Workshop SeqBio 2013

Montpellier, November 25-26 2013

Disentangling homeologous contigs in allo-tetraploid assembly: application to durum wheat

V. Ranwez, Y. Holtz, G. Sarah, M. Ardisson, S. Santoni,
S. Glémin, M. Tavaud, J. David

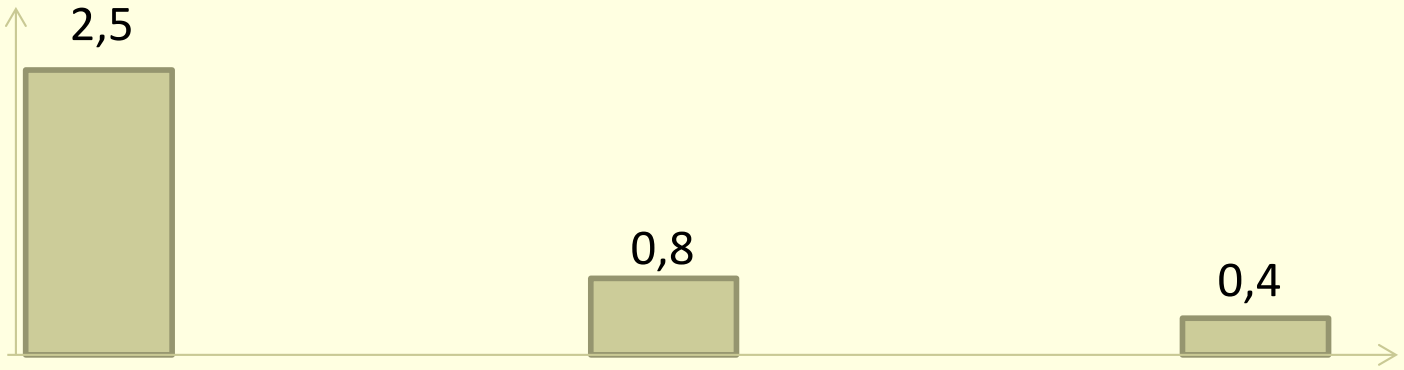
Domestication and diversity

- Domestication & selection drastically reduces diversity

Wild Type:
T. t. dicoccoides

Domestication:
T. t. dicoccum

Durum wheat:
T. t. durum



Genetic diversity estimated through polymorphism π (10^{-3})
(Adapted from Haudry 2008 phD)

Domestication and diversity

- Genetic diversity is important for future selection:

- Pre-breeding project: reintroducing diversity
  - Crossing elites and core collection of *T. turgidum* (wild & domesticated)
  - Controlling outcrossing via a male sterility gene (ms gene)
  - Improving the population via a soft selection (eliminating the weakest)

- After 17 generations
  - ⇨ High phenotypic diversity

- Linking genotype & phenotype
  - ⇨ **Search for SNPs**

- DATA
  - ⇨ RNAseq for 106 accessions (>4 selfing generations)

Durum Wheat Genome: a complex history

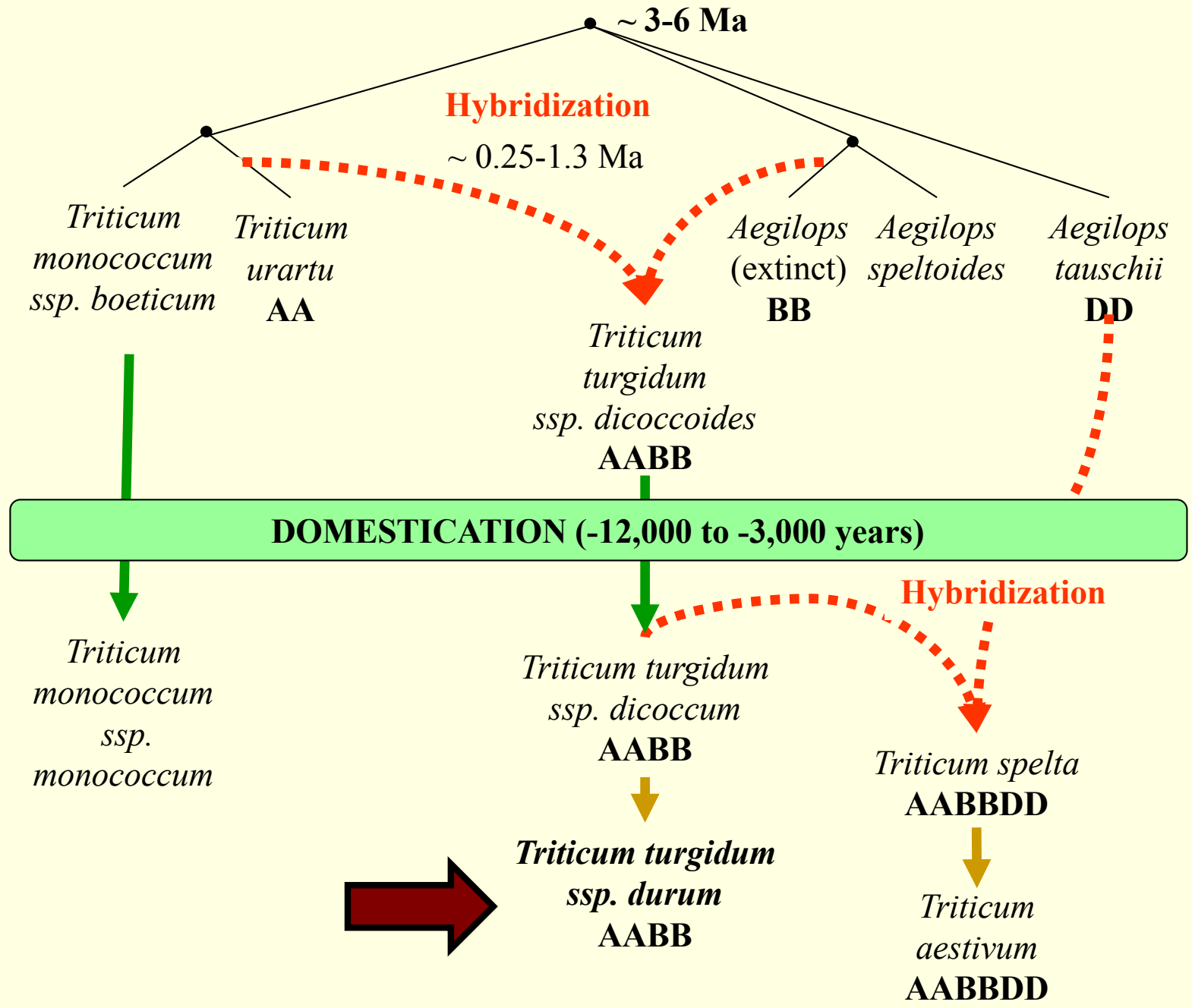
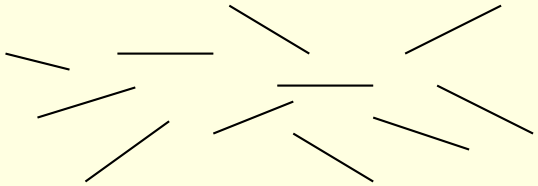


