Bioinformatics tools and resources to study plant genomes

Lieven Sterck

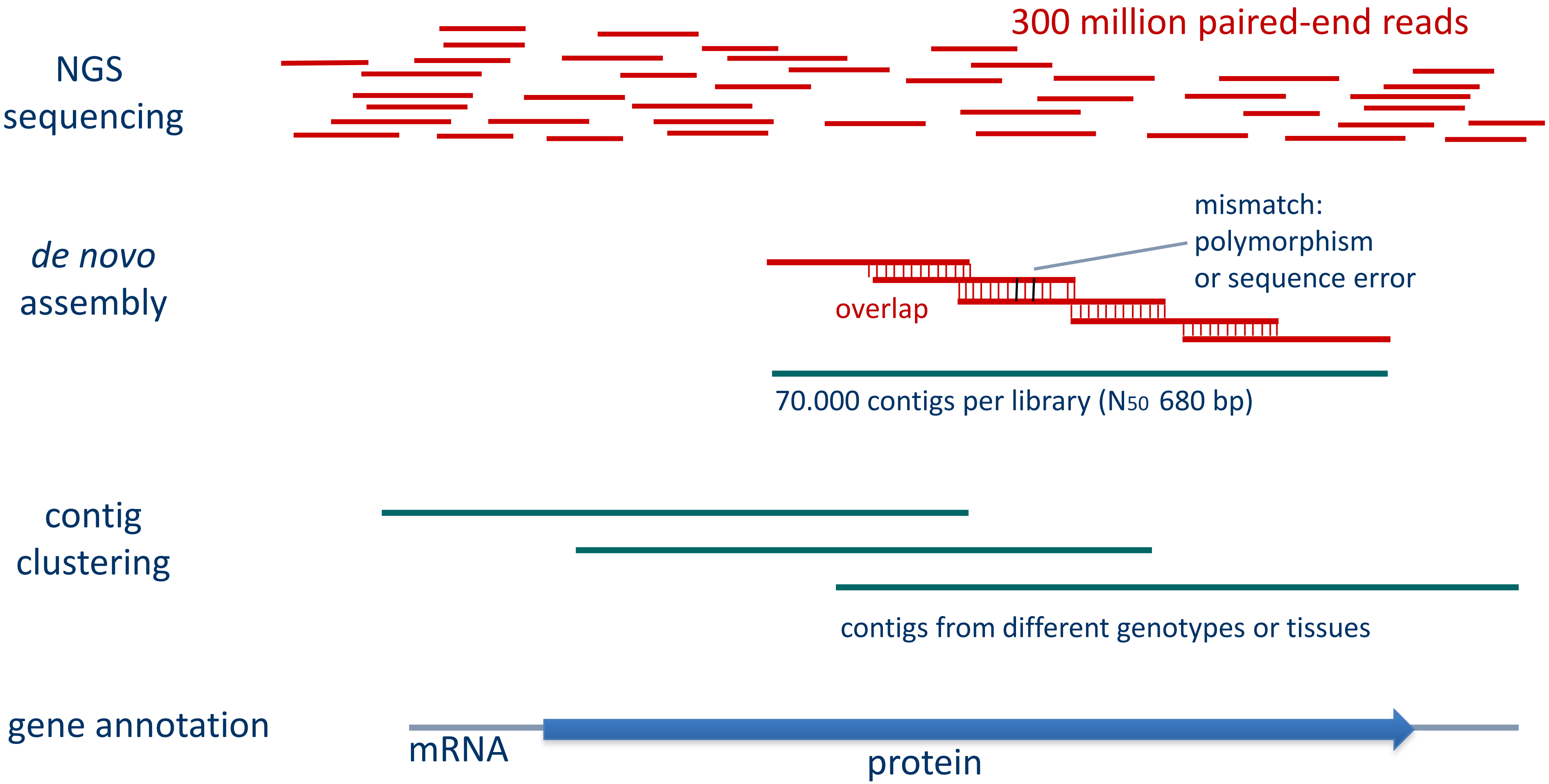
Bioinformatics & Systems Biology, VIB, Belgium

