

Genomic and Epigenomic Patterns in Novel Heterotic Pools of Winter Rapeseed (*Brassica napus*)



Jenny HueyTyng Lee
Plant Breeding Department
Justus Liebig University, Giessen

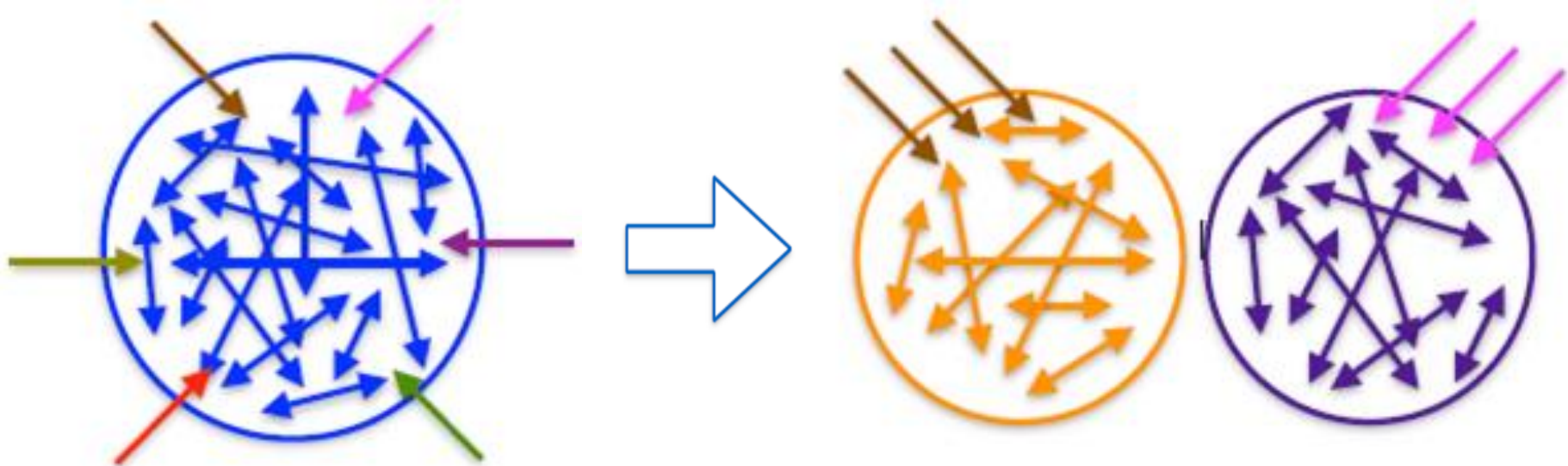


IRC Berlin 2019

Photo credit: Annaliese Mason

Heterosis is not fully exploited in oilseed rape

- Hybrid breeding is one of the landmark contributors to crop yield increase
- Heterosis is when a hybrid has phenotypic superiority over parents in particular traits
- Can we optimise hybrid breeding in oilseed rape?