Table of content

Polyploidy induces chimeric contigs

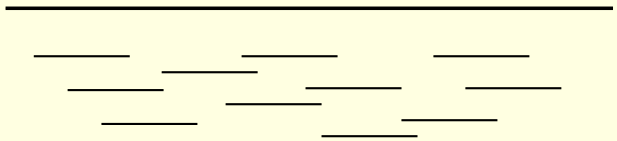
Diploidy

Transcripts of Gene 1



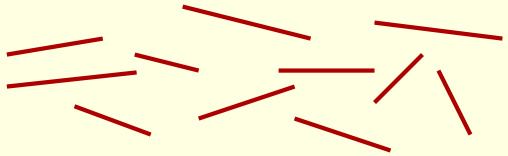
Assembling

Contig of Gene 1

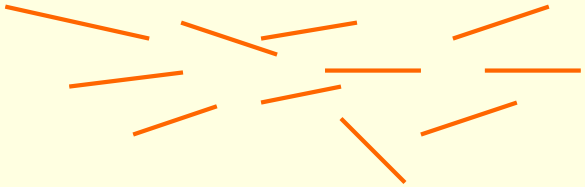


Polyploidy

Transcripts of Gene 1
Genome A

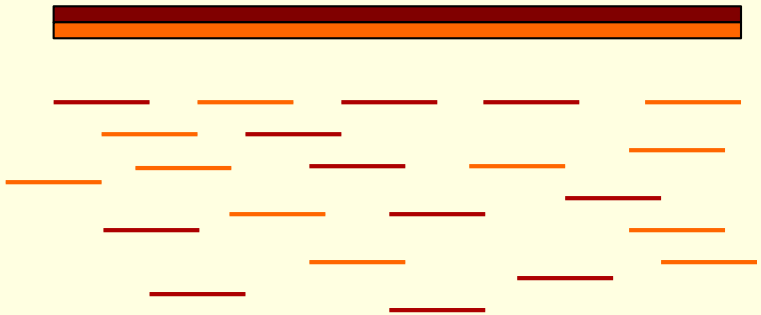


Transcripts of Gene 1
Genome B

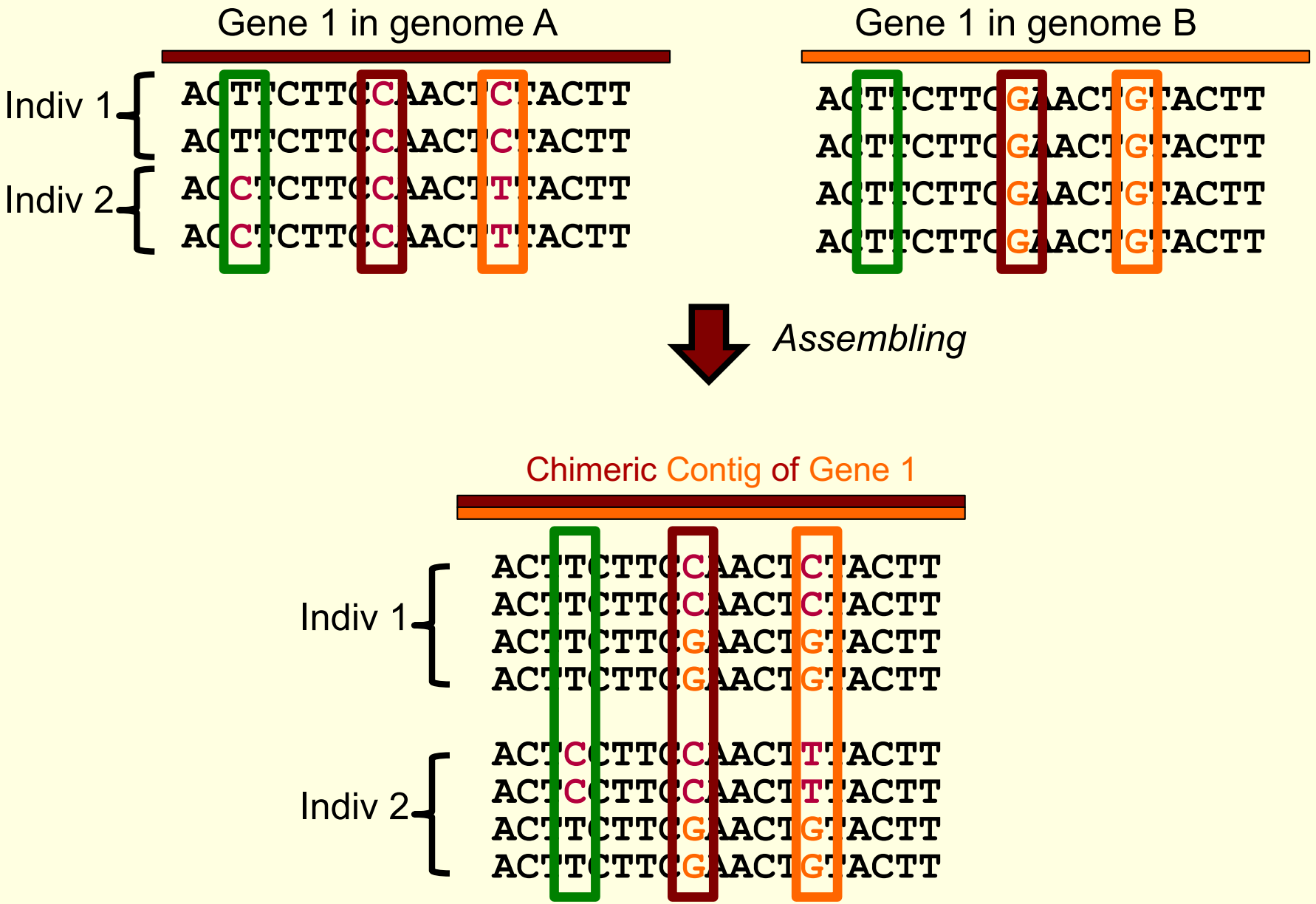


Assembling

Chimeric Contig of Gene 1



