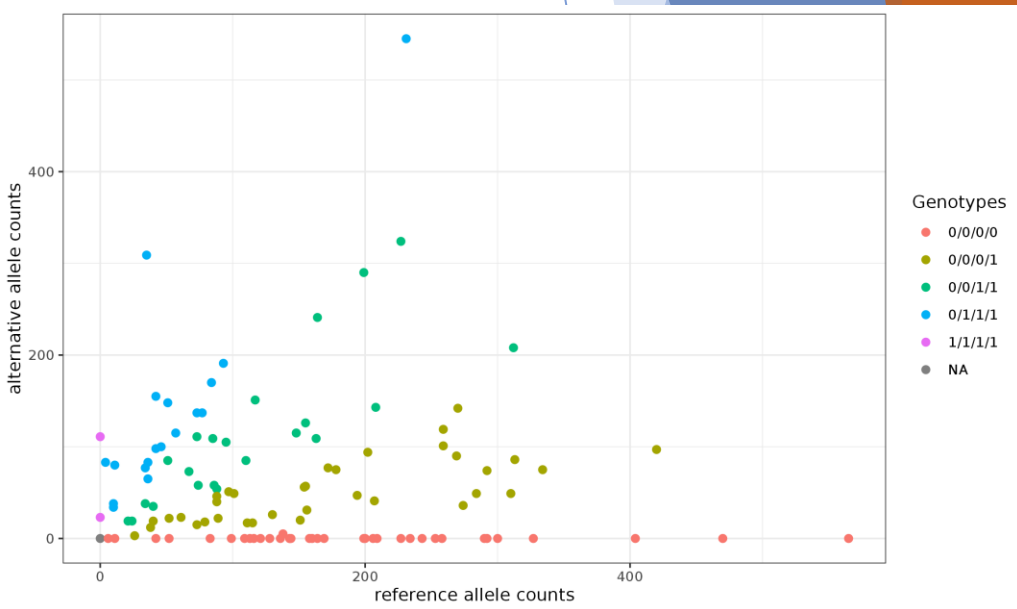
Diploid (mean depth 6)
N = 200
Aa x Aa



Diploid (mean depth 96)
N = 138
Aa x Aa



Tetraploid (mean depth 83)
N = 114
AAaa x AAaa



TOOLS FOR POLYPLOIDS

STATISTICAL GENETICS LAB
ESALQ / USP

Dosage calling

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

TOOLS FOR POLYPLOIDS

STATISTICAL GENETICS LAB · ESALQ / USP

Dosage calling

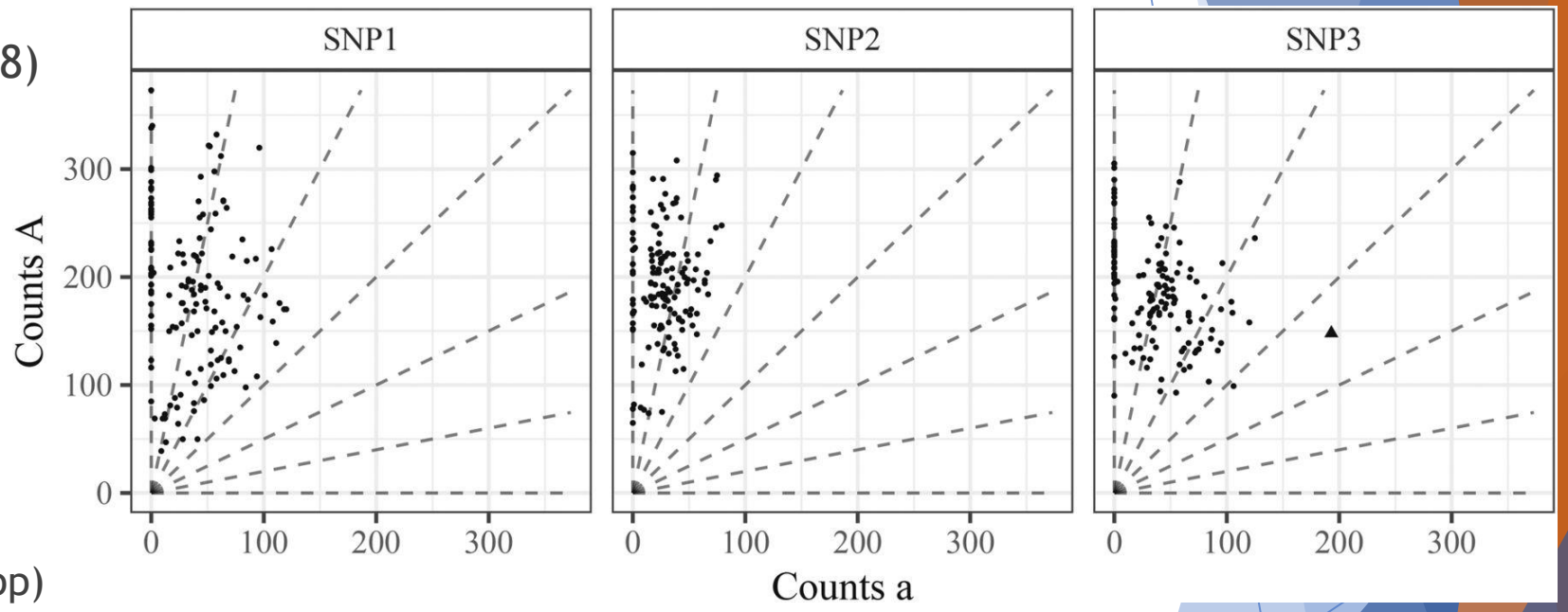
- updog (Gerard et al., 2018)
  - Any ploidy
  - Allelic bias
  - Overdispersion
  - Sequencing errors
  - Outliers
  - Population structure