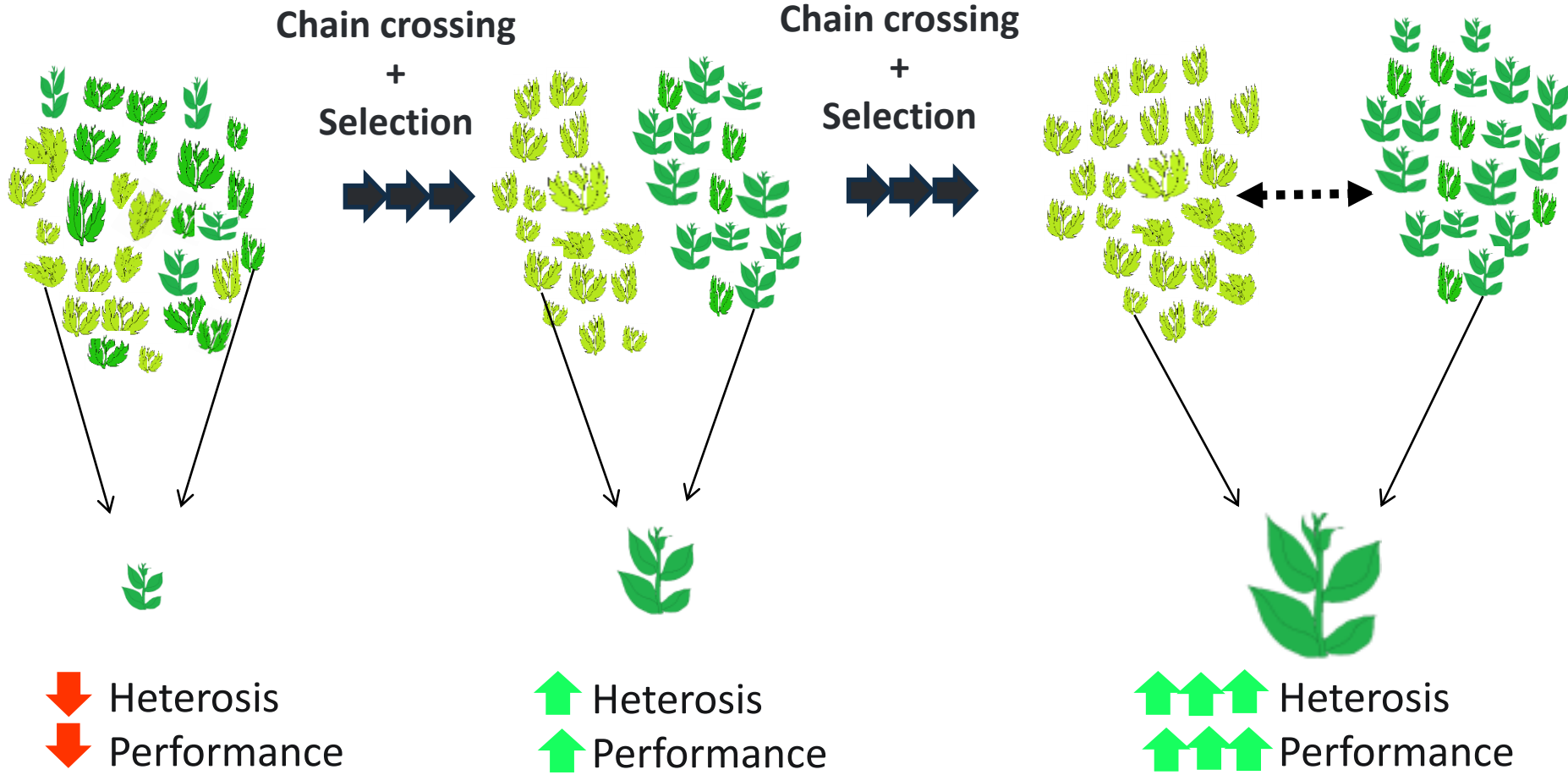


One common breeding pool
Diversity low
LD conservation high

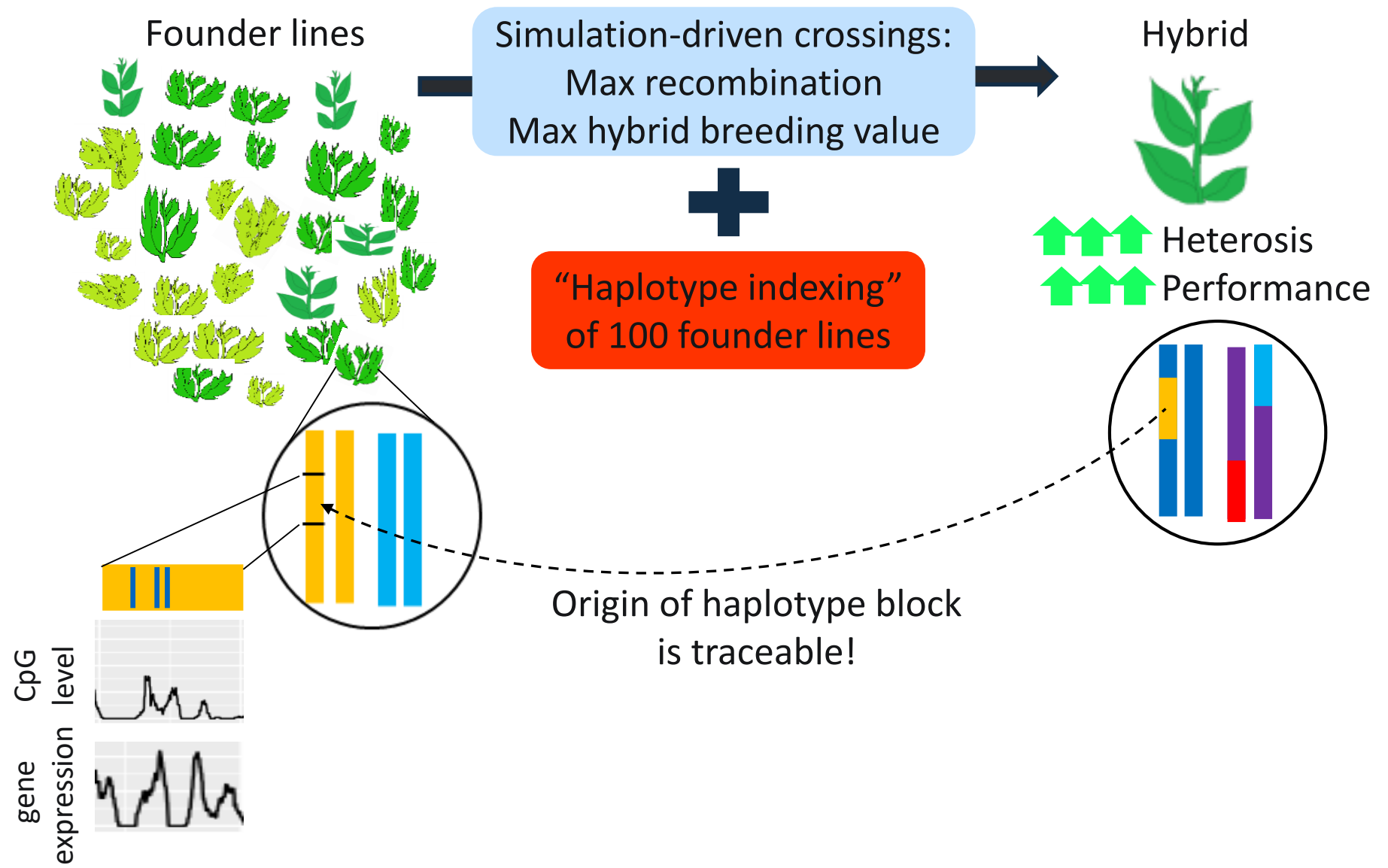
Separate heterotic pools
Better hybrid vigour

Rapidly create heterotic pools

2x50 elite
founder lines



Perfect dataset to develop novel prediction models 4



April 2018

June 2019, bioRxiv

Trends in Genetics

CellPress
REVIEWS

Inferring the landscape of recombination using recurrent neural networks

Review

Supervised Machine Learning for Population Genetics: A New Paradigm

Daniel R. Schrider^{1,*} and Andrew D. Kern^{1,*}

Jeffrey R. Adrion^{1,†}, Jared G. Galloway^{1,†}, Andrew D. Kern¹

¹Institute of Ecology and Evolution, University of Oregon

As population genomic datasets grow in size, researchers are faced with the daunting task of making sense of a flood of information. To keep pace with this explosion of data, computational methodologies for population genetic inference are rapidly being developed to best utilize genomic sequence data. In this review we discuss a new paradigm that has emerged in computational population genomics: that of supervised machine learning (ML). We review the fundamentals of ML, discuss recent applications of supervised ML to population genetics that outperform competing methods, and describe promising future directions in this area. Ultimately, we argue that supervised ML is an important and underutilized tool that has considerable potential for the world of evolutionary genomics.

Highlights

ML methods are powerful approaches that have revolutionized many fields, but their use in population genetics inference is only beginning.

These methods are able to take advantage of high dimensional input – an important asset for population genetics inference – and are often more robust than other statistical approaches.

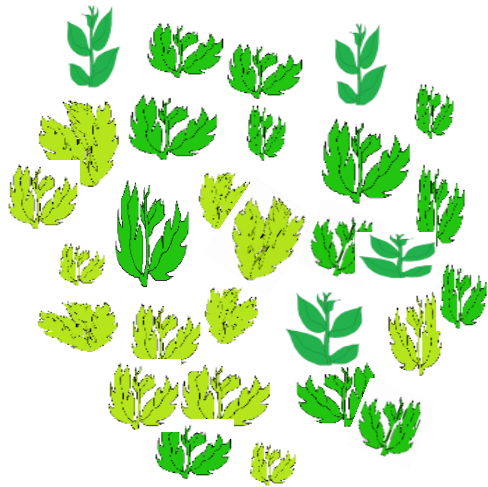
The early applications of ML to population genetics demonstrate that they outperform traditional approaches.