chimeric contigs interfere with SNP calling



SNP calling despite polyploidy

- Excess of predicted SNPs
  - How to avoid this problem?

1. By detecting erroneous SNPs
   - Excess of heterozygotes at those sites (but removes true SNPs)
2. By disentangling Homeologous gene copies
   - Implicitly: mapping on diploid relatives
   - Explicitly: splitting problematic contigs

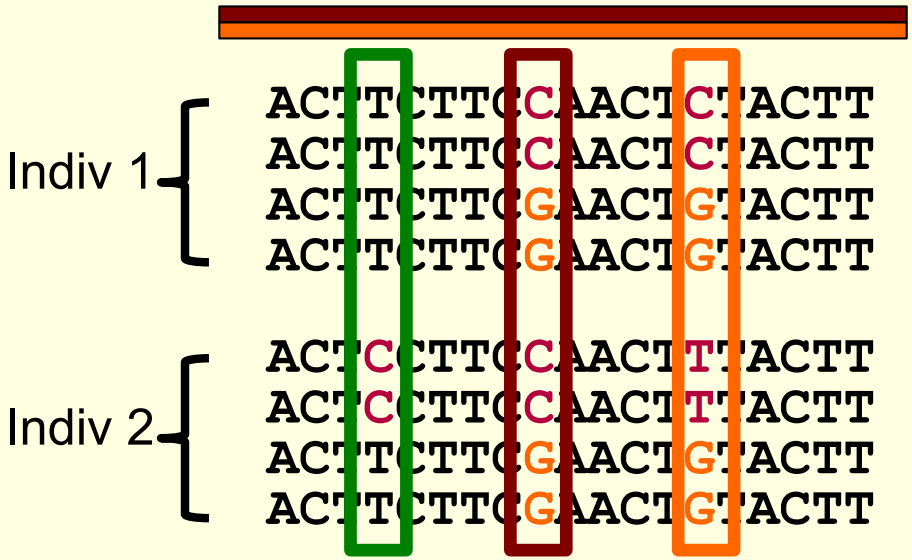
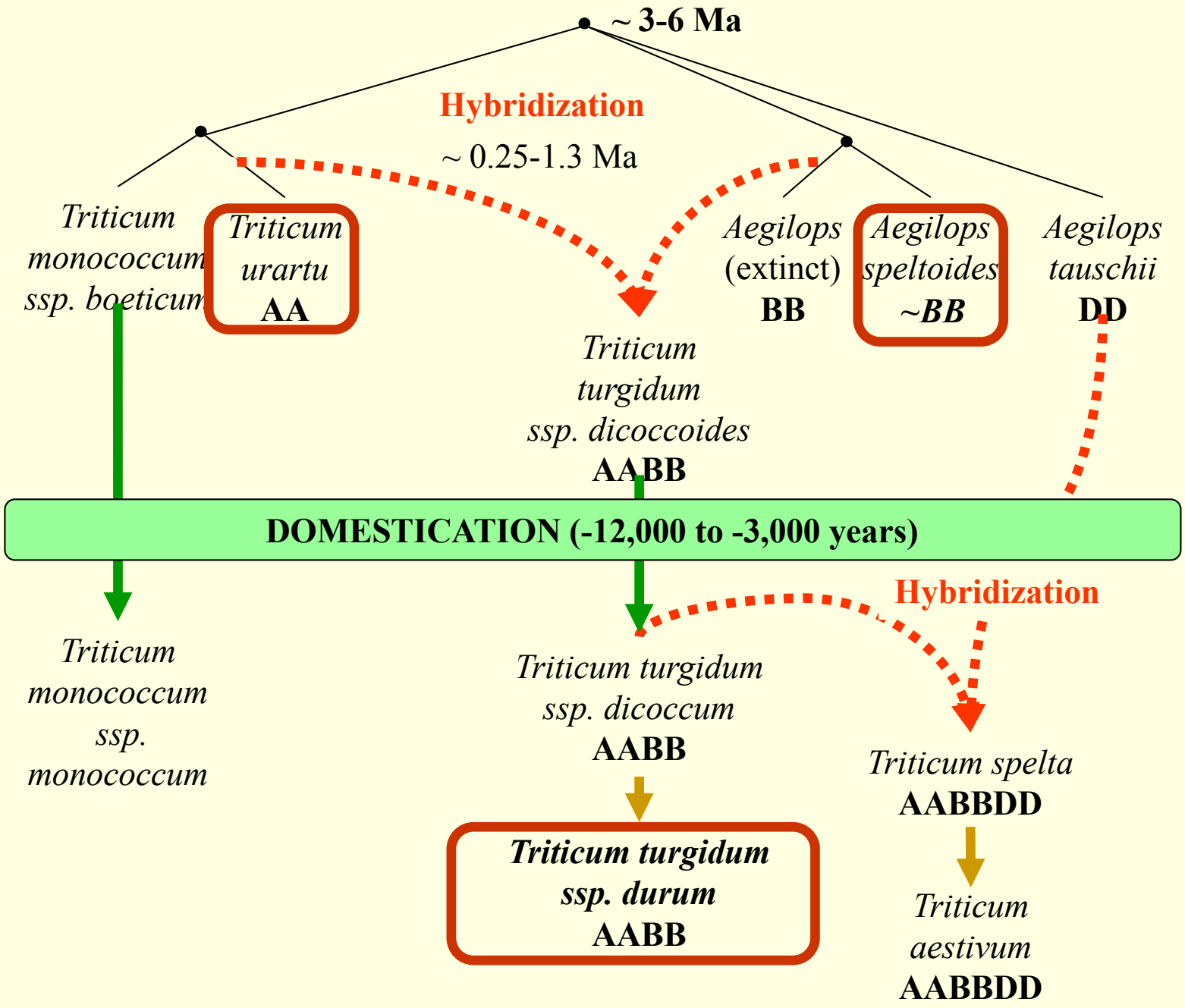