GHENT UNIVERSITY

VIB

OGA : a tool to analysing heterozygous crops

De novo assembly of transcript sequences: RNA-seq in 14 *Lolium perenne* genotypes



NGS sequencing

300 million paired-end reads

de novo assembly

overlap

mismatch:
polymorphism
or sequence error

70.000 contigs per library (N_{50} 680 bp)

contig clustering

contigs from different genotypes or tissues

gene annotation

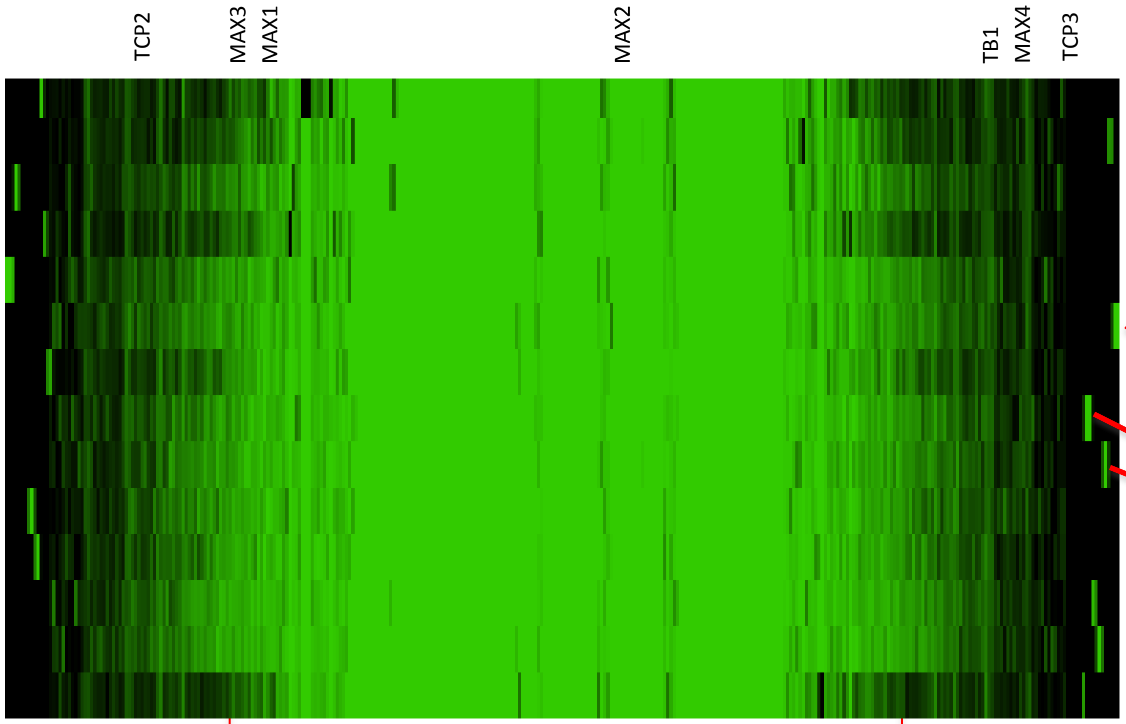
mRNA

protein

overlap between transcriptomes of 14 Lp genotypes

reciprocal Blast of all 26.552 *Brachypodium* genes to contigs of *Lp* cDNA libraries

# genes / cDNA lib	
10.465	Lp12
10.989	Lp13
11.800	Lp03
11.042	Lp11
12.529	Lp09
12.540	Lp07
11.962	Lp10
12.605	Lp16
12.514	Lp19
12.508	Lp20
12.377	Lp25
12.723	Lp27
12.757	Lp23
10.831	Lp14