In this review we introduce ML to a biology audience, discuss examples of their application to evolutionary and population genetics, and lay out future directions that we view as promising.

Machine Learning for Population Genetics

Population genetics over the past 50 years has been squarely focused on reconciling molecular genetic data with theoretical models that describe patterns of variation produced by a combination of evolutionary forces. This interplay between empiricism and theory means that many advances in the field have come from the introduction of new stochastic population genetic models, often of increasing complexity, that describe how population parameters (e.g., recombination or mutation rates) might generate specific features of genetic polymorphism (e.g., the **site frequency spectrum**, SFS; see [Glossary](#)). The goal, broadly stated, is to formulate a model that describes how nature will produce patterns of variation that we observe.

Abstract Accurately inferring the genome-wide landscape of recombination rates in natural populations is a central aim in genomics, as patterns of linkage influence everything from genetic mapping to understanding evolutionary history. Here we describe ReLERNN, a deep learning method for accurately estimating a genome-wide recombination landscape using as few as four samples. Rather than use summaries of linkage disequilibrium as its input, ReLERNN considers columns from a genotype alignment, which are then modeled as a sequence across the genome using a recurrent neural network. We demonstrate that ReLERNN improves accuracy and reduces bias relative to existing methods and maintains high accuracy in the face of demographic model misspecification. We apply ReLERNN to natural populations of African *Drosophila melanogaster* and show that genome-wide recombination landscapes, while largely correlated among populations, exhibit important population-specific differences. Lastly, we connect the inferred patterns of recombination with the frequencies of major inversions segregating in natural *Drosophila* populations.