    (F1, S1, HW, F1pp, S1pp)

Overdispersion

Allelic bias

Outlying observations



Gerard et al., 2018

TOOLS FOR POLYPLOIDS

STATISTICAL GENETICS LAB · ESALQ / USP

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)

TOOLS FOR POLYPLOIDS

STATISTICAL GENETICS LAB · ESALQ / USP

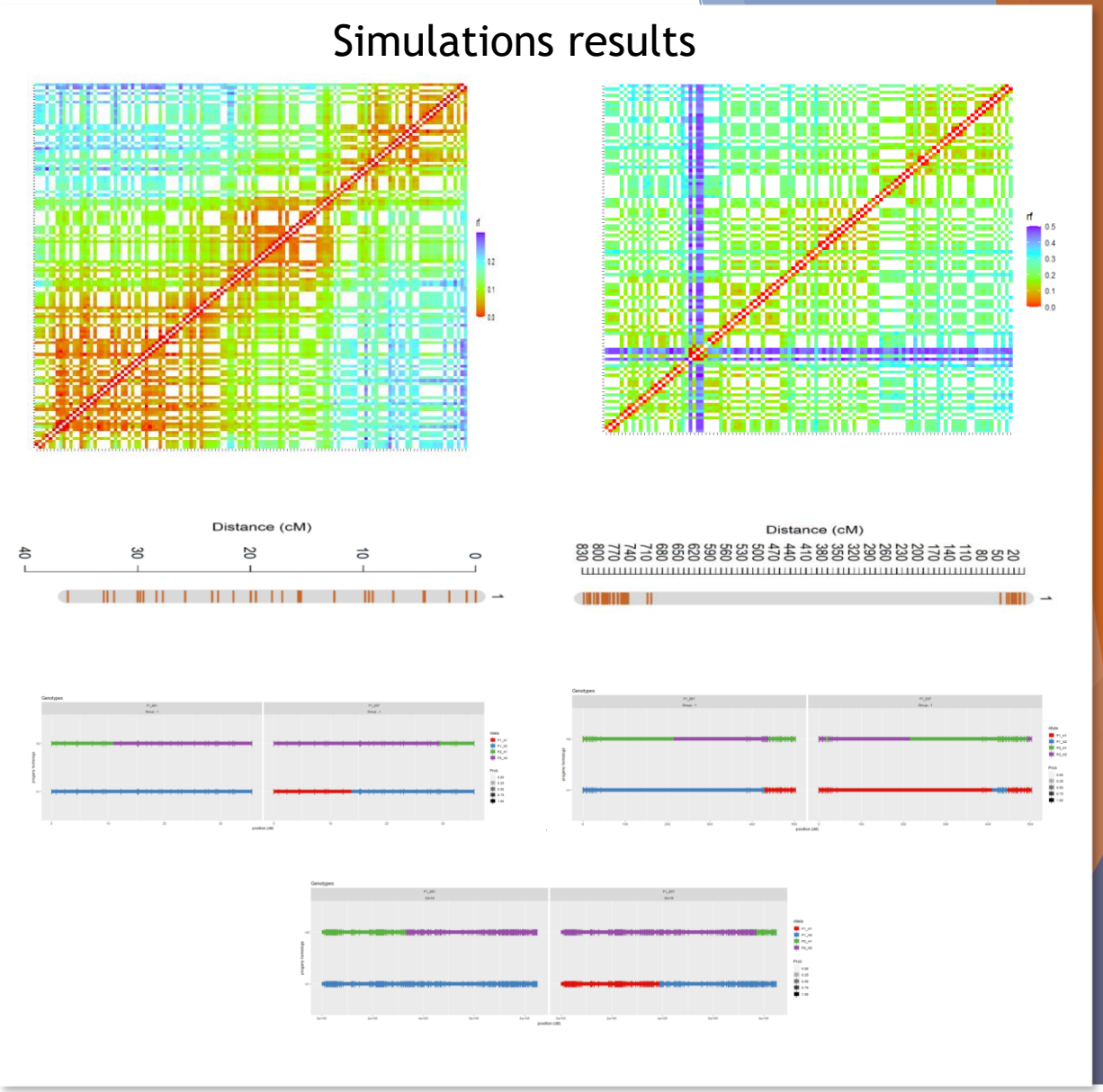
Which is the best pipeline?