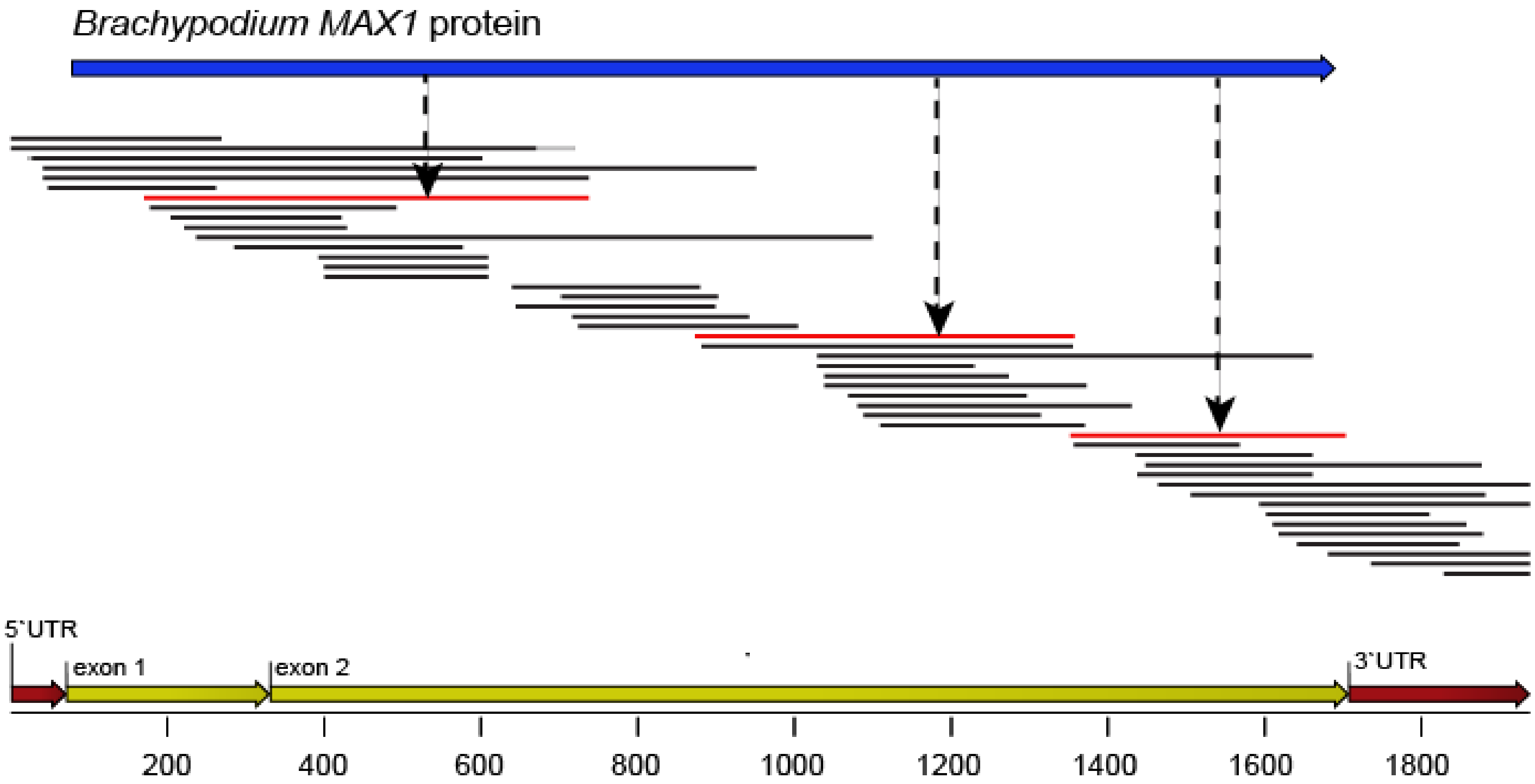


genes represented in single genotype

genes represented in many genotypes

any library contains only 65% of all expressed genes

Assembling all genotypes at once



consensus reference transcript

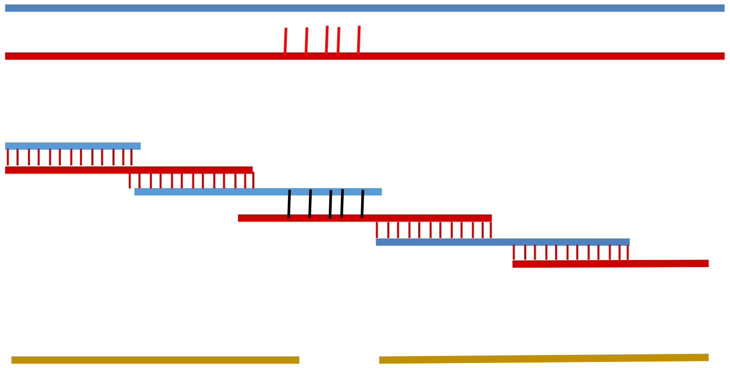
choose your reference sequence:
1 genotype is not optimal