**2x50
elite founder lines**

15X whole genome
sequencing

20X bisulphite
sequencing

mRNA sequencing

miRNA sequencing

Genome wide SNPs

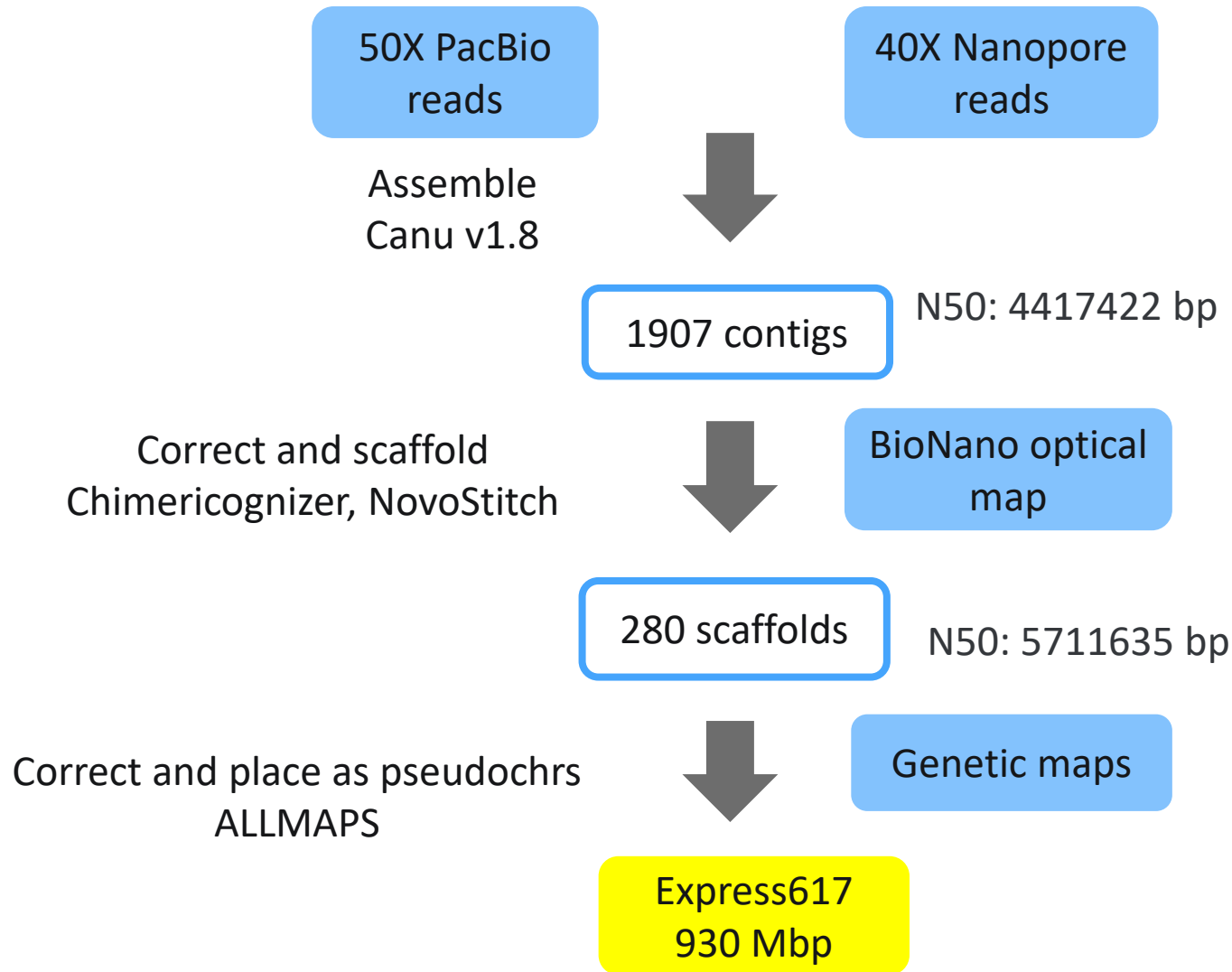
Genome wide
methylation patterns

Protein-coding gene
expression

smallRNA expression

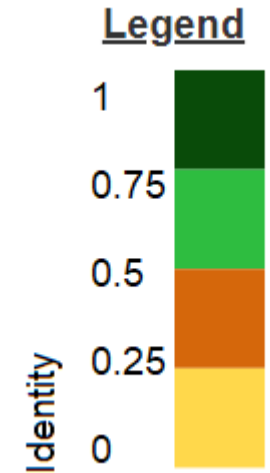
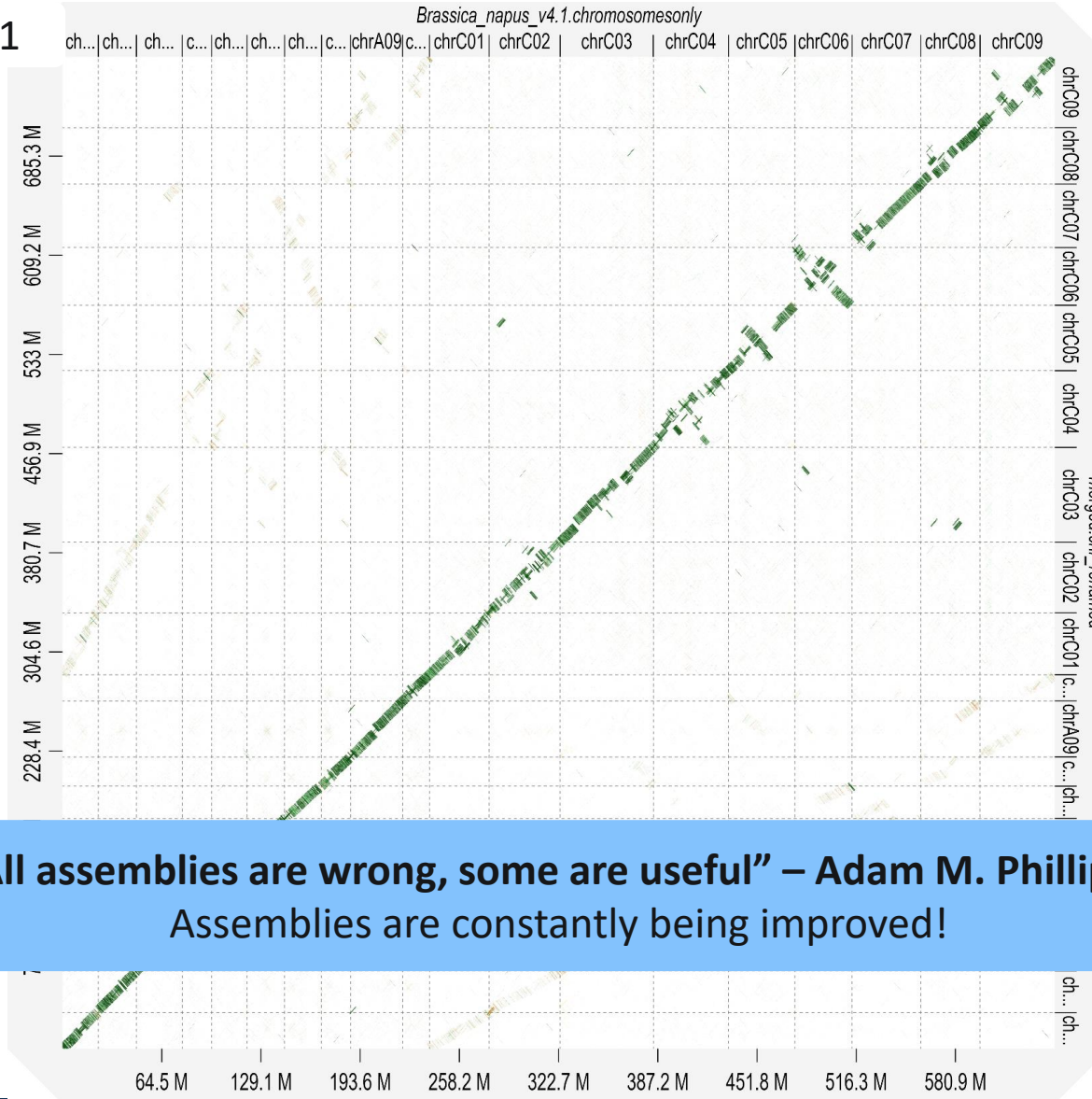
Construction of Express617 reference genome