- Challenges:
  - Many software, many dependencies
  - Different input and output formats
  - Computational resources
  - Explore and visualize results
  - Feedback for developers
  - Adapt to software updates
  - Reproducibility
  - Quality criteria
- Useful tools:
  - Containers (Docker and singularity)
  - Workflow Description Language (WDL)
  - GitHub
  - HPC and Google Cloud
  - Shiny
  - Linkage map

TOOLS FOR POLYPLOIDS

STATISTICAL GENETICS LAB · ESALQ / USP

Maps as benchmarks

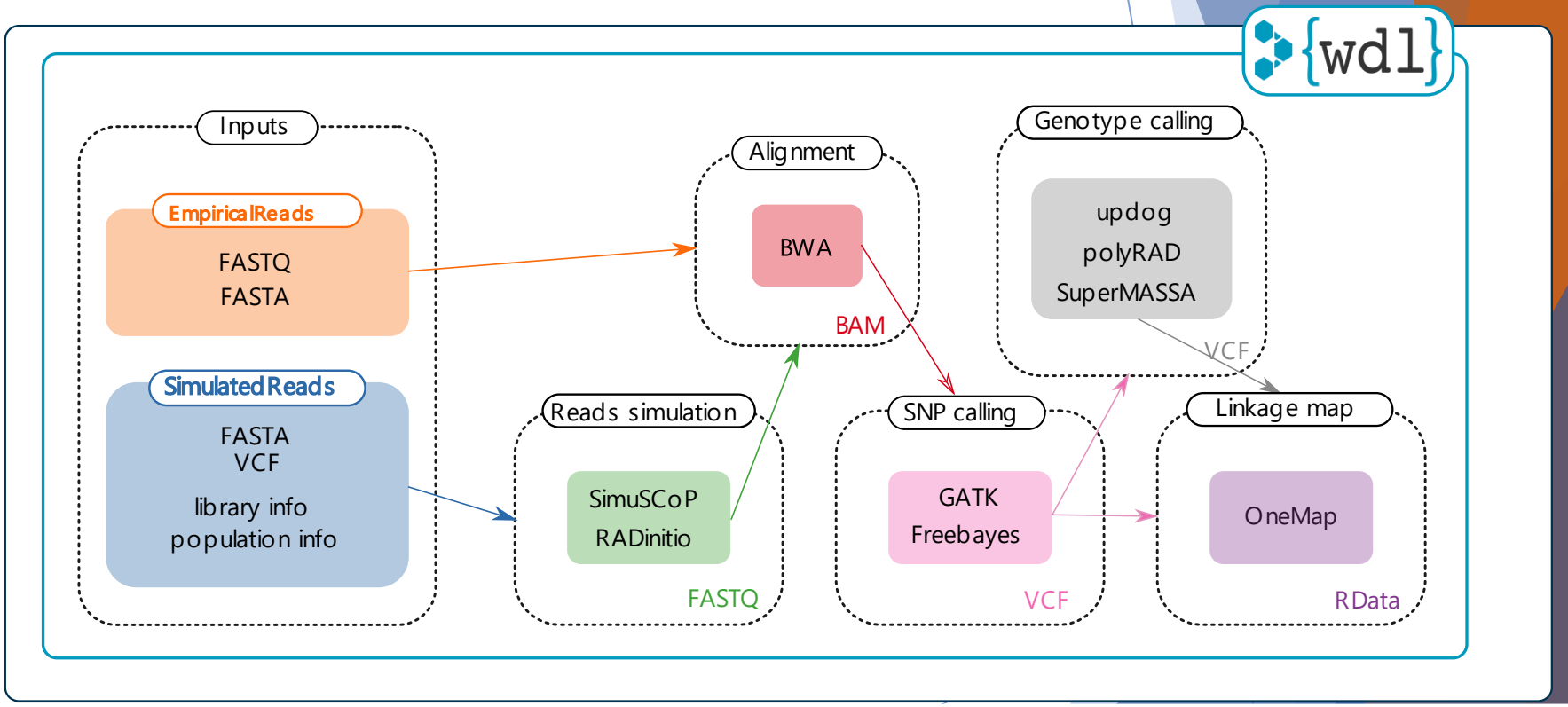


TOOLS FOR POLYPLOIDS

STATISTICAL GENETICS LAB
ESALQ / USP

Implementation

- Join several bioinformatics and statistical analyses
- Workflow Description Language (WDL - “widdle”) – Broad Institute (human genome research)
- Best practices
- EmpiricalReads2Map
- SimulatedReads2Map
- Diploid species