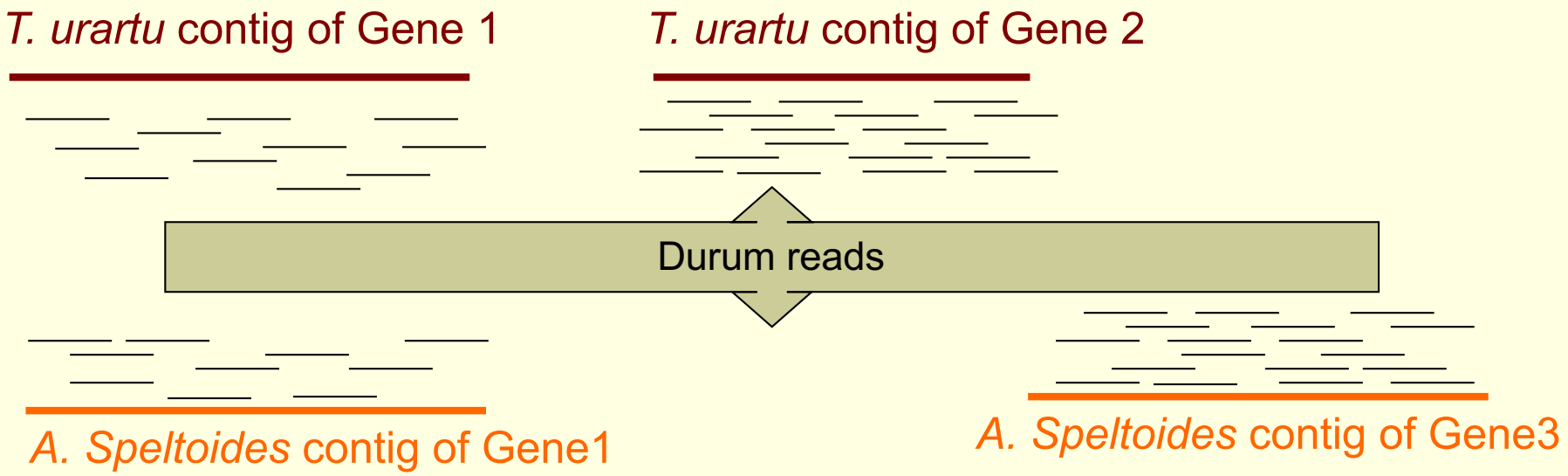
Table of content

Durum Wheat Genome: a complex history

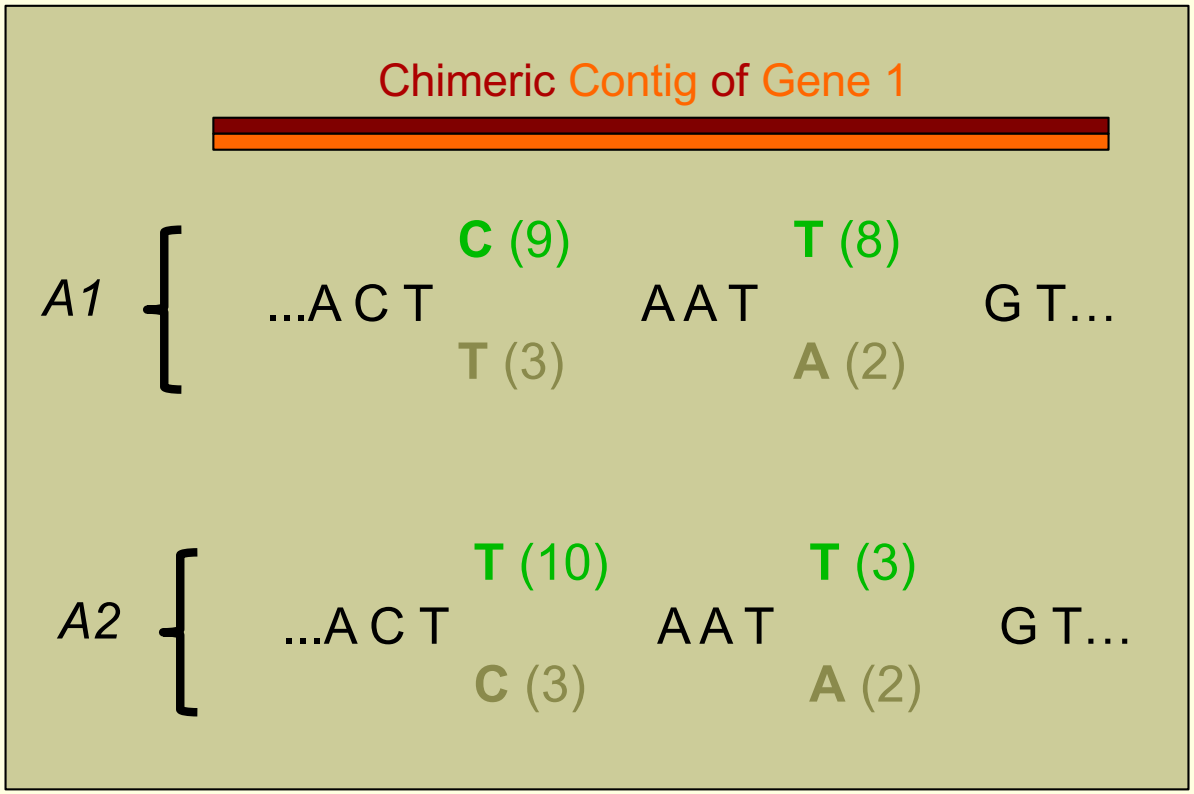


Implicit disentangling: mapping on diploid relatives

- Strategy
  1. Sequencing and assembling contigs of diploids relatives
  2. Mapping *T. durum* reads on those contigs
- Advantage:
  - ⇨Avoid confusion when contigs are available for both relatives
- Limits:
  - ⇨No advantage if the gene is available for only one relative
  - ⇨Diploids are only relatives, a strict mapping cannot be used



Explicit disentangling: model



Likelihood of CT?

➔Associated expression ratio
r_{1CT} in A1: (9/12 +8/10)/2= 0.76
r_{2CT} in A2: (3/13 +3/5)/2 ~0.415

➔Probability for one site:

$$P(9C \mid 12\,reads, r_{1CT}) = \binom{12}{9} r_{1CT}^{9} (1 - r_{1CT})^{3}$$

➔Likelihood of CT:

$$Lk(D \mid CT) = \prod_{Ai} P(D_{Ai} \mid r_{iCT})$$

➔Probability for one accession:

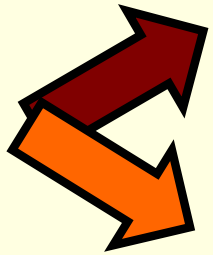
$$P(D_{A1} \mid r_{1CT}) = P(9C \mid 12\,reads, r_{1CT}) * P(8T \mid 10\,reads, r_{1CT})$$

Explicit disentangling: validation

- Once the most likely pattern (e.g. CT) is found the contig sequence is split into two new contigs

$Contig_{lk1}$= **ACTTTCCTACTGGCAACAGTAACG**

$Contig_{Orig}$= ACTTTCCTA C/T TGGCAACAG A/T CAACG



$Contig_{lk2}$= **ACTTTCCTATTGGCAACAGAAACG**

- Strategy

1. Assemble *T. durum* reads (and create associated mapping)
2. For each contig with an excess of heterozygosity
   1. Search for likely homeologous contigs within those reads
   2. Replace current contig by the two predicted homeologous ones
3. Map *T. durum* reads on this new set of contigs
4. Use standard SNP calling on this mapping

Explicit disentangling: overview

- Advantages:
  - Avoid most of the confusion caused by homeologous genes
  - No need to have diploid relatives (except for validation)

- Current limits:
  - Slower than mapping on diploid relatives (two mappings)
  - Less effective if homeologous genes are similarly expressed