libraries are fragmented:

low coverage or

short stretch of SNPs breaks contigs:

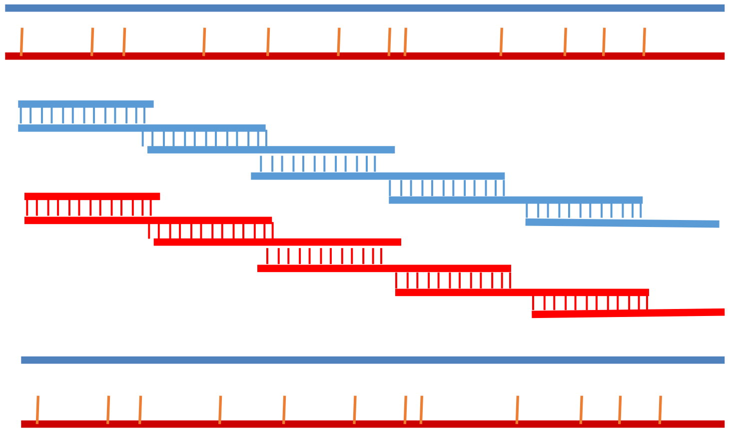
3-4 fragments per gene



libraries are redundant:

long stretch of SNPs: 2 different sequences

two alleles in two contigs



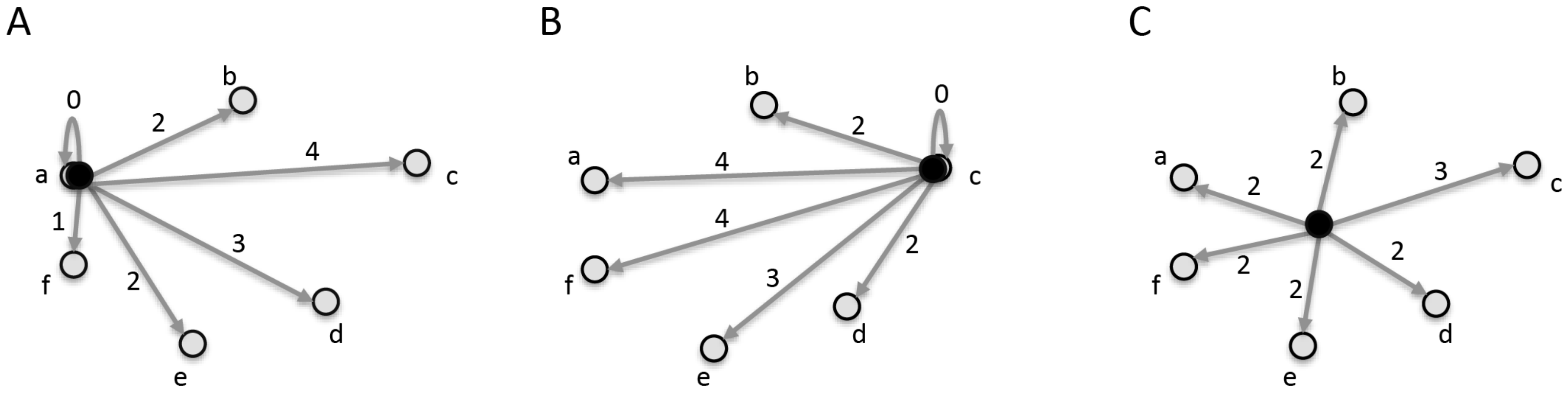
depends on the allele pair:

fragmentation is different for each genotype

libraries are incomplete: any library only has 65% of all expressed genes

choose your reference sequence: a consensus of all alleles is better

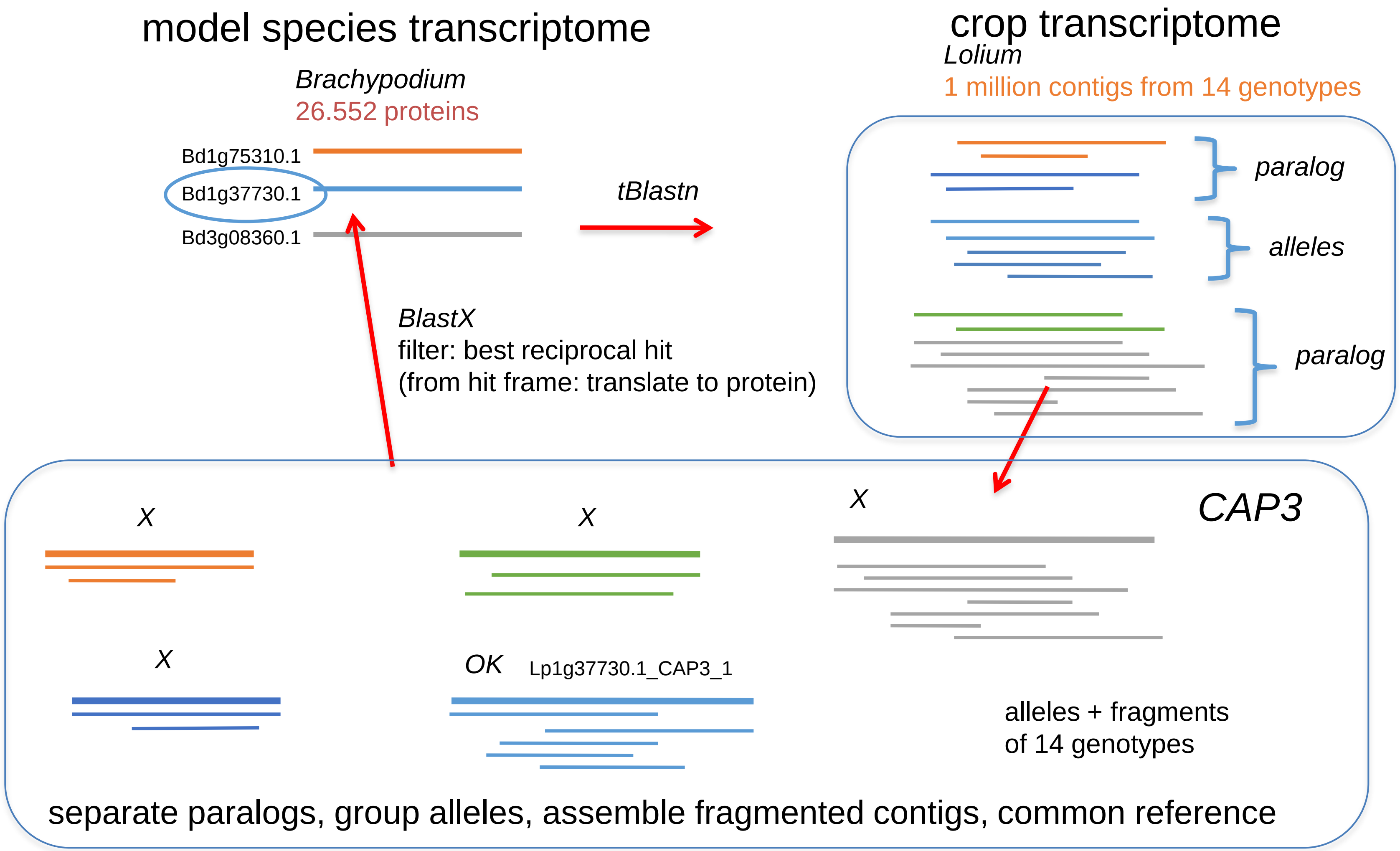
SNP identification depends on genetic distance between reference and allele:
threshold for number of mismatches