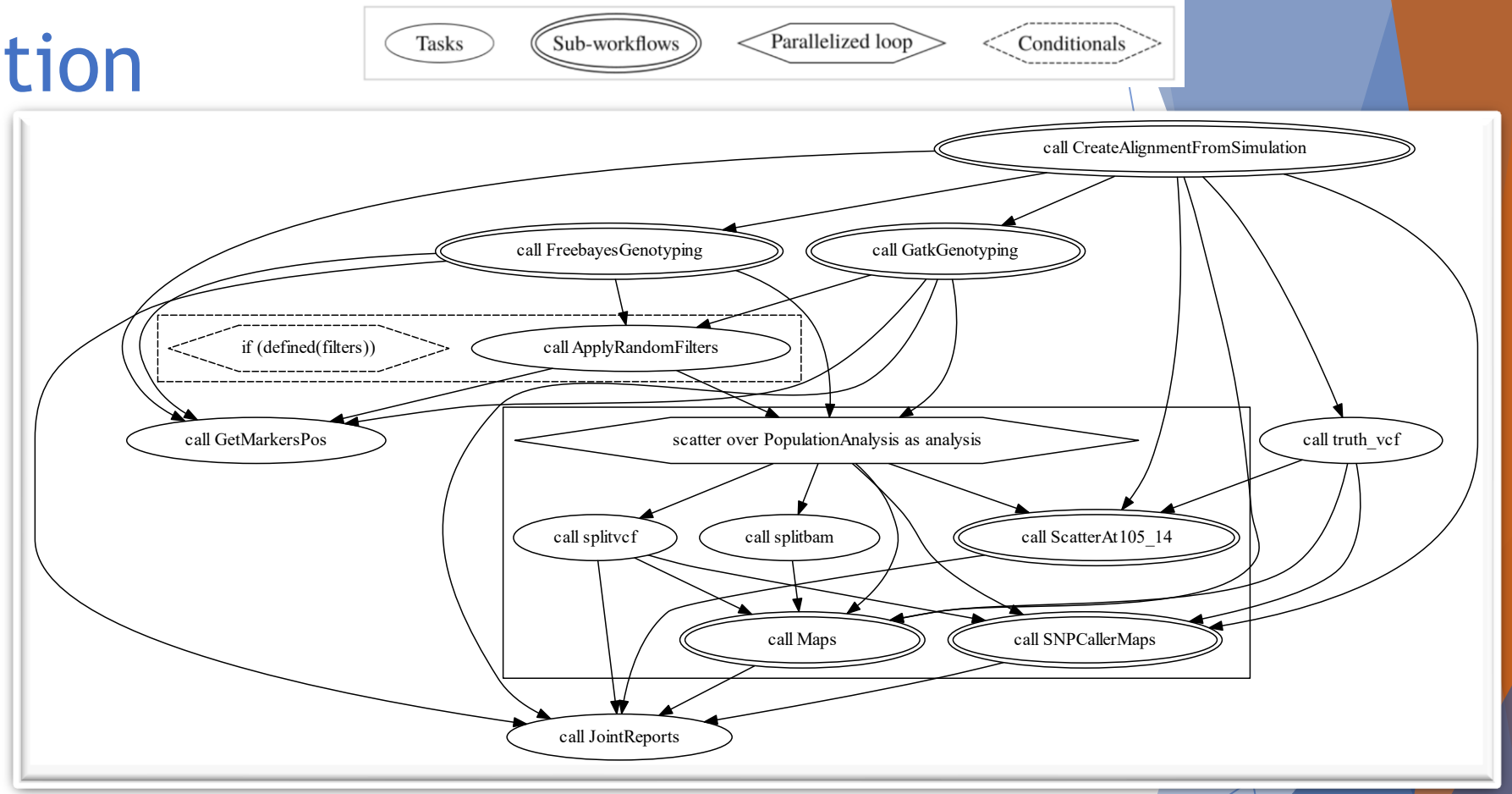


TOOLS FOR POLYPLOIDS

STATISTICAL GENETICS LAB · ESALQ / USP

Implementation

- Workflows
  - Sub-workflows
    - Tasks




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

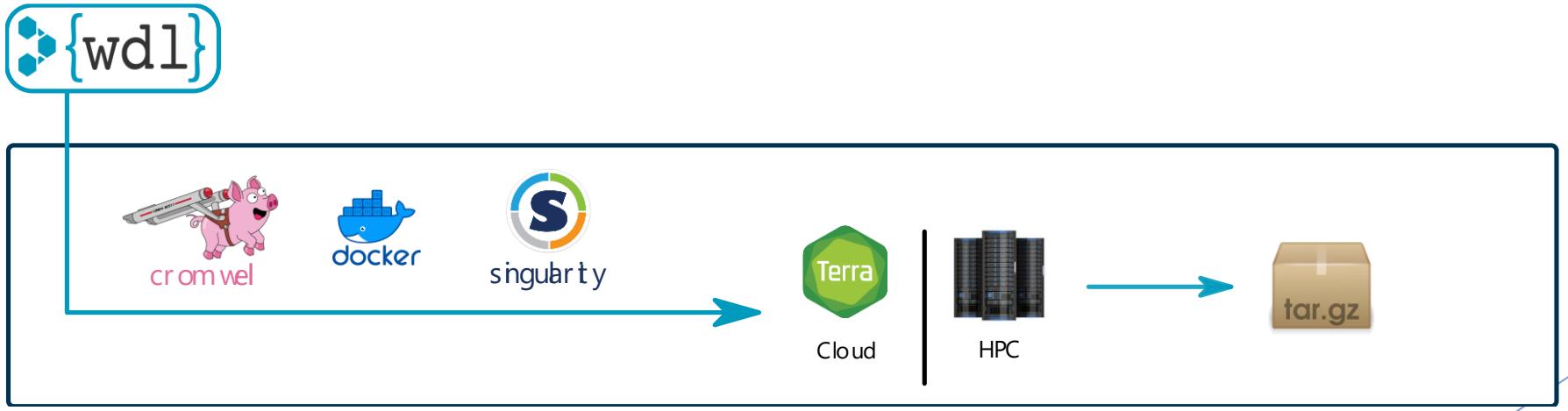
TOOLS FOR POLYPLOIDS

STATISTICAL GENETICS LAB · ESALQ / USP

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
```



TOOLS FOR POLYPLOIDS

STATISTICAL GENETICS LAB
ESALQ / USP

More about WDL and usage in Cloud

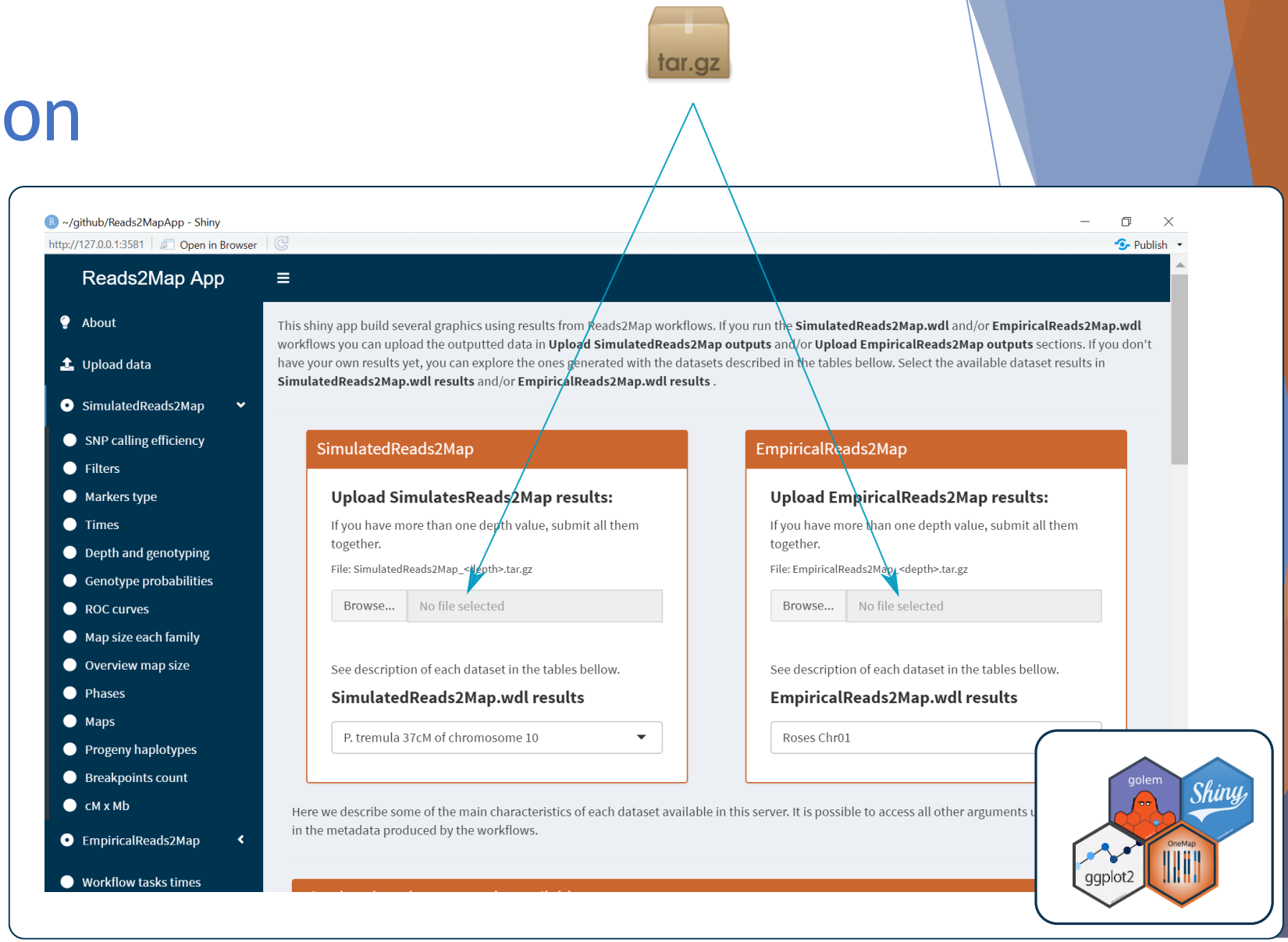
Streaming live on YouTube
https://youtu.be/3AMJ-LlWRtE

TOOLS FOR POLYPLOIDS

STATISTICAL GENETICS LAB
ESALQ / USP

Implementation