7



The Express617 genome assembly

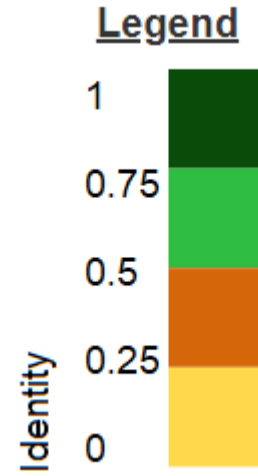
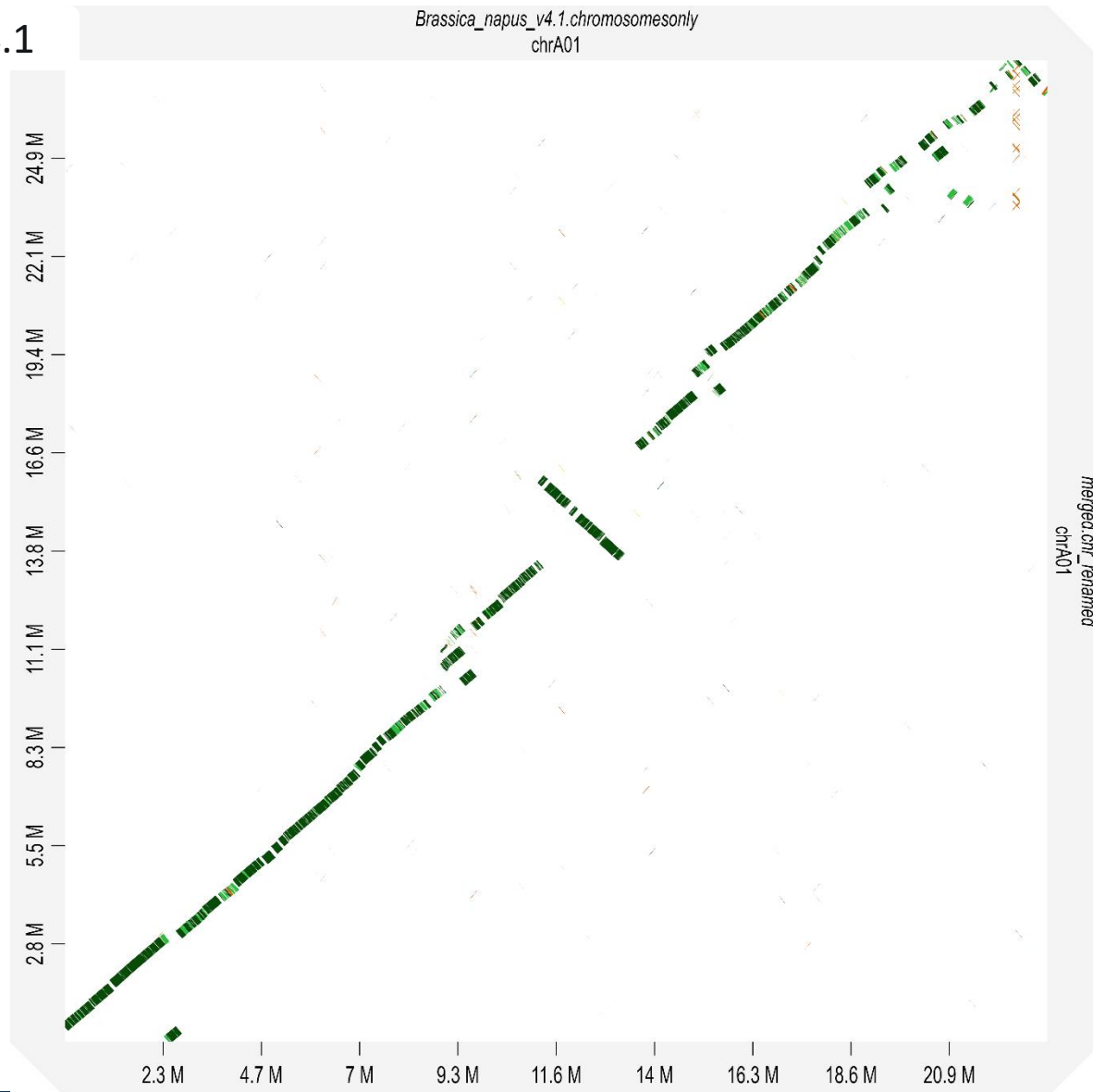
Darmor 4.1

