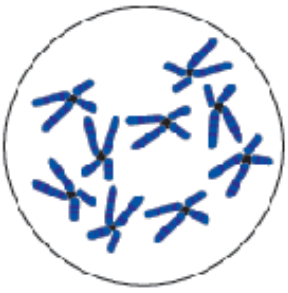
Structural genome variation and its consequences for rapeseed genetics and breeding

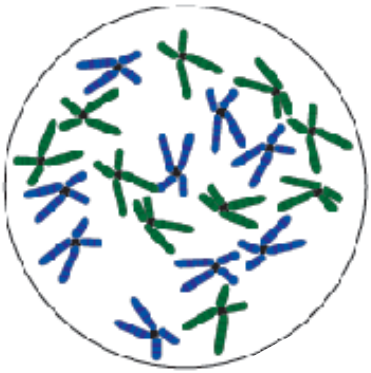
Rod Snowdon
Department of Plant Breeding, Justus Liebig University, Giessen, Germany

Brassica napus: A recent, anthropogenic allopolyploid

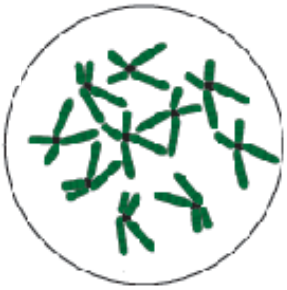


Brassica oleracea
2n = 2x = 18
Genome CC

Mediterranean cabbages and kales



Brassica napus
2n = 4x = 38
Genome AACC



Brassica rapa
2n = 2x = 20
Genome AA

Asian cabbages, Sarson and turnips

Rapeseed/canola, kale & swede/rutabaga

Originated under cultivation, just a thousand years ago, from very few founding hybridisation events – wild forms unknown

Rags to riches in three decades: A “Cinderella” crop…

Today the world’s No. 2 oilseed crop

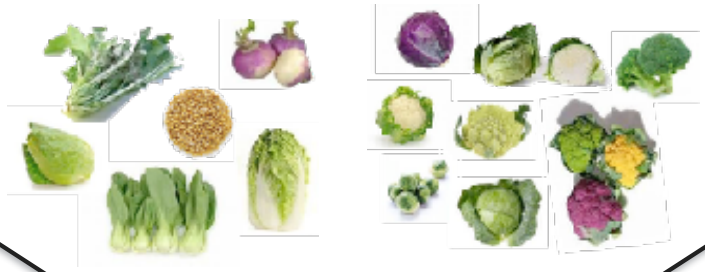
Healthiest vegetable oil for human nutrition, equally suitable as a biofuel

High-quality, protein-rich extraction meal for livestock feed

Vital component in cereal crop rotations

A paradoxical polyploid fairy-tale?

Diploid progenitor species



Strong allopolyploidisation bottleneck

***De novo* allopolyploid**
Very few species founders

Seed quality selection
Extreme breeding bottlenecks

Extreme erosion of genetic diversity essential for breeding

However: Unexpected **adaptive capacity** and **breeding success**

Rod Snowdon, JLU Giessen

Learning from *Brassica napus*

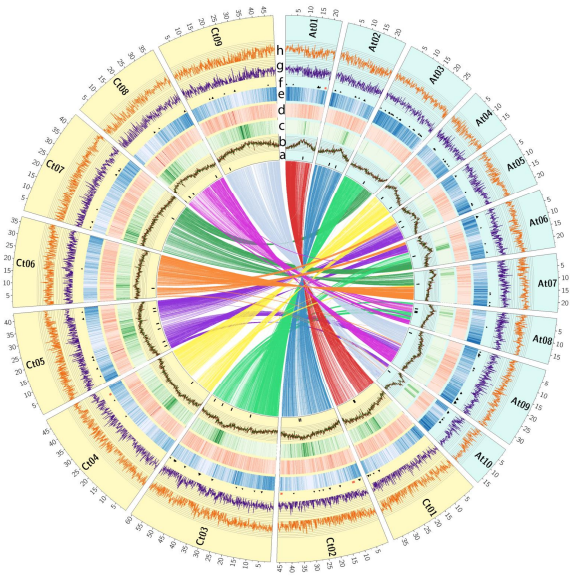
Retrospective learning

- How could allopolyploids (ancestral and recent) achieve sufficient *de novo* diversity (despite significant polyploidisation bottlenecks) to adapt to completely new environments and become highly successful crops?

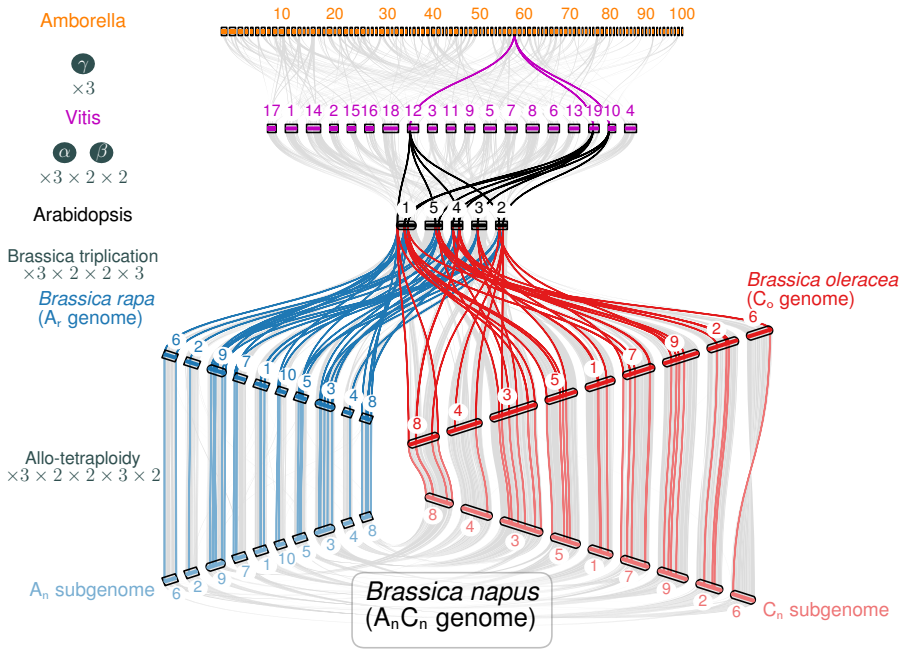
Breeding applications

- Can the *B. napus* story help us to more efficiently use *de novo* allopolyploids (synthetics) to reinstate diversity into depleted crop gene pools, driving adaptation to newly emerging environmental challenges?

The complex, allopolyploid genome of *Brassica napus*



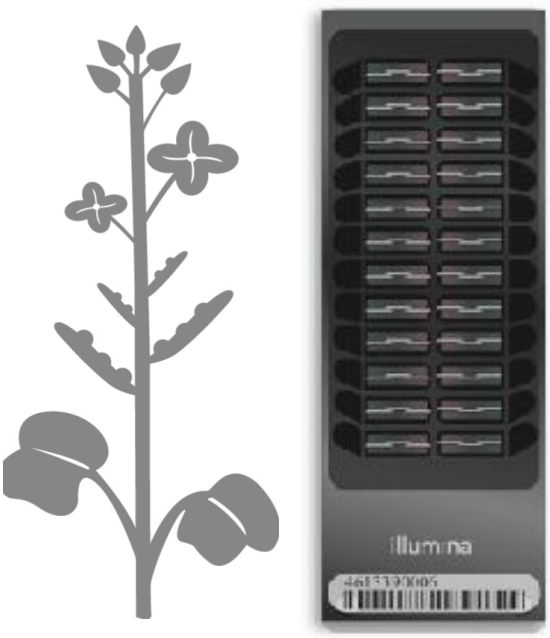
Chalhoub et al., *Science* (2014)



- Allopolypoid from hybrid between between two ancestral hexaploids
- Extensive inter- and intra-subgenomic duplication and homoeology

High-throughput genotyping in a complex polyploid

Brassica 60k SNP Infinium consortium array (released 2012)



Clarke et al., *TAG* (2016)