Table of content

Explicit disentangling: validation

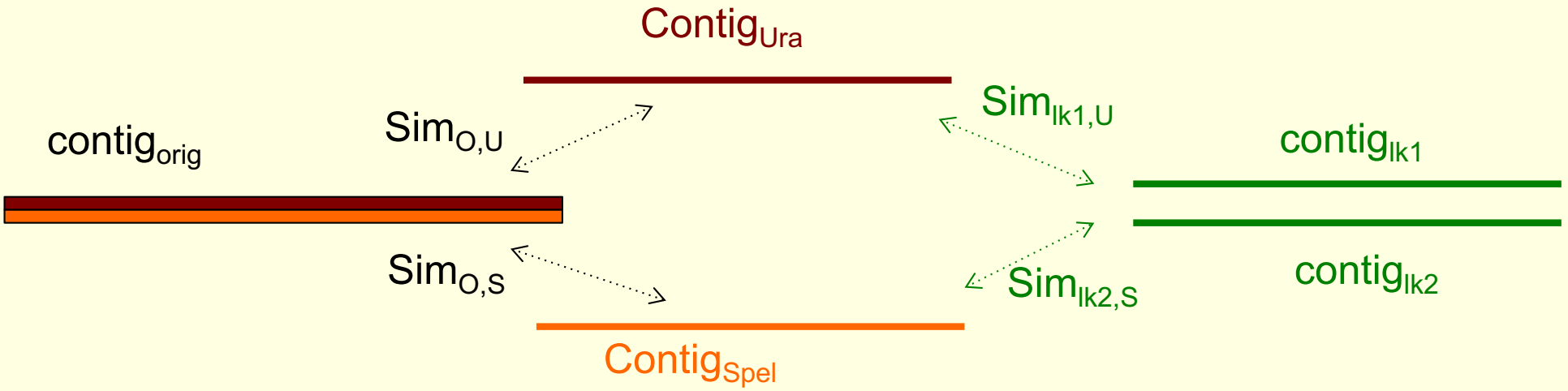
- Validation using diploids relatives

  - **Focusing on 3709** $Contig_{Orig}$ for which we found
    1 homologous Urartu contig: $Contig_{Ura}$
    1 homologous Speltoides contig: $Contig_{Spe}$
    0 or 1 other homologous $Contig_{Orig}$

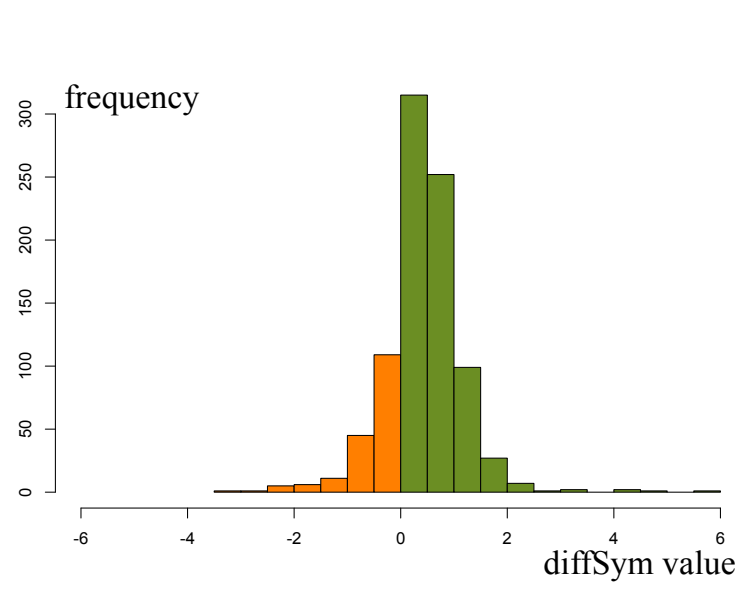
  - **3083 contigs with at least one questionable site**

  - **New contigs in 967** cases, if the method works they should be more similar to $Contig_{Ura}$ and $Contig_{Spe}$

Explicit disentangling: validation

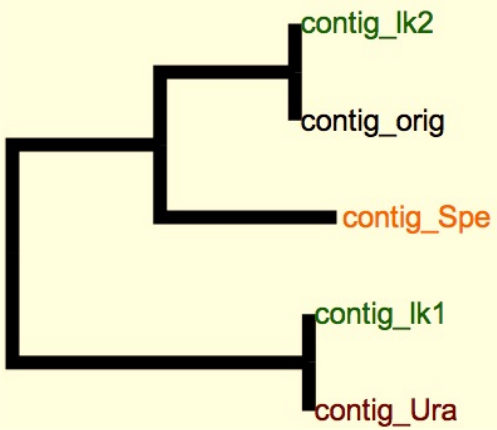


Validation: diffSym=(Sim$_{lk1,U}$ + Sim$_{lk2,S}$) – (Sim$_{O,U}$ + Sim$_{O,S}$)