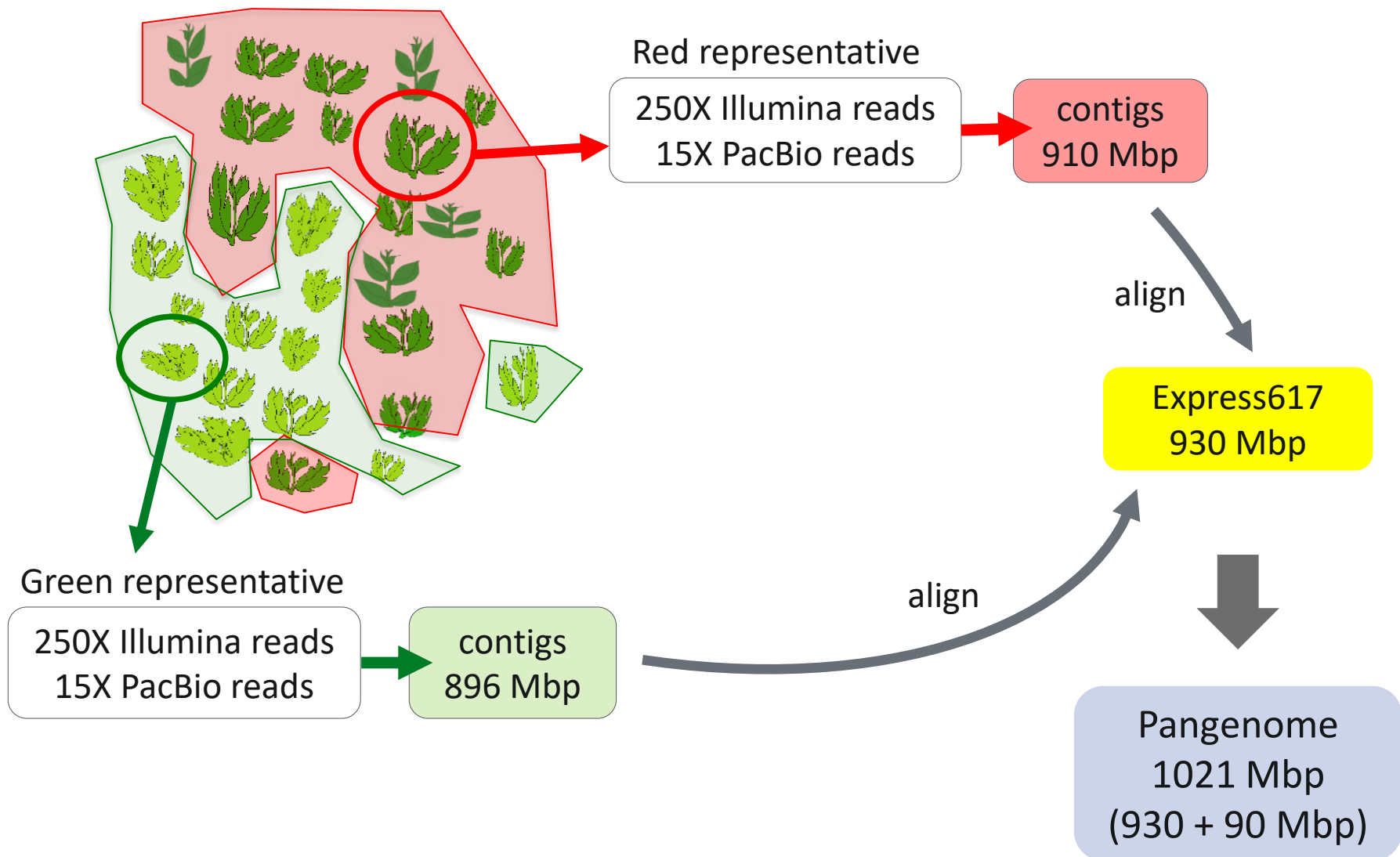


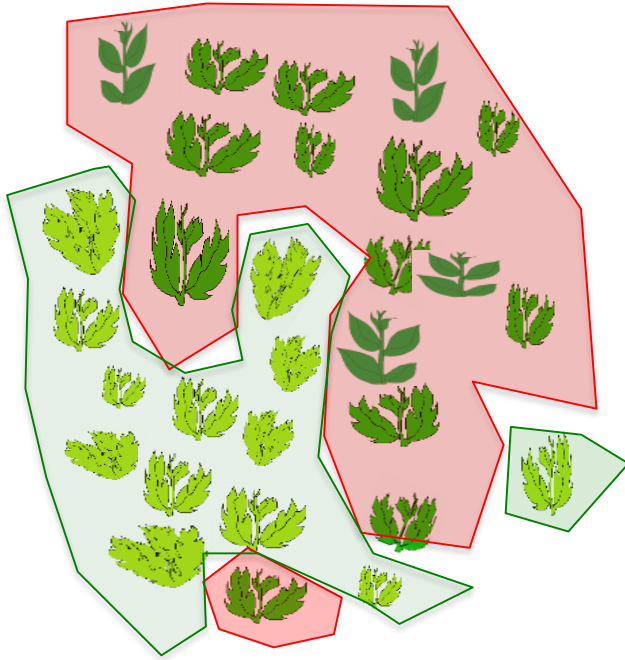
“All assemblies are wrong, some are useful” – Adam M. Phillippy
Assemblies are constantly being improved!