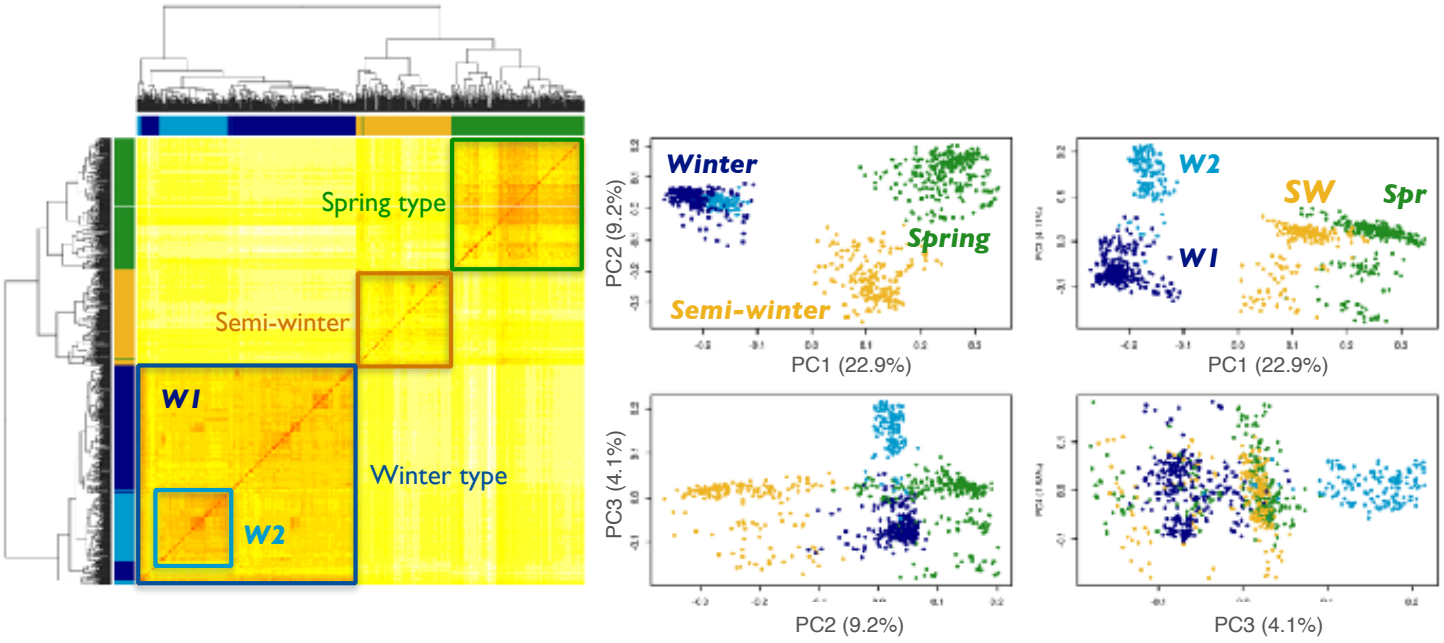
- Highly successful community designed SNP genotyping array
- Widely used in rapeseed/canola genetics and breeding

However:

- ~40% of the Infinium assays address multiple loci (hemi-SNPs)
- Hence, most downstream analysis (mapping, GWAS, GS etc.) works best with a filtered panel of **~25,000 single-locus SNPs**

Disturbingly low SNP diversity in major breeding pools?

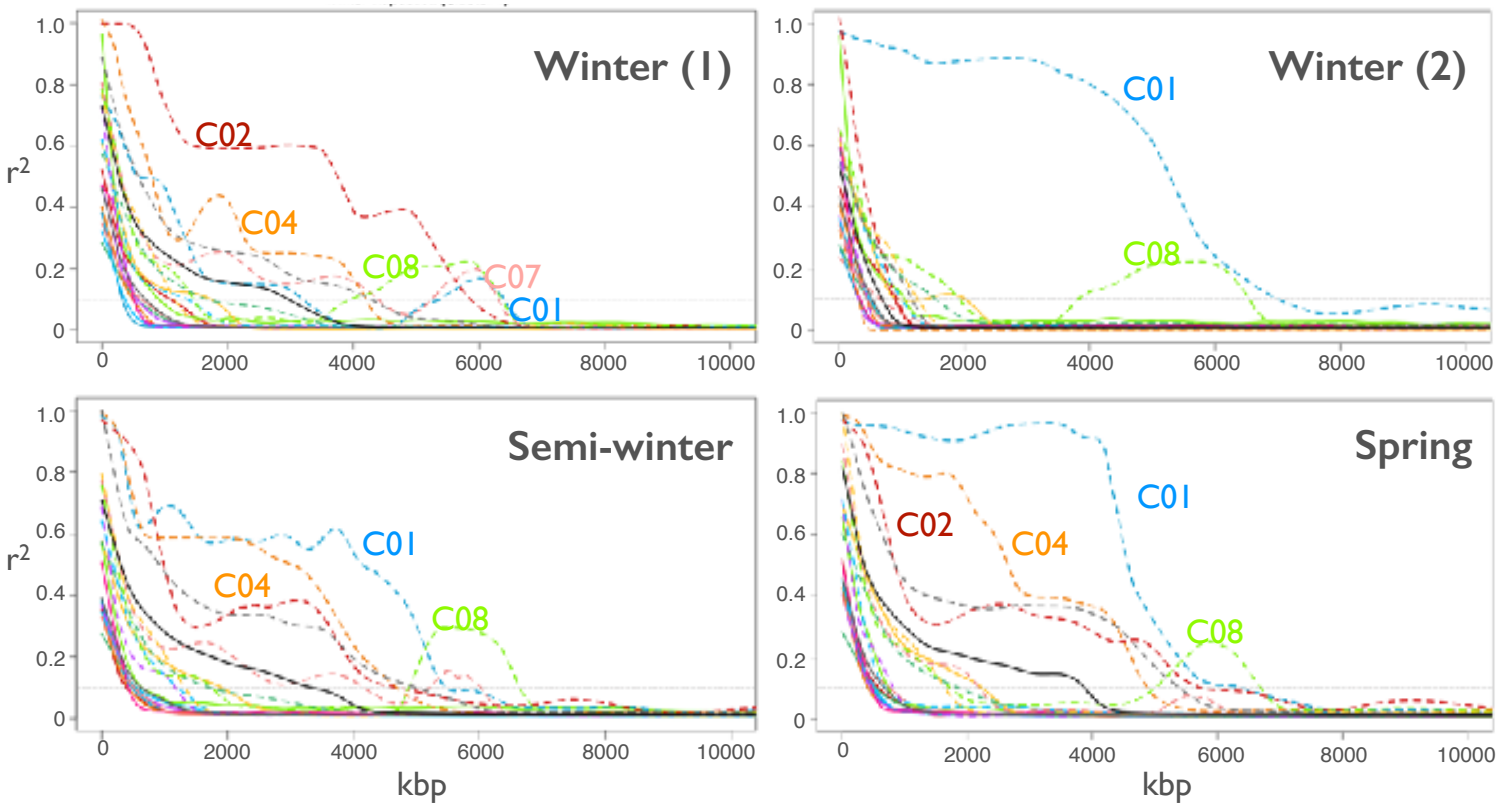
Genetic diversity analysis in a species-wide germplasm collection (n~850) based on **~20,000 single-copy, polymorphic SNPs**



- SNP diversity is relatively high between ecogeographic pools, but appears to be very low within the major gene pools

Chromosomal and genomic bias in allelic diversity

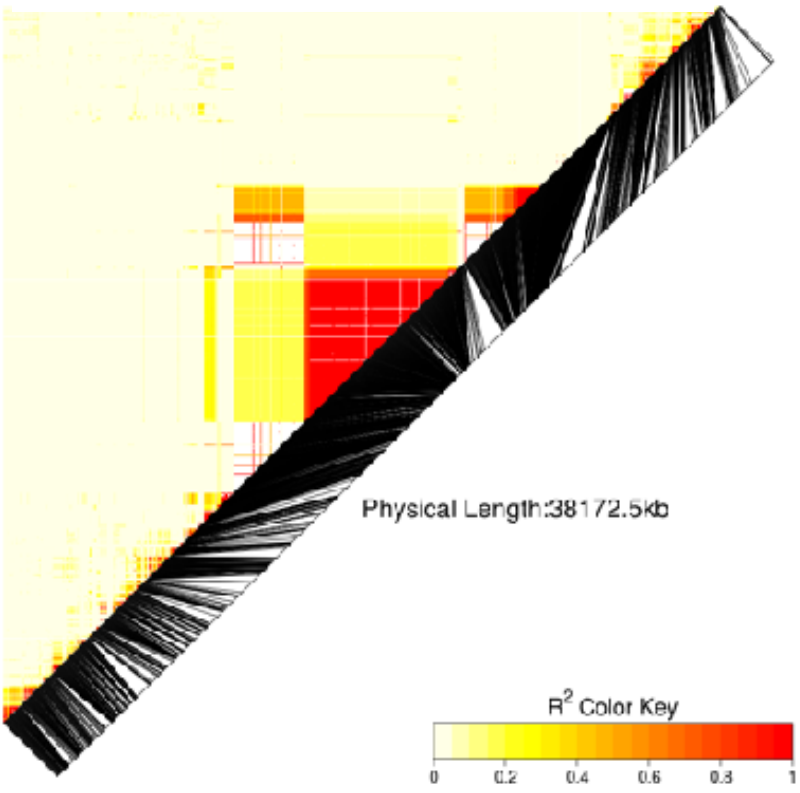
Decay of linkage disequilibrium in different *B. napus* breeding pools