in ‘sequence space’ , the consensus of all alleles is somewhere in the middle

create reference with equal distance to all alleles
balanced chance to detect all SNPs at relatively **higher** stringency

“reference guided” construction of *Lp* transcriptome



model species transcriptome

Brachypodium

26.552 proteins

Bd1g75310.1

Bd1g37730.1

Bd3g08360.1

tBlastn

crop transcriptome

Lolium

1 million contigs from 14 genotypes

paralog

alleles

paralog

BlastX

filter: best reciprocal hit

(from hit frame: translate to protein)

CAP3

X

X

X

X

OK Lp1g37730.1_CAP3_1

alleles + fragments

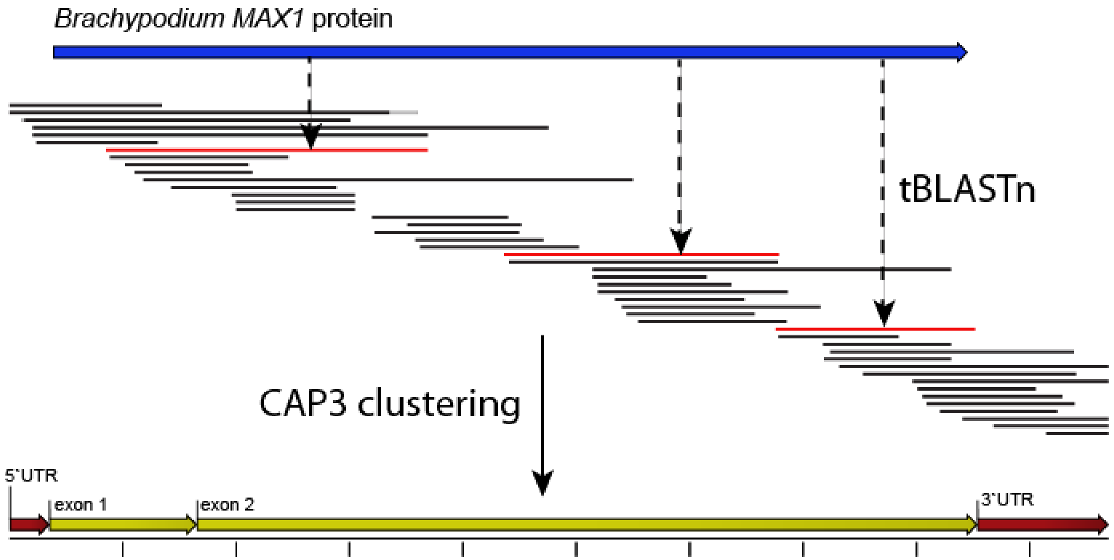
of 14 genotypes

separate paralogs, group alleles, assemble fragmented contigs, common reference

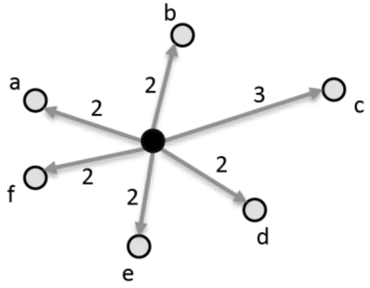
create your reference sequence:
cluster all alleles, gene-by-gene

select all alleles from all genotypes, align, and take the consensus
overlaps fragmented sequences, solves the redundancy in alleles