- Visualization and exploration



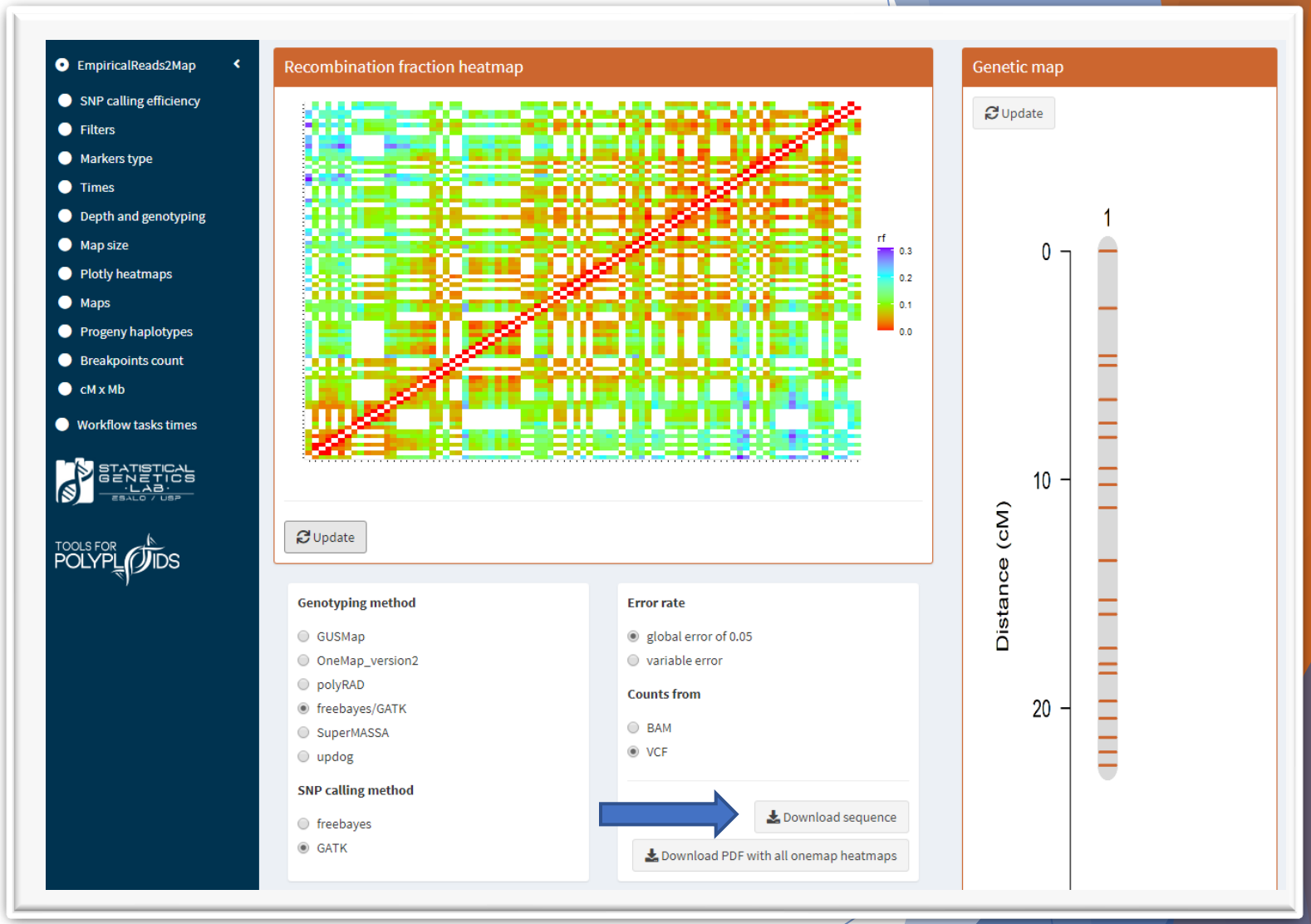
TOOLS FOR POLYPLOIDS

STATISTICAL GENETICS LAB
ESALQ / USP

Example results -Diploids

- Outputted maps:
  - Empirical: 34
  - Simulated: 68

- Test only a subset of one group and repeat the pipeline to others



Results for different scenarios

- Playing with parameters:
  - 8160 simulated maps
  - 816 empirical maps

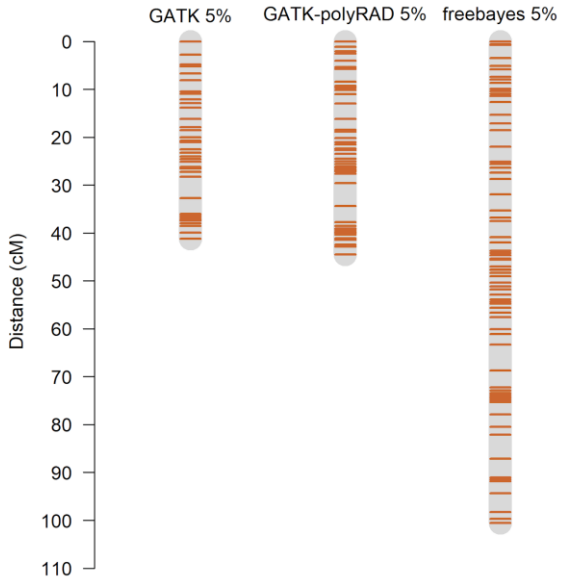
- Effect of filters
- Effect of multiallelic markers
- Effect of contaminants
- Effect of segregation distortion
- Comparison with GUSMap (Bilton et al, 2018)
- Select best pipeline

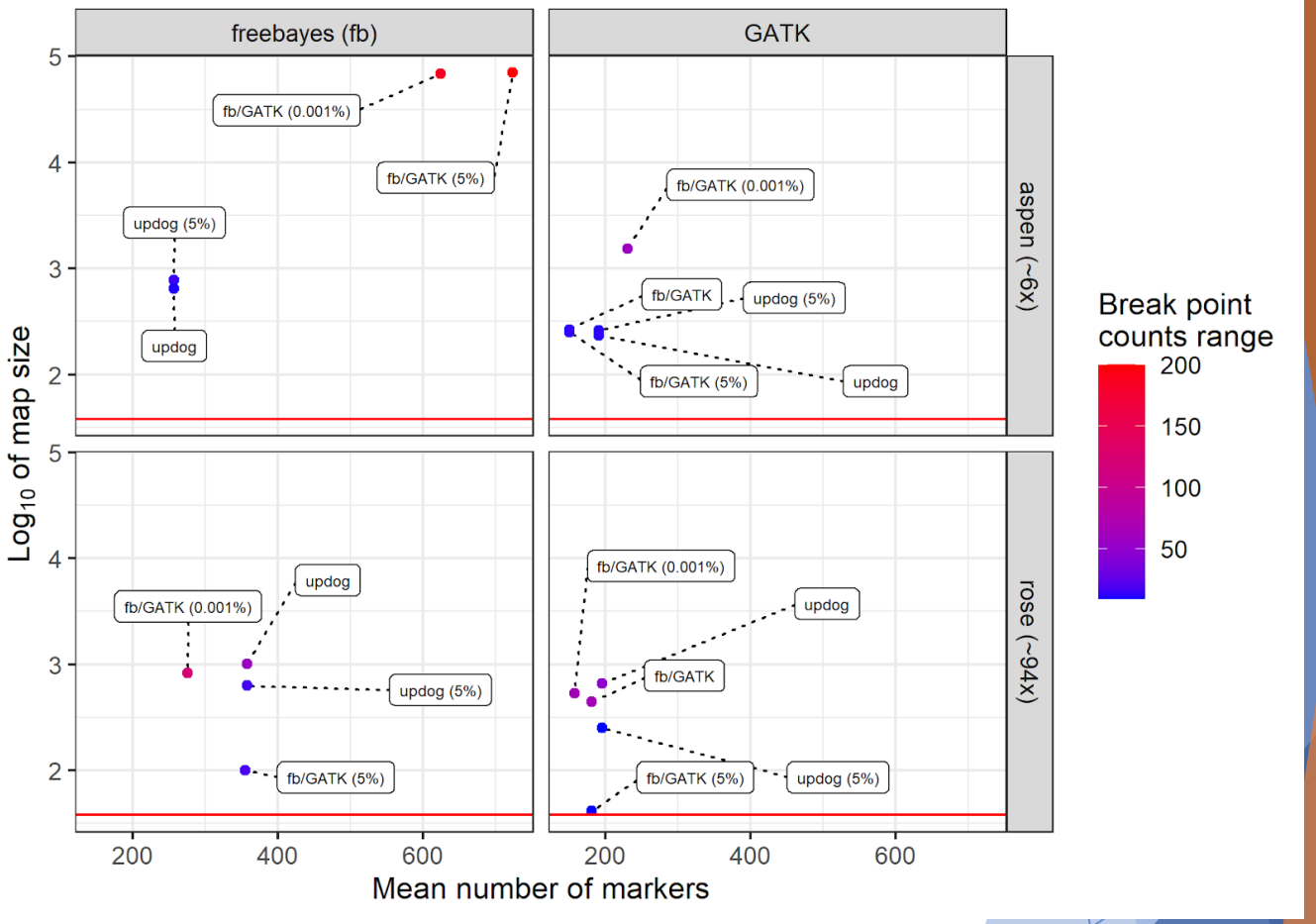