Qian et al. *BMC Genomics* (2014), Jan et al. *PLOSone* (2015), Schiessl et al. *Front Plant Sci* (2014)

Extreme selection signatures on some chromosomes

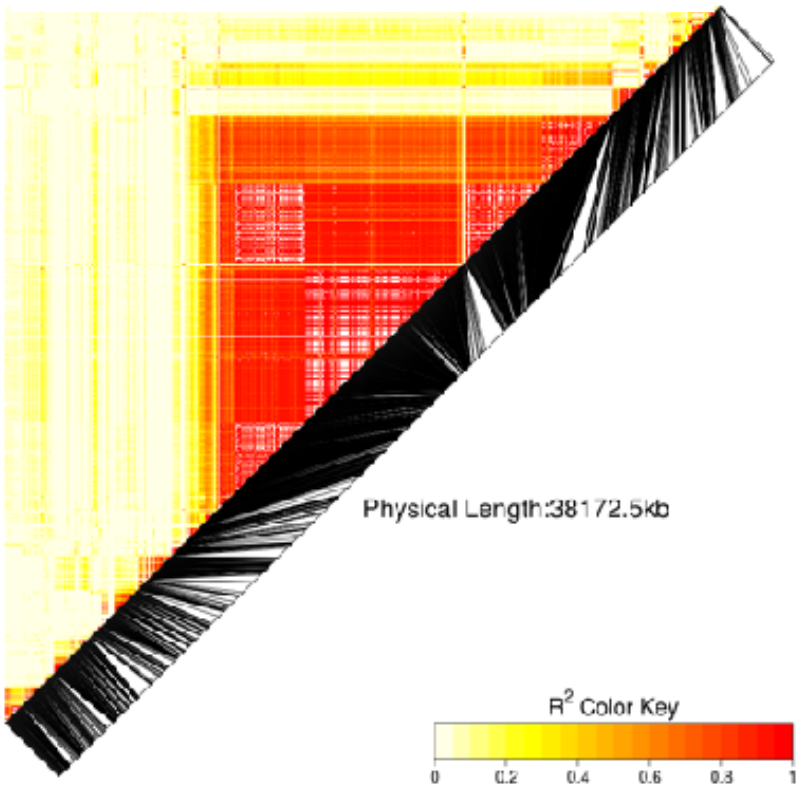
Chromosome C02, winter-type (1)



Christian Werner & Lunwen Qian
Rod Snowdon, JLU Giessen

Extreme selection signatures on some chromosomes

Chromosome C02, winter-type (2)

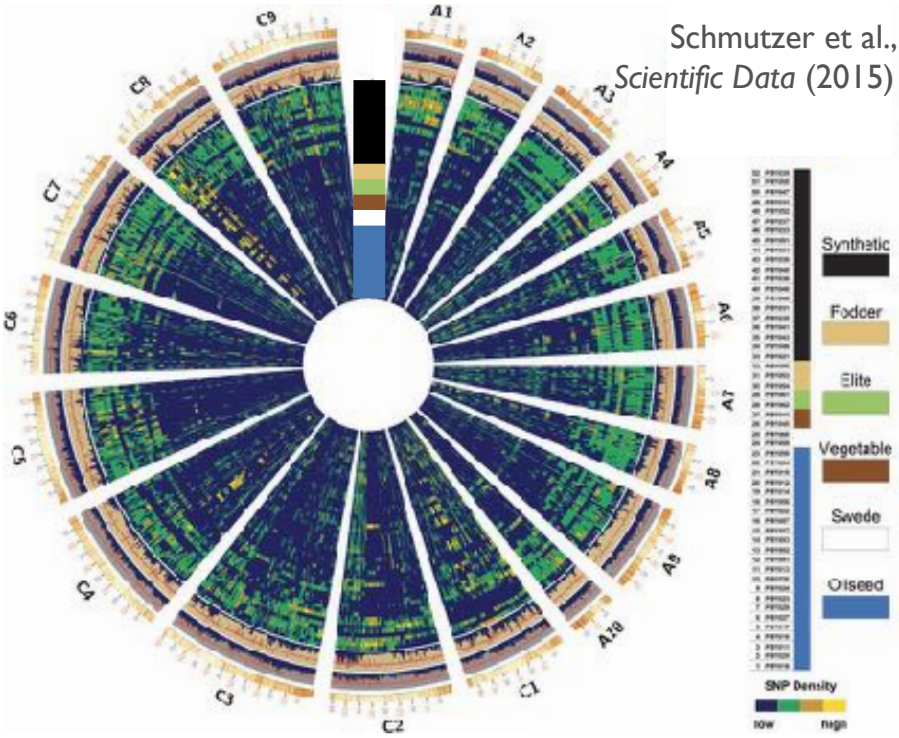


Genomic resequencing: SNP diversity actually not bad?

PreBreed-Yield
BnNAM Consortium

20 highly diverse synthetic &
32 exotic, natural *B. napus*
(BnNAM founders)

~15x coverage, 100bp PE
~4M high-quality SNPs

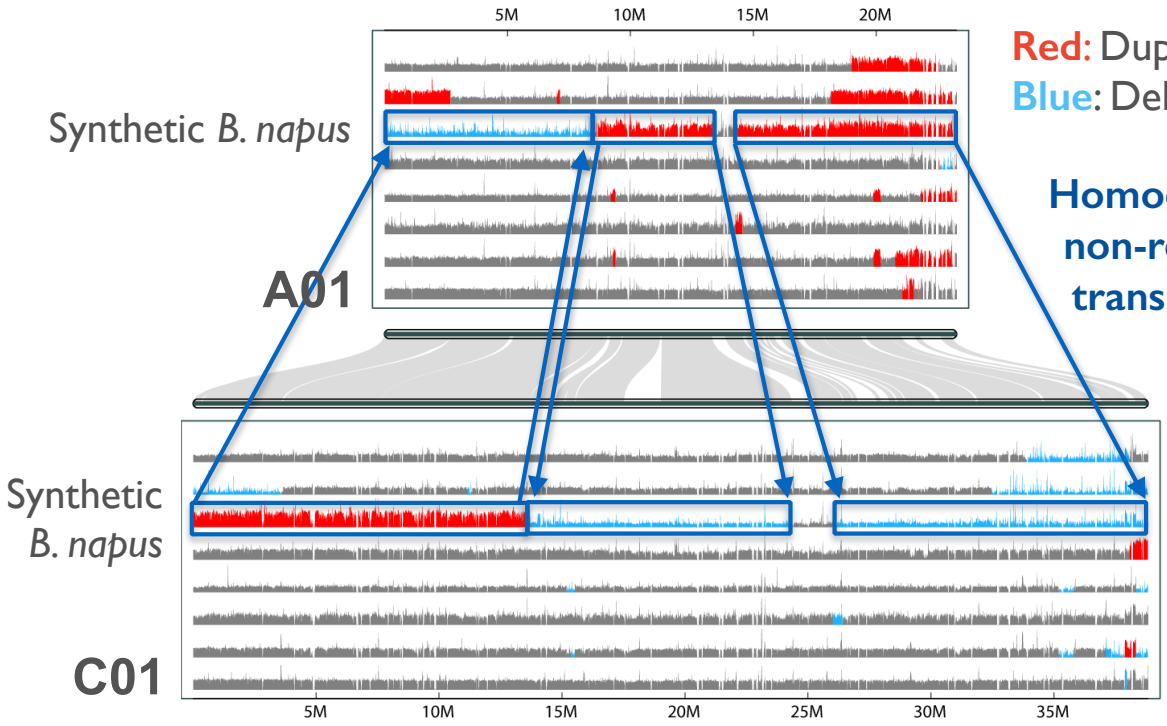


Schmutzer et al.,
Scientific Data (2015)

- **Contradiction between array and sequence data:** Is decisive sequence variation obscured in duplicated genes/genome regions?