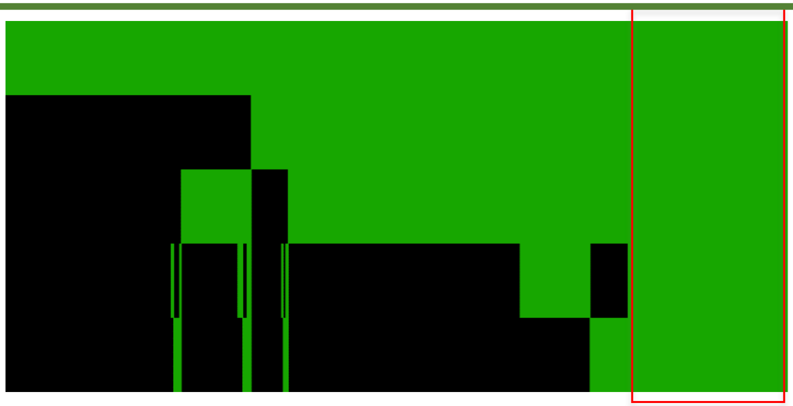
gene-by-gene: use all proteins of the closest related complete genome as guidance



consensus reference transcript



interspecific markers in *Lolium, Festuca, Phleum*



26.552 *Brachypodium* genes

18.213 *Lolium* genes (6 tissues)   (Torben Asp, U-Aarhus)

19.329 *Lolium* genes (14 genotypes)

8.413 *Festuca* genes (2 genotypes) (Simen Sandve, UMB)

7.399 *Phleum* genes (2 genotypes) (Simen Sandve, UMB)