TOOLS FOR POLYPLOIDS

STATISTICAL GENETICS LAB
ESALQ / USP

Results for different scenarios

- Selecting best





TOOLS FOR POLYPLOIDS

STATISTICAL GENETICS LAB · ESALQ / USP

Thanks!

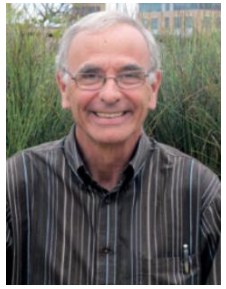
- Many people
- Authors:

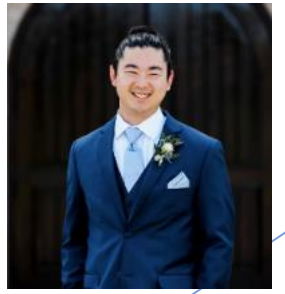
Cristiane Taniguti
Lucas Taniguti
Gabriel Gesteira
Thiago Oliveira
Jeekin Lau
Getúlio Ferreira
Rodrigo Amadeu
David Byrne
Oscar Riera-Lizarazu
Guilherme Pereira
Marcelo Mollinari
Augusto Garcia












Contact: chtaniguti@tamu.edu

Project Members

Other funding agencies

This project is funded by USDA NIFA Specialty Crop Research Initiative Award # 2020-51181-32156 (09/01/20 - 08/31/24)