B. napus: Widespread structural genome variation

Chalhoub et al., *Science* (2014)

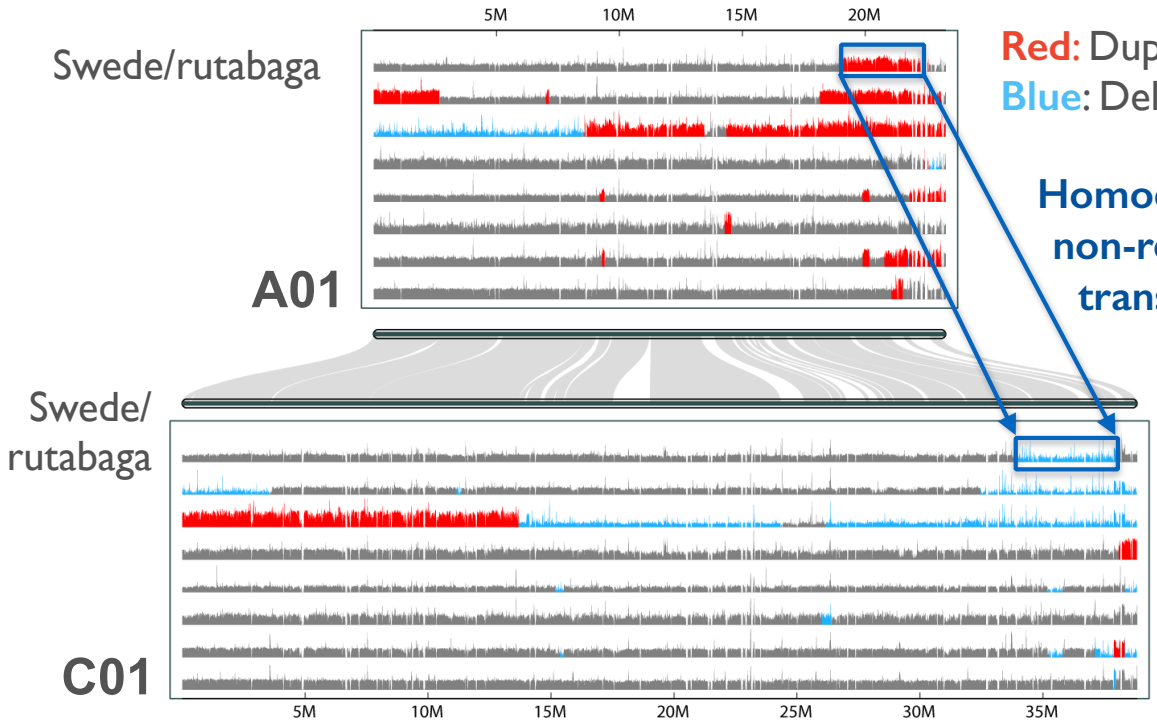


Red: Duplicated segments
Blue: Deleted segments

Homoeologous non-reciprocal translocations

B. napus: Widespread structural genome variation

Chalhoub et al., *Science* (2014)

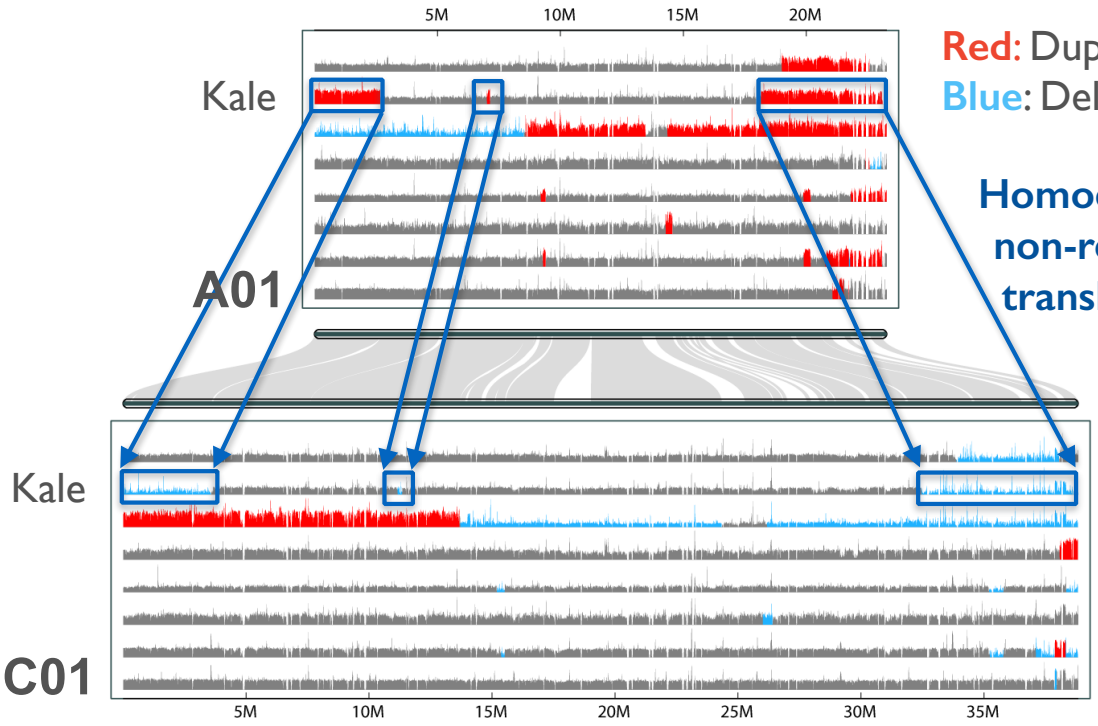


Red: Duplicated segments
Blue: Deleted segments

Homoeologous non-reciprocal translocation

B. napus: Widespread structural genome variation

Chalhoub et al., *Science* (2014)

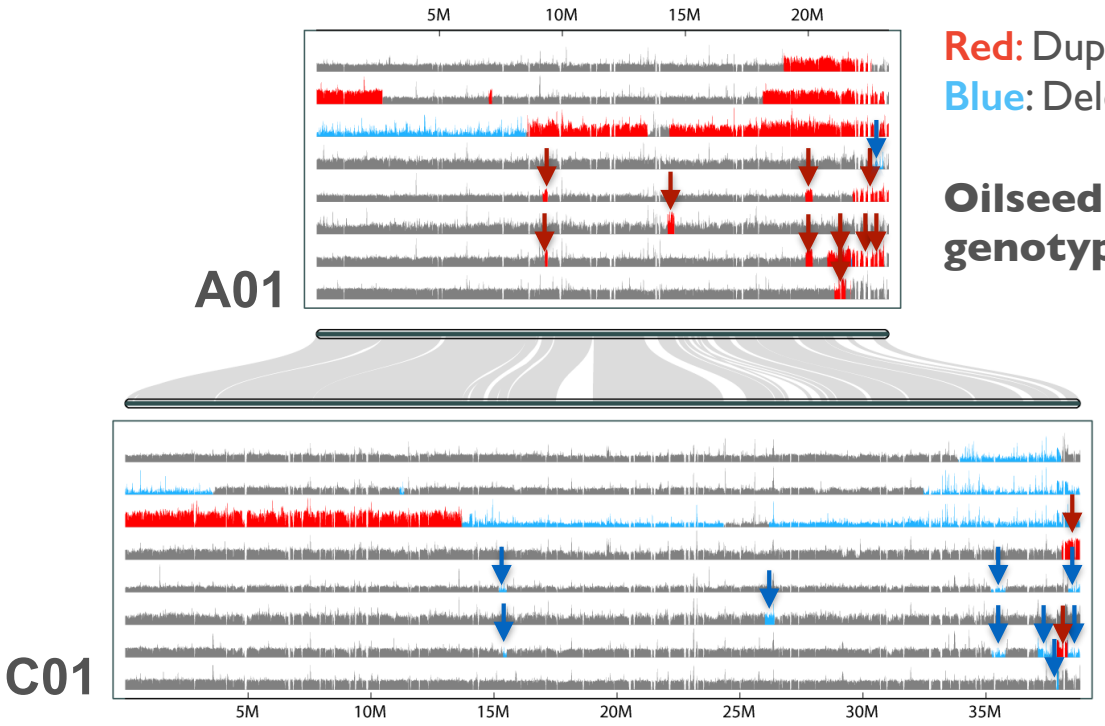


Red: Duplicated segments
Blue: Deleted segments

Homoeologous non-reciprocal translocations

B. napus: Widespread structural genome variation

Chalhoub et al., *Science* (2014)