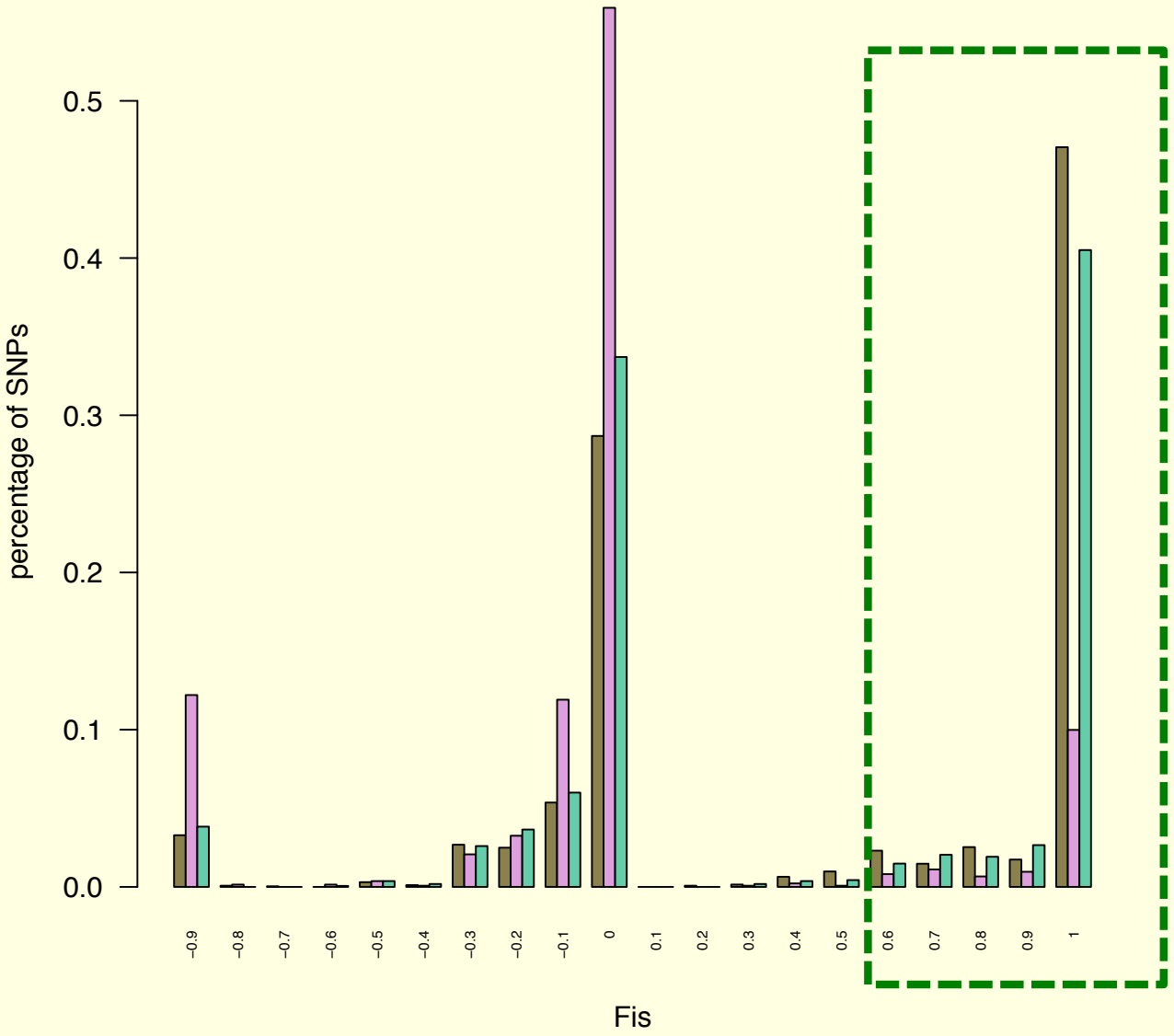


Significant improvement:

- paired t-test: p-value < 2.2e-16
- Average diffSym 0.43



Validation using FIS



Diploid relatives

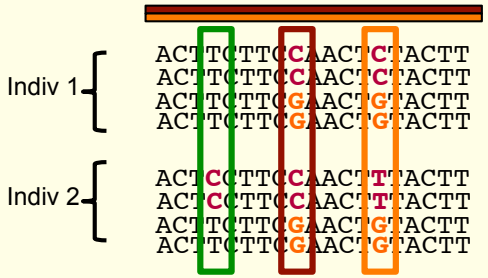
De novo

HomeoSplitter

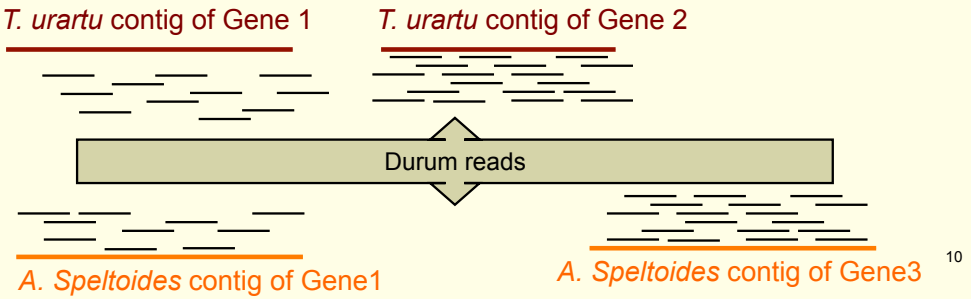
$$F_{is} = 1 - \frac{H_{obs}}{H_e}$$

Expected F_{is} for durum Wheat

- Polyploidy should not be ignored during SNP calling

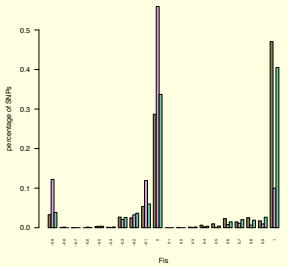


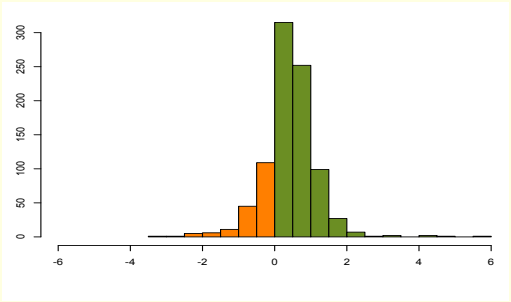
- Mapping on diploid relatives helps but is insufficient