Other Project Members

Plant & Food Research
Rangahau Ahumāra Kai

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

TOOLS FOR POLYPLOIDS

STATISTICAL GENETICS LAB
ESALQ / USP

This project is funded by USDA NIFA Specialty Crop Research Initiative Award # 2020-51181-32156 (09/01/20 - 08/31/24)

Other Collaborators

Genetwister

DAICEL arbor biosciences

VacCAP

Neuhouse Farms

CAPES

bejo

CIAT

Spring Meadow

Royal Van Zanten

American Garden Rose Selections™

Plant Sciences PSI Inc. Advancing Agriculture through Science

IITA Transforming African Agriculture

FAPESP

CIP INTERNATIONAL POTATO CENTER

Potatoes USA

Woolf Roses L.L.C.

star ROSES and PLANTS

Dekker chrysanten

BAILEY

CSS FARMS

Weeks Roses

DÜMMEN ORANGE

illumina

PACIFIC BERRY BREEDING

DELIFLOR Chrysanten

CNPq Conselho Nacional de Desenvolvimento Científico e Tecnológico

Driscoll's

TOOLS FOR POLYPLOIDS

STATISTICAL GENETICS LAB · ESALQ / USP

This project is funded by USDA NIFA Specialty Crop Research Initiative Award # 2020-51181-32156 (09/01/20 - 08/31/24)

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

TOOLS FOR POLYPLOIDS

STATISTICAL GENETICS LAB · ESALQ / USP

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

TOOLS FOR POLYPLOIDS

STATISTICAL GENETICS LAB · ESALQ / USP