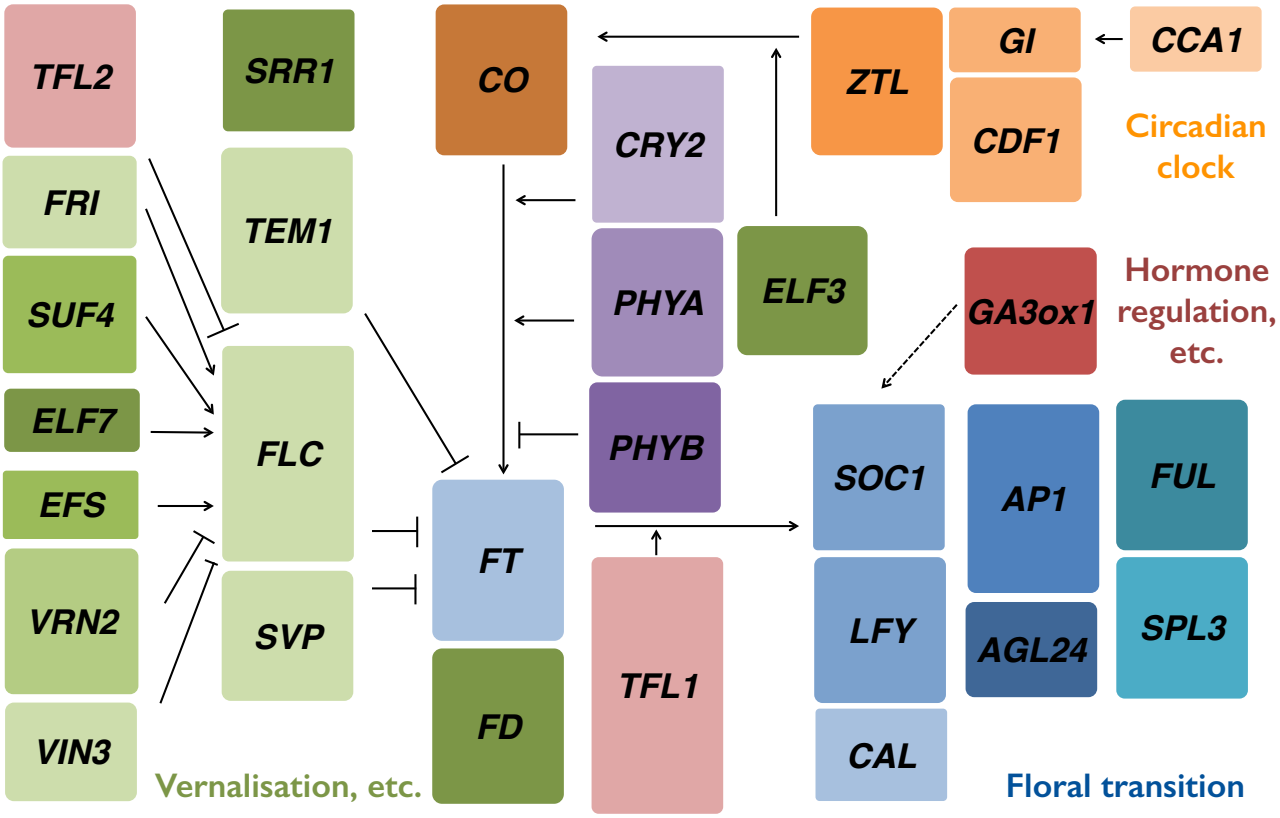
Red: Duplicated segments
Blue: Deleted segments

Oilseed genotypes

Oilseed genotypes

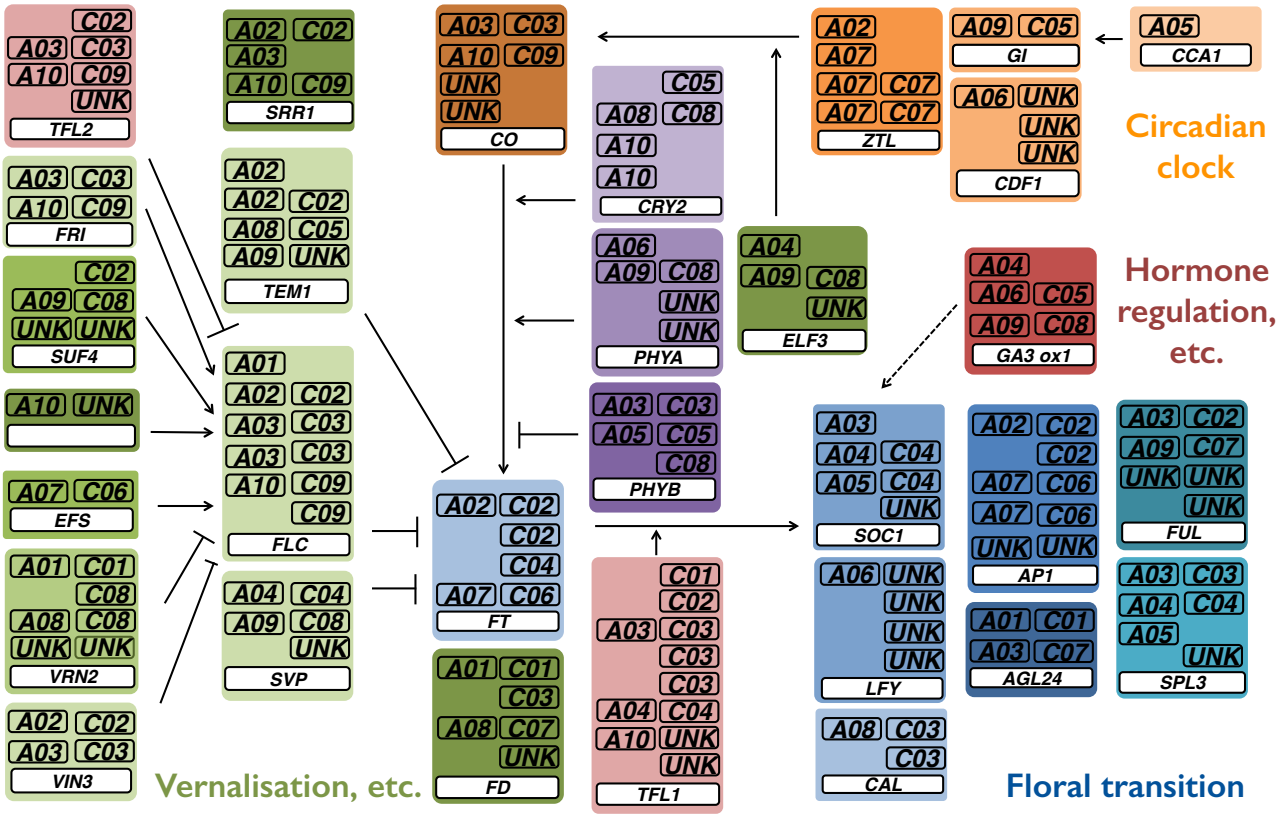
- Also widespread small-scale rearrangements, including frequent gene conversion events (right down to single-nucleotide exchanges)