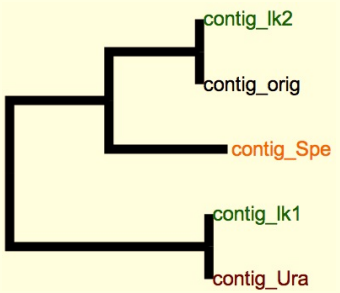


10

- Explicitly disentangling homeologous contigs seems to be a very promising solution

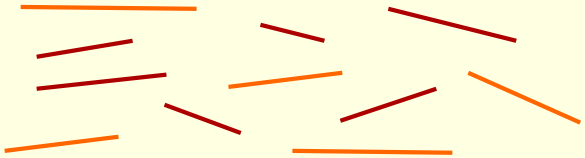






Current Pipeline

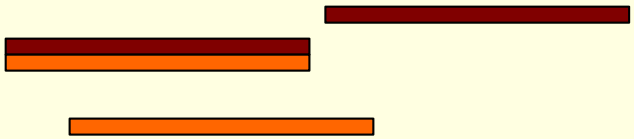
Transcripts of the EPO population



About 2 billions reads for 106 accessions

Assembling

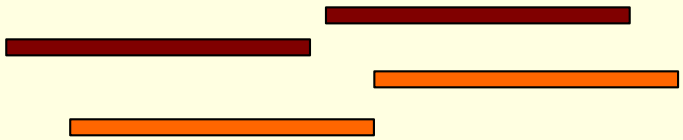
Consensus Assembly



70000 Contigs (with Chimeric)

Mapping + HoméoSplitter

New Consensus Assembly



80000 clean Contigs

Mapping + SNP Calling

About 90000 SNPS !

Disentangling homeologous contigs in allo-tetraploid assembly: application to durum wheat.

V Ranwez, Y Holtz, G Sarah, M Ardisson, S Santoni, S Glémin, M Tavaud-Pirra. BMC Bioinformatics 14 (Suppl 15), S15 (RECOMB-CG 2013 special issue)

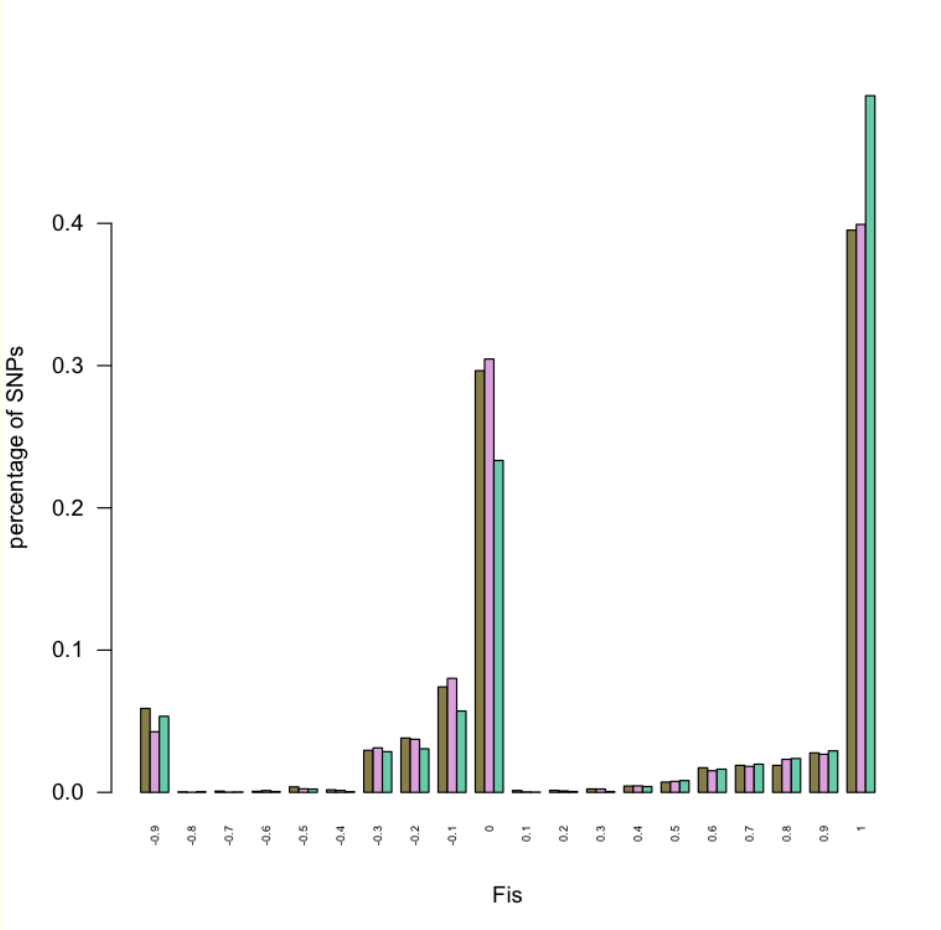
http://bioweb.supagro.inra.fr/homeoSplitter/

Thank you for your attention

Our reads on *Krasileva et al 2013 de novo contigs*

Our reads on *Krasileva et al 2013 de novo contigs split by their phasing*

Our reads on *Krasileva et al 2013 de novo contigs split by HomeoSplitter*



$$F_{is} = 1 - \frac{H_{obs}}{H_e}$$