The Express617 genome assembly

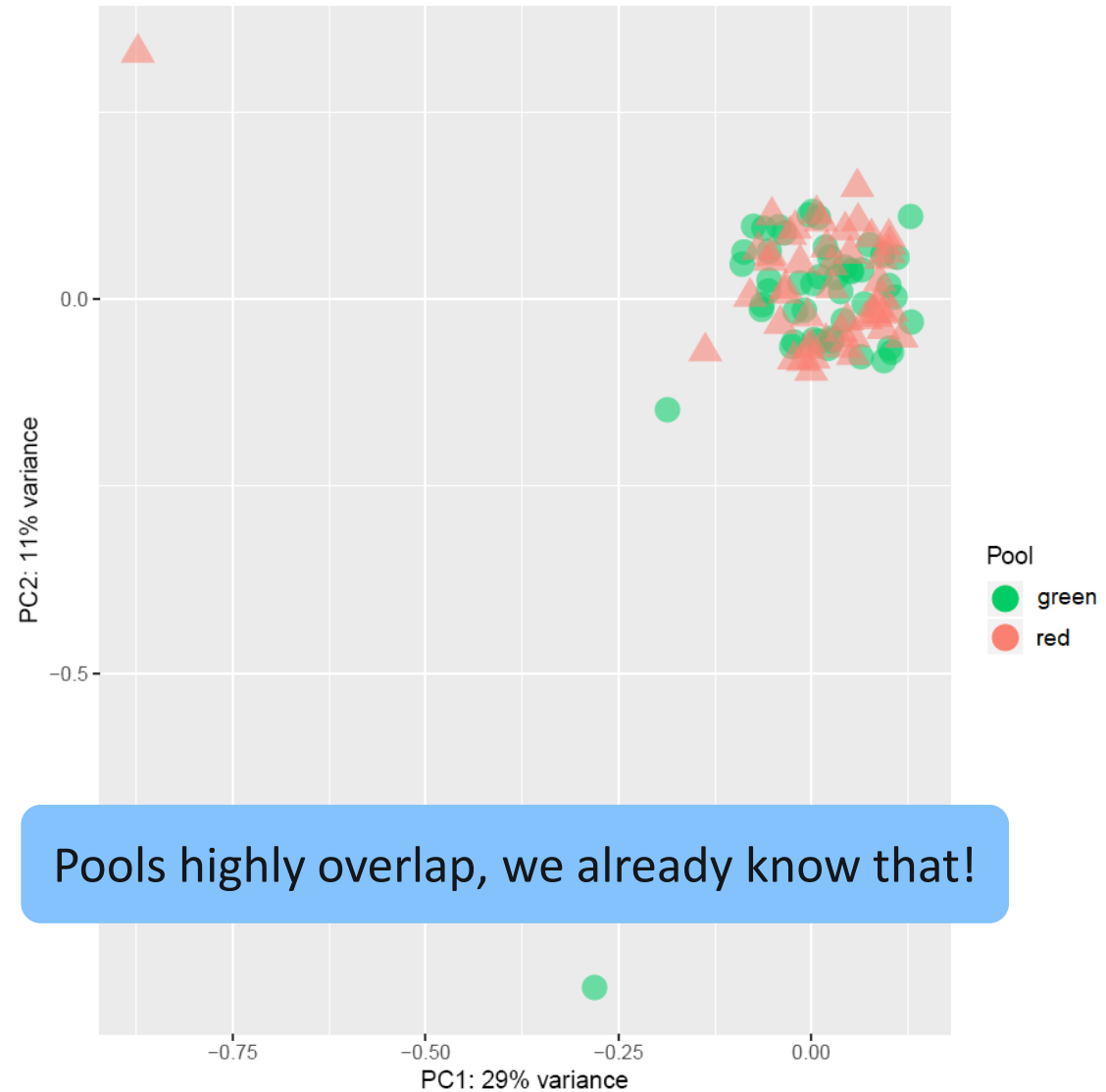
Darmor 4.1



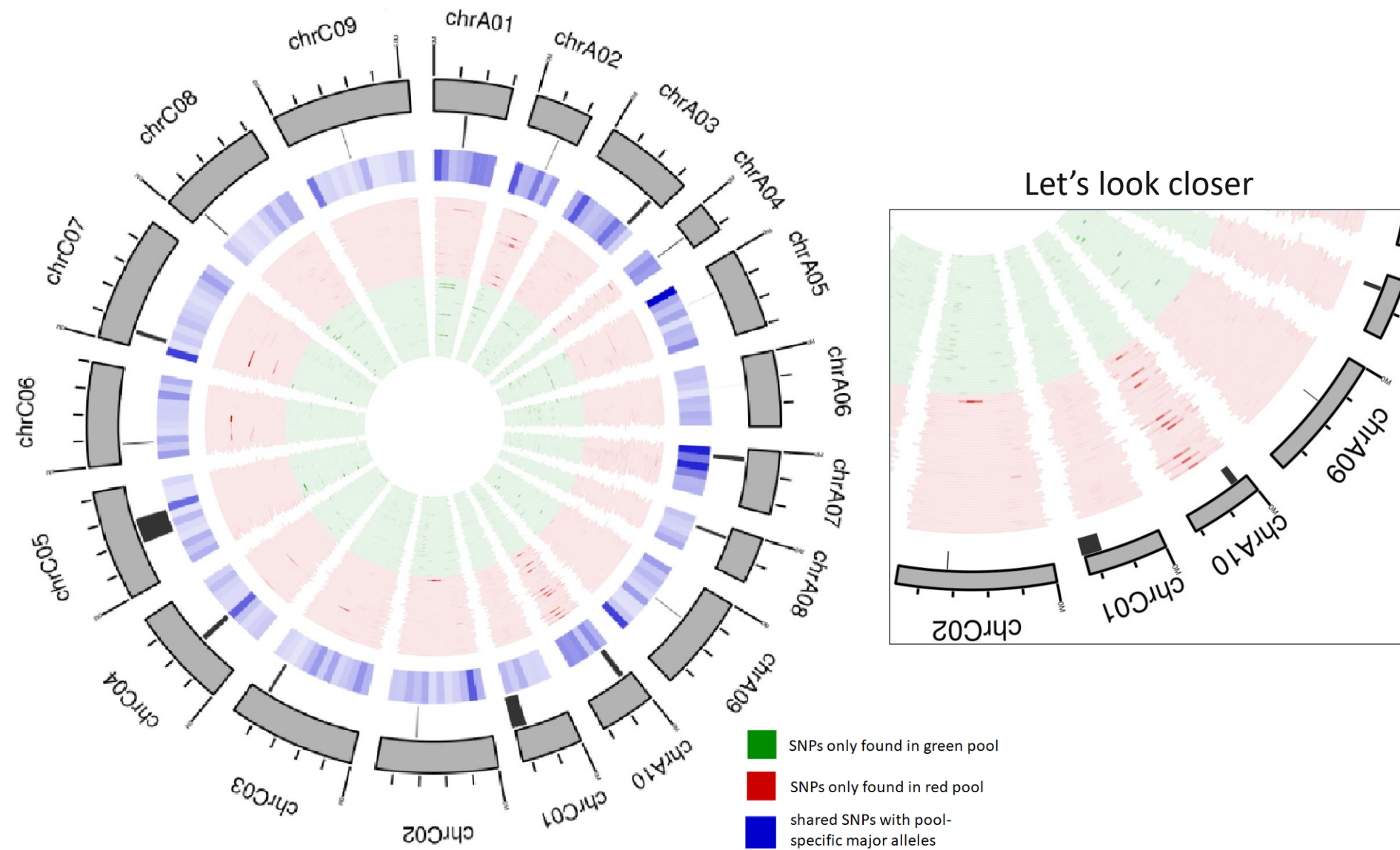




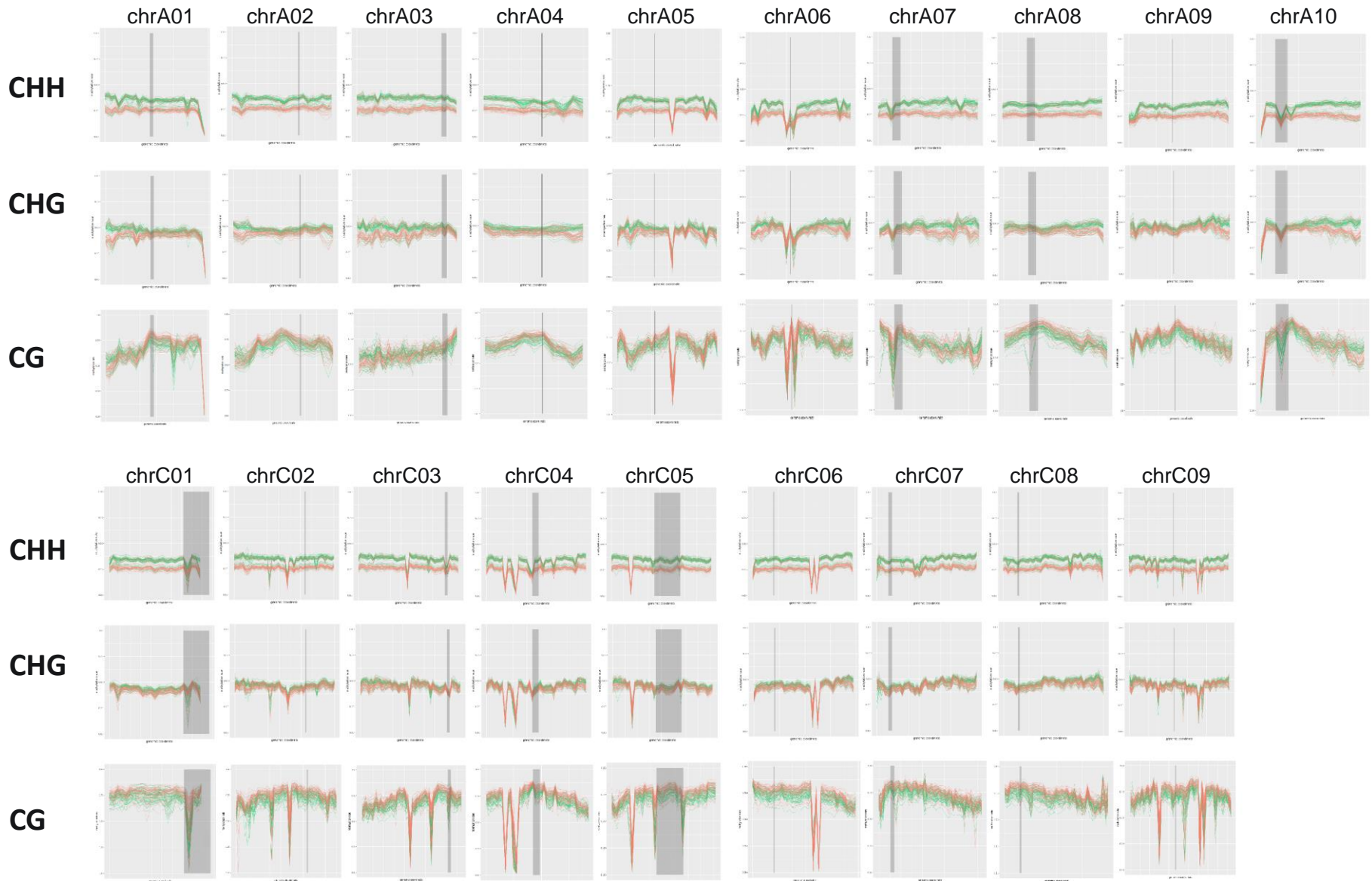
Nearly 70% of SNPs are shared between pools