Structural variation: Adaptive potential from PAV/CNV



Example: The complex Arabidopsis flowering-time regulatory pathway

Bn: Multiple copies increase tolerance for mutation

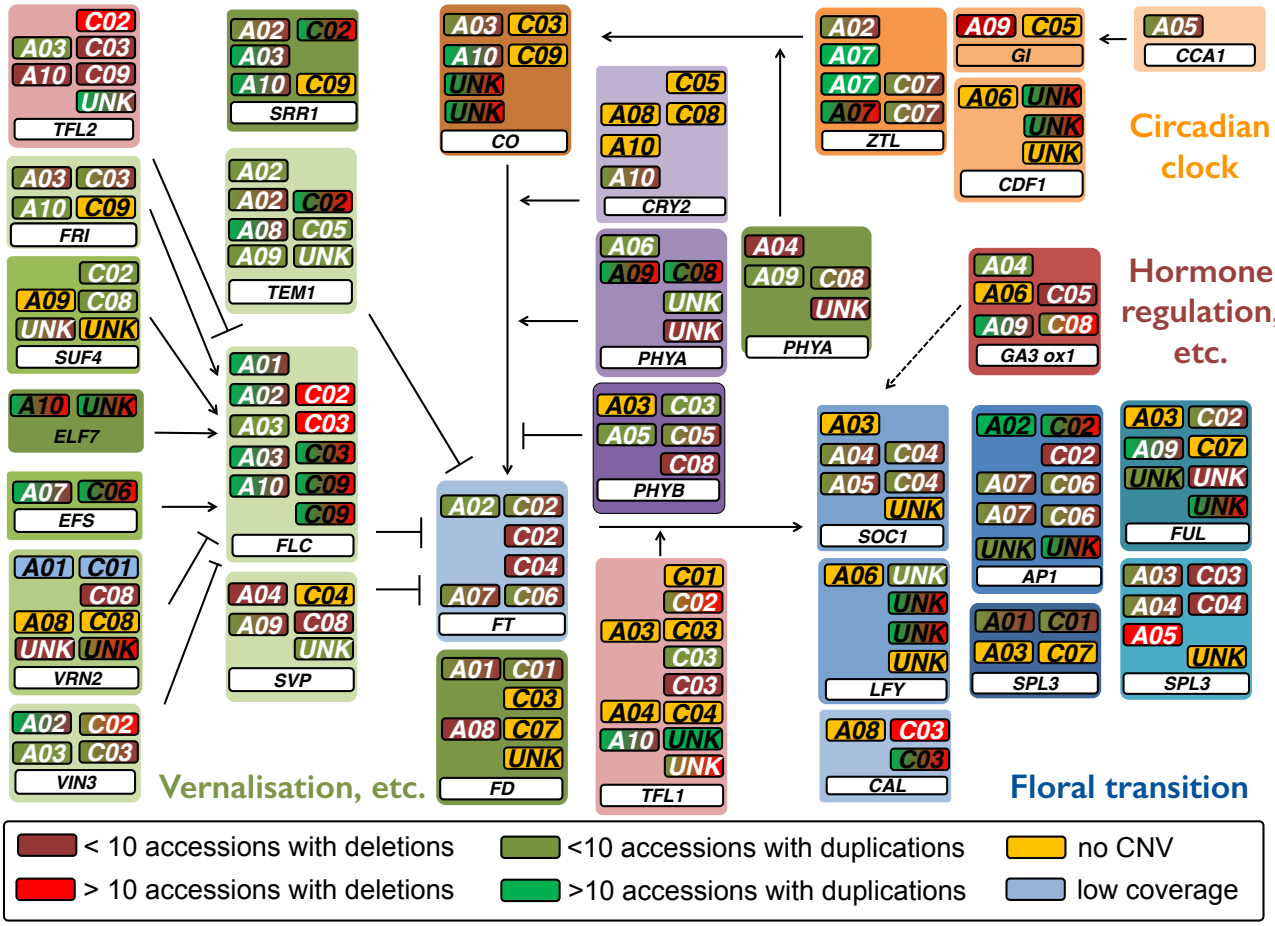


Species-wide sequence capture:

~600*x* deep-sequencing of 32 gene families in 280 *B. napus* accessions

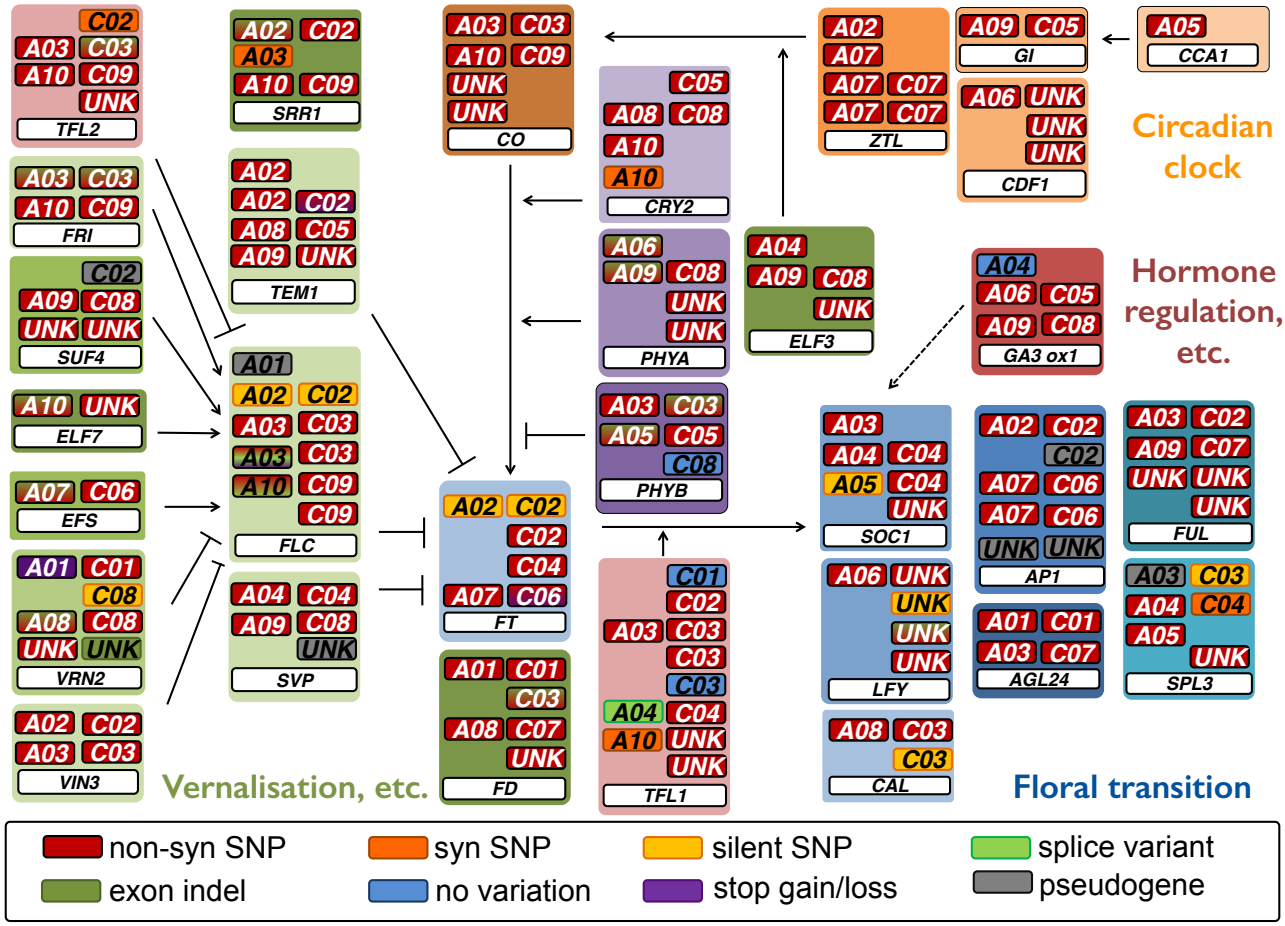
Schiessl et al., *Sci Data* (2017), *Sci Rep* (2017)

Not one gene family without PAV in at least one copy



Schiessl et al., *Sci Data* (2017), *Sci Rep* (2017)

Extensive, functionally relevant SNP variation



Schiessl et al., *Sci Data* (2017), *Sci Rep* (2017)

Consequences of genome rearrangements in *B. napus*

Positive

- Genome structural diversification is widespread and ongoing in *B. napus* and is a valuable source of untapped diversity for breeding
- Significant potential influence on quantitative trait control
- Expected contribution to additive heterosis and "heterotic haplotypes"

Negative

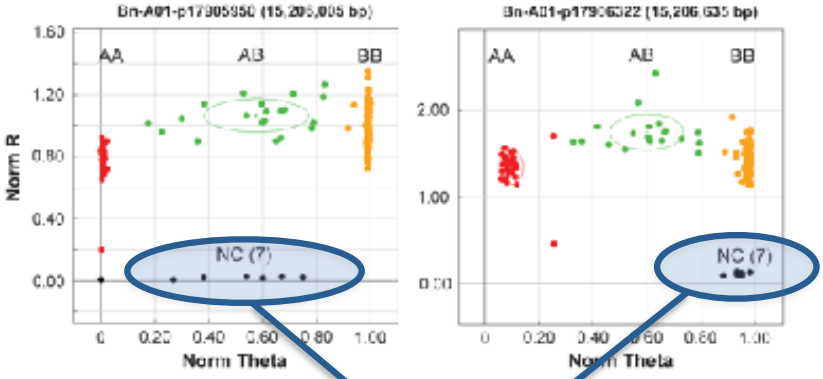
- SNP markers in regions with CNV or PAV are likely to be discarded by standard QC parameters in auto-calling procedures, due to excessive "failed calls" …
- and even if you do keep those markers in your dataset, downstream software will frequently throw them out of your analysis anyway because they show excessive "missing data"

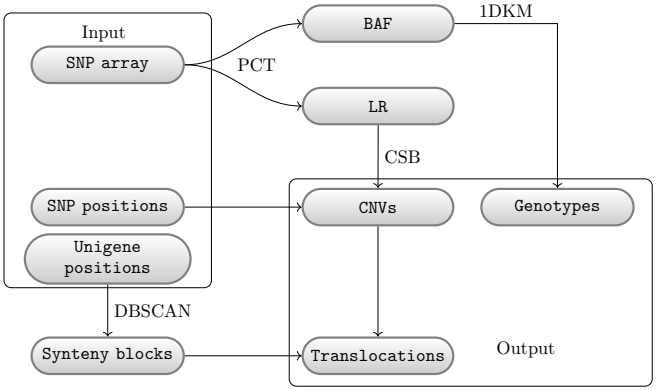
Challenge

- Need for a simple and cost-effective technique to recover "missing data" from deletions to enable their use for quantitative trait elucidation and breeding

Identifying structural variants from SNP array data

gsrc: An R package for genome structure rearrangement calling





1DKM	1-Dimensional K-Means Clustering
BAF	B-Allele Frequency
CSB	Circular Binary Segmentation
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
LR	Log R ratios
PCT	Polar Coordinates Transformation

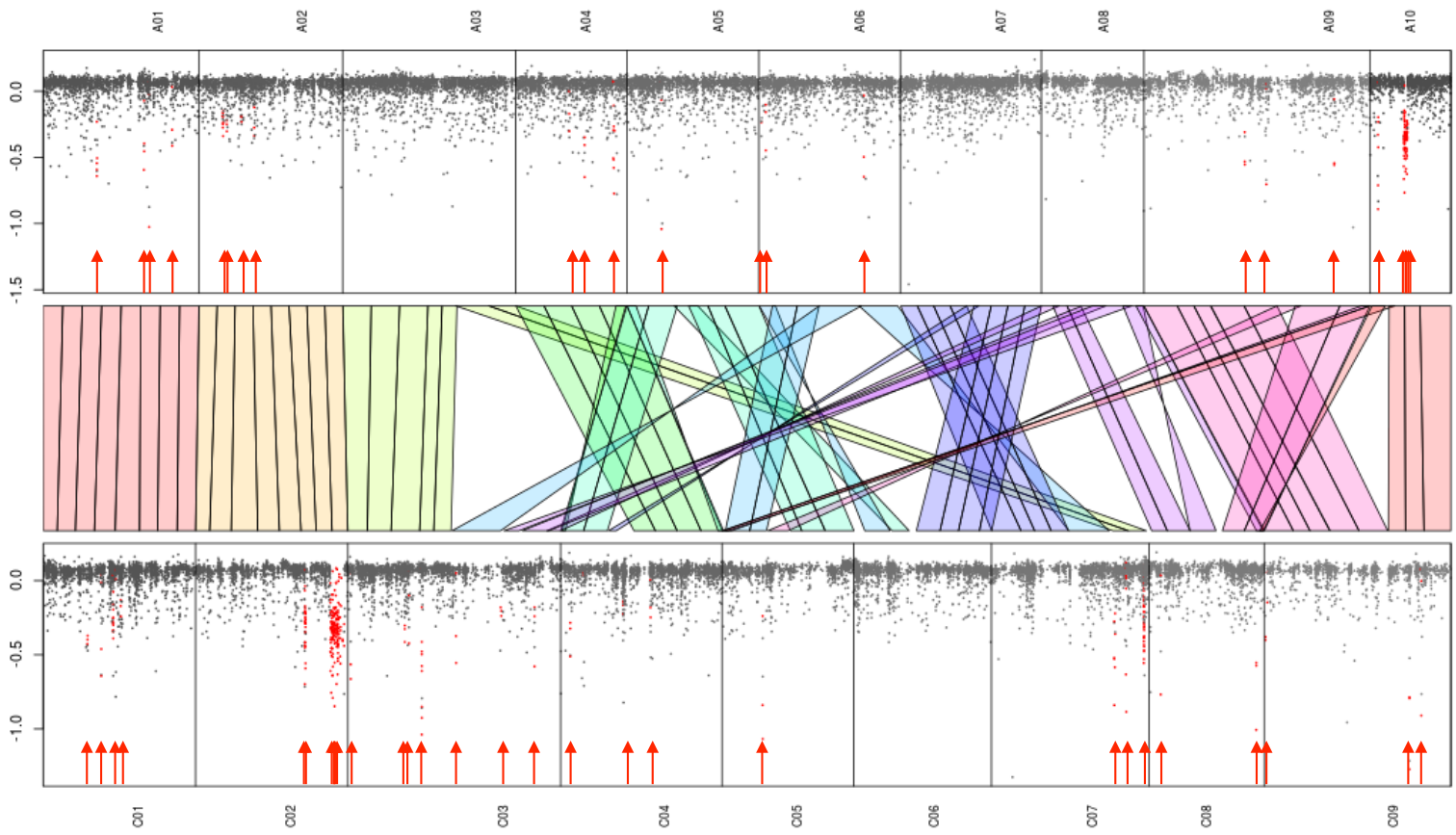
Physically adjacent SNPs which consistently "fail" in the same genotypes represent deletions in those genotypes and can be confidently called as **SNAP**s

(Single-nucleotide absence-presence)

~ 5-20% of assays on *Brassica* 60k SNP array

Grandke et al., *Bioinformatics* (2016)

Auto-calling of deletions from Illumina 60k-SNP data



Grandke et al., *Bioinformatics* (2016)

Reanalysis of SNP array data for structural variants

High-throughput calling of structural variants from array data

- Additional **auto-called output matrices** describing **PAV** (SNAP) & **CNV**
- Inclusion of SNAP data enables **consideration of additive allele effects that are normally excluded from mapping or breeding datasets** due to excessive "missing" or falsely-segregating marker data
- **Meets No. 1 breeder criterion: Cheap, reliable and quick!**
- **…and samples the most variable, trait-relevant parts of the genome**

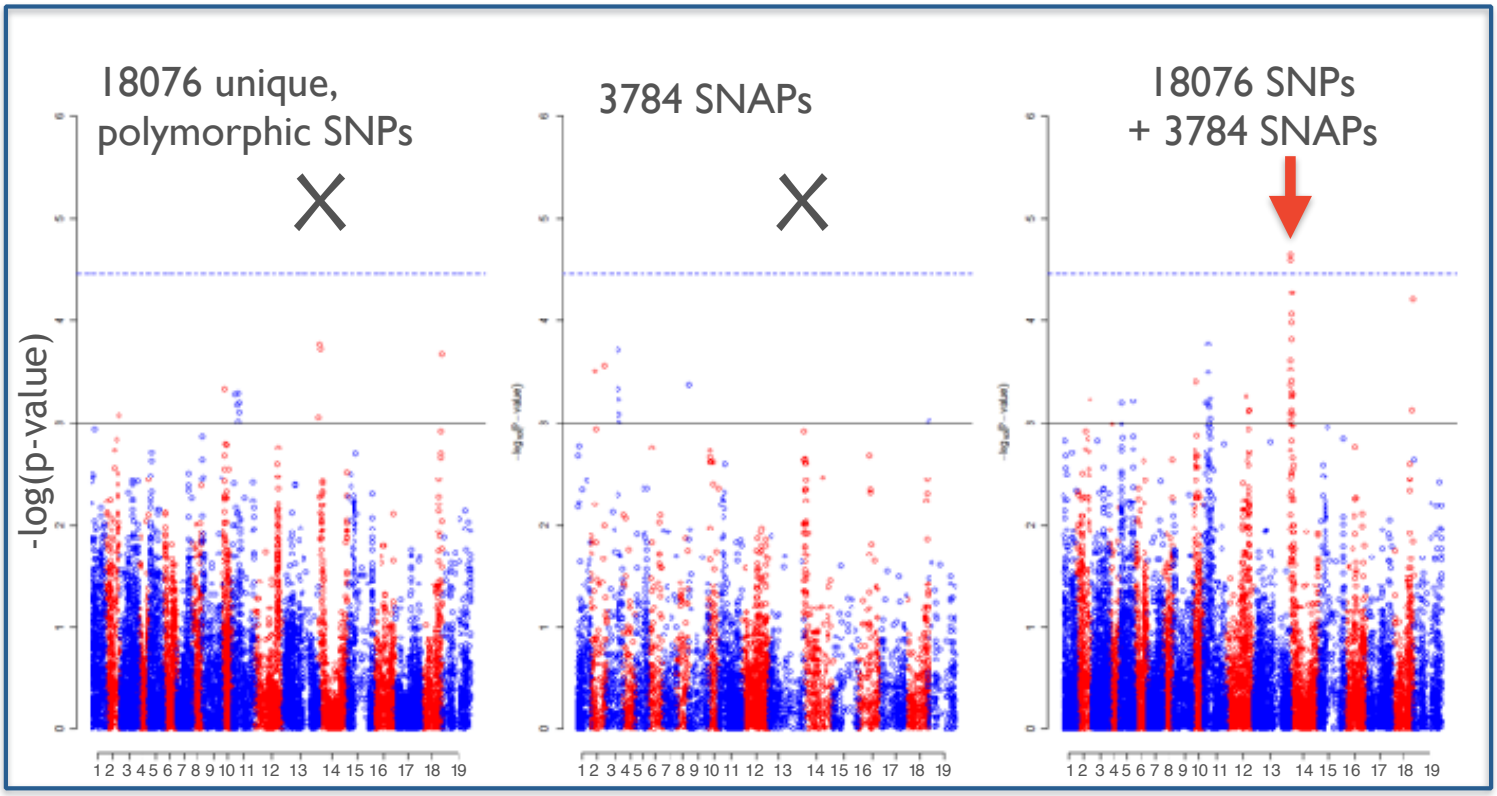
Work in progress

- Reanalysis of historical Illumina 60k array raw data from >5000 genotypes (biparental mapping populations, *B. napus* diversity panels, BnNAM populations, breeding lines, F1-hybrids, etc.)
- Inclusion of **PAV data matrices** as an **additional dimension in genetic maps, QTL/GWAS analysis, genomic prediction models**, etc.