SNPs show pool specificity

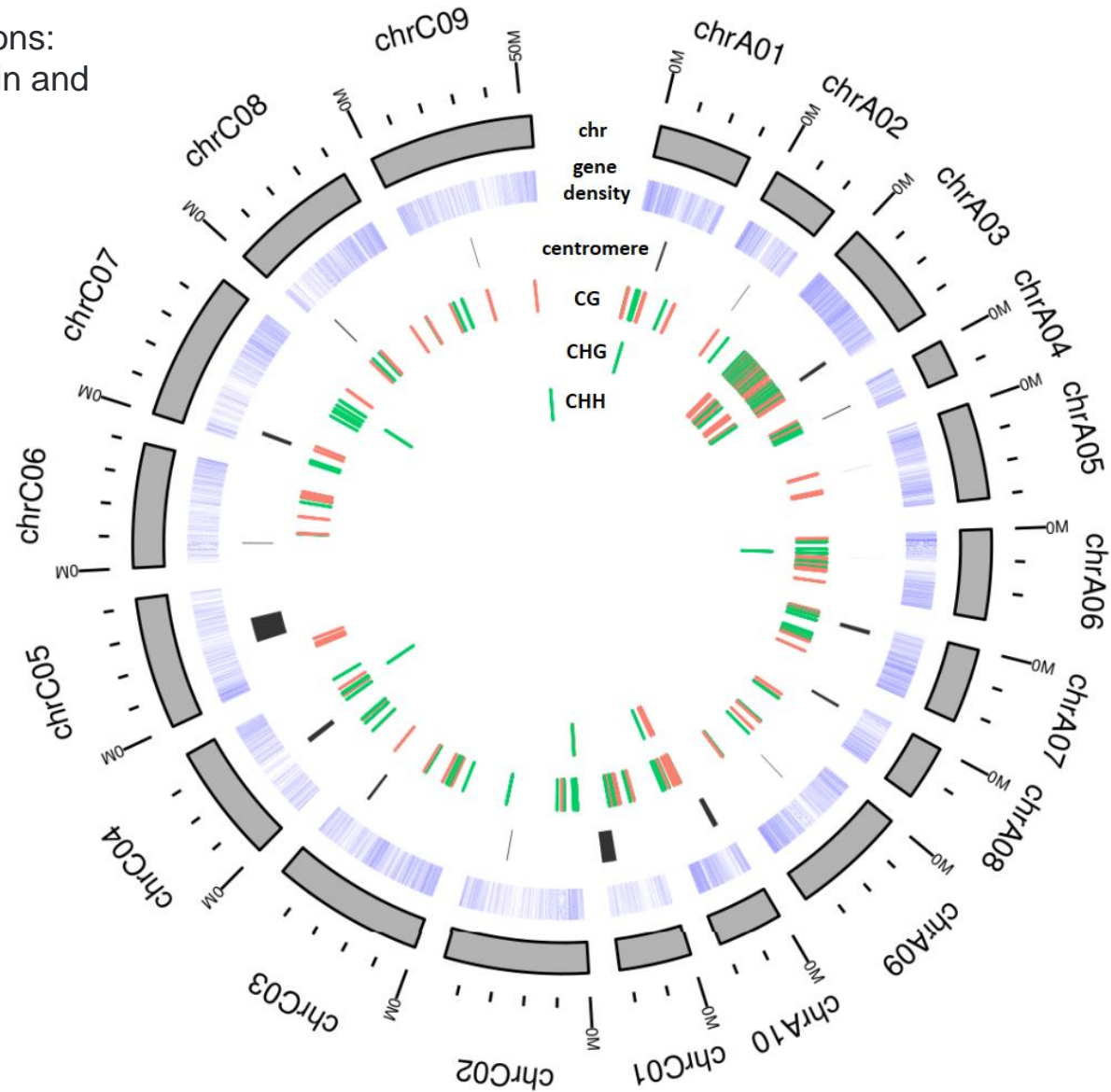


Founder lines genome-wide methylation rate

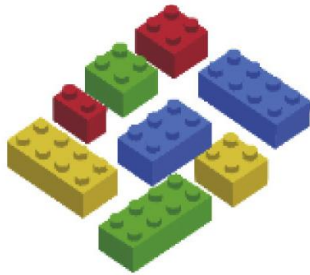


Methylation rate for CG CHG and CHH show pool specificity

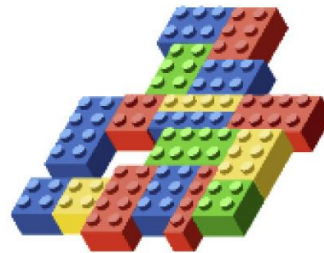
Differentially methylated regions:
Rate difference $\geq 50\%$ per bin and
Score test p value ≤ 0.01



- We show that SNPs and methylation patterns in founders are pool-specific
- The fixation of these variants contributes to successful pool separation
- Together with mRNA and miRNA, we hope to use haplotype index data to better predict heterotic potential with ML approaches



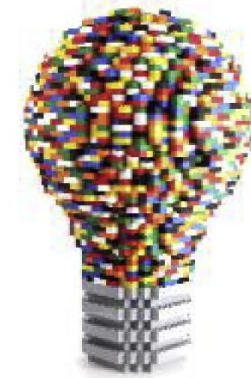
Preference
Experience



Statistics test
Mechanization
Fertilizer/pesticide



Marker assisted selection
Transgenic technology



Knowledge driven
Multi-dimension data

5000 BC - 1900

1st Green
revolution

2nd Green
revolution

Present – 3rd Green
revolution

Adapted from Fernie and Yan, 2019

Rod Snowdon



@pbgiessen



Gunhild Leckband
Amine Abbadi
Tobias Kox
Felix Dreyer



Björn Usadel
Asis Hallab



Matthias Frisch
Daniel Krenzer
Eva Herzog

Marius Kloft
Robert Vandermeulen