CNV/PAV data reveals “invisible” QTV

Example: Quantitative disease resistance

- GWAS MLM(K): BnNAM subpopulation, n=250



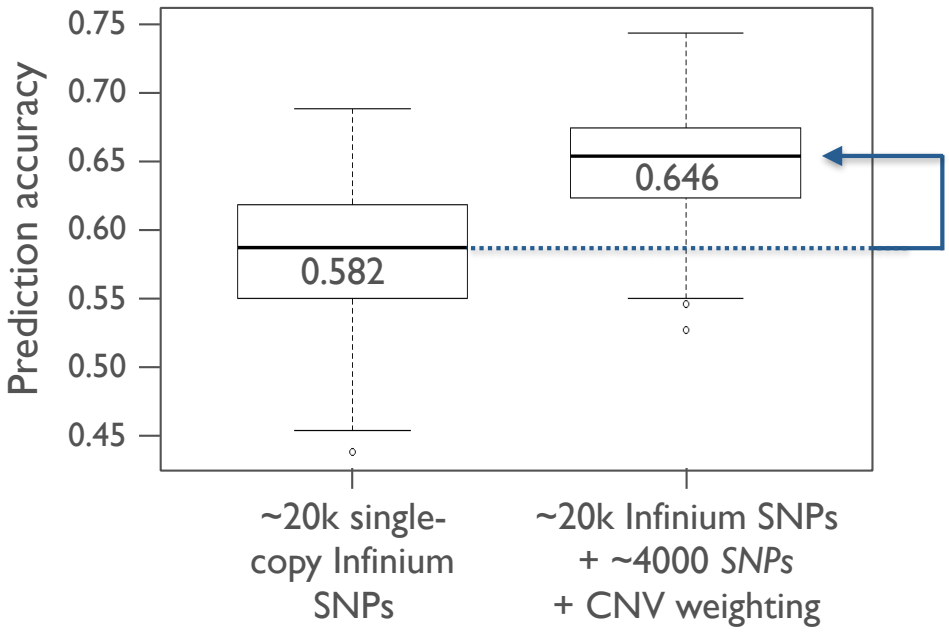
Fabian Grandke, Iulian Gabur, Christian Obermeier

Rod Snowdon, JLU Giessen

Structural variants contribute to heterosis

Example: Genomic prediction of hybrid yield performance

- rrBLUP for hybrid yield performance in 450 test hybrids
- Training population n=300, yield data from 12 environments
- Prediction using *in silico* F1 genotype data



6.4% jump in prediction accuracy for hybrid yield by using SNP & SNAP marker calls that consider structural genome variants

Using structural genome variation in rapeseed breeding

SNAPs – discovering and using the value in "missing" data

- Considerably improved elucidation of quantitative traits in *B. napus*
- Genomics-based diversification of heterotic pools
- Targeted fixation of complementary, additive heterotic effects driven by widespread deletions

Enrichment of diversity by introgressions from synthetics

- Exploiting the extended pan-genome, elevation of gene diversity
- Increased recombination of new diversity in recalcitrant chromosome regions

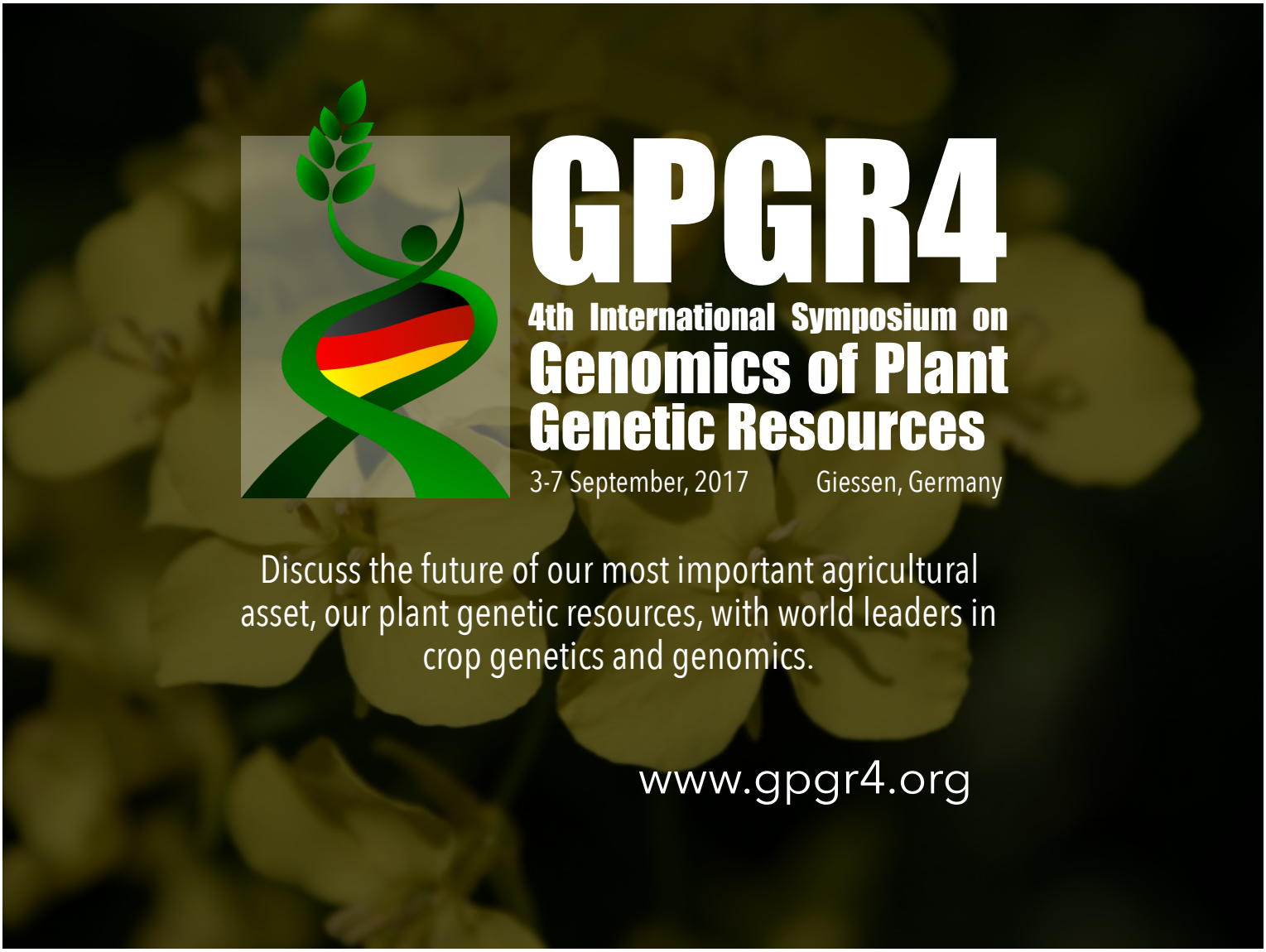
Maintain strong selection progress despite breeding bottlenecks

- Genomic predictions implementing SNP and SNAP data
- Exploitation of quantitative resistances induced by segmental deletions

Department of Plant Breeding, JLU Giessen

Boulos Chalhoub
Regine Delourme
Anne-Marie Chévre

Ian Bancroft


GPGR4
4th International Symposium on
Genomics of Plant
Genetic Resources
3-7 September, 2017
Giessen, Germany
Discuss the future of our most important agricultural asset, our plant genetic resources, with world leaders in crop genetics and genomics.
www.gpgr4.org