5.426 genes in all 4 species

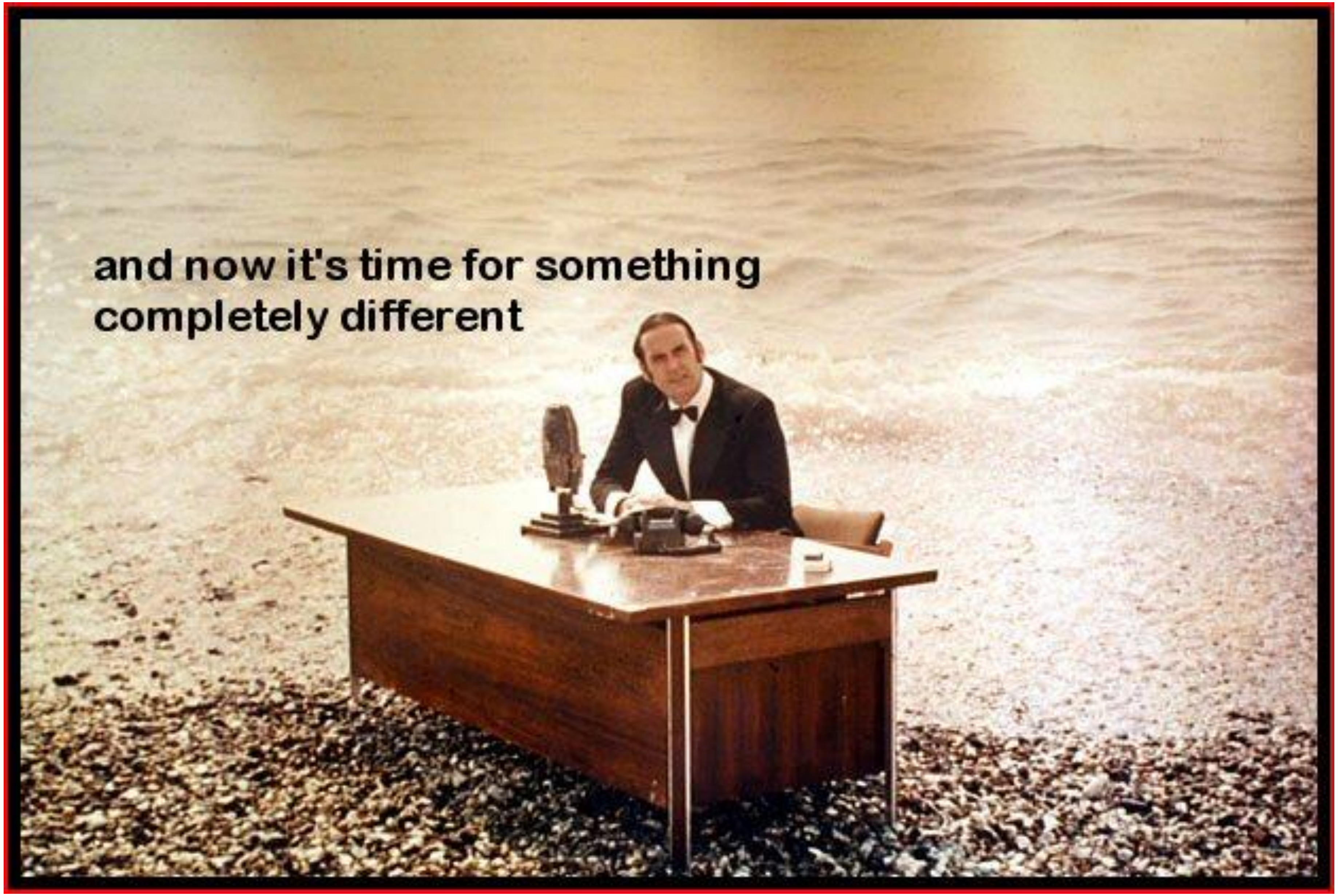
```
Bradi1g03160.1 AAAATACTGAAGAAGTATGACAAAAGAACAGGCGGATTGCTCAGCCTACCTTTCACTCAACGTGCTCGCCACGAACCATTTTTCACCACTGAACCCTTAACAAGGCTTGTTCGAGAATGTGAGGCTA

     Consensus AAGATACTNAANAAGTANGACAAAAGAACAGGNGGCGTGCTCAGCCTACCNTTCACNCAACGNGCTCGNCACNAACCATTNTTCACCACNGAACCNTTAANAAGGCTTGTTNGAGAATGTGAGGNTA
      Lolium_1 AAGATACTGAAGAAGTATGACAAAAGAACAGGCGGCGTGCTCAGCCTACCTTTCACTCAACGCGCTCGTCACGAACCATTCTTCACCACTGAACCCTTAAGAAGGCTTGTTAGAGAATGTGAGGTTA
      Lolium_2 AAGATACTGAAGAAGTATGACAAAAGAACAGGCGGCGTGCTCAGCCTACCTTTCACTCAACGCGCTCGTCACGAACCATTCTTCACCACTGAACCCTTAAGAAGGCTTGTTAGAGAATGTGAGGTTA
      Lolium_3
     Festuca_1 AAGATACTCAAGAAGTATGACAAAAGAACAGGCGGCGTGCTCAGCCTACCTTTCACTCAACGCGCTCGTCACGAACCATTCTTCACCACAGAACCCTTAAGAAGGCTTGTTAGAGAATGTGAGGTTA
      Phleum_1 AAGATACTCAAGAAGTACGACAAAAGAACAGGTGGCGTGCTCAGCCTACCATTCACCCAACGTGCTCGGCACCAACCATTTTTCACCACTGAACCTTTAACAAGGCTTGTTCGAGAATGTGAGGCTA
      Phleum_2 AAGATACTCAATAAGTACGACAAAAGAACAGGTGGCGTGCTCAGCCTACCATTCACCCAACGTGCTCGGCACCAACCATTTTTCACCACTGAACCTTTAACAAGGCTTGTTCGAGAATGTGAGGCTA
      Phleum_3 AAGATACTCAAGAAGTACGACAAAAGAACAGGTGGCGTGCTCAGCCTACCATTCACCCAACGTGCTCGGCACCAACCATTTTTCACCACTGAACCTTTAACAAGGCTTGTTCGAGAATGTGAGGCTA
```

Phleum intraspecific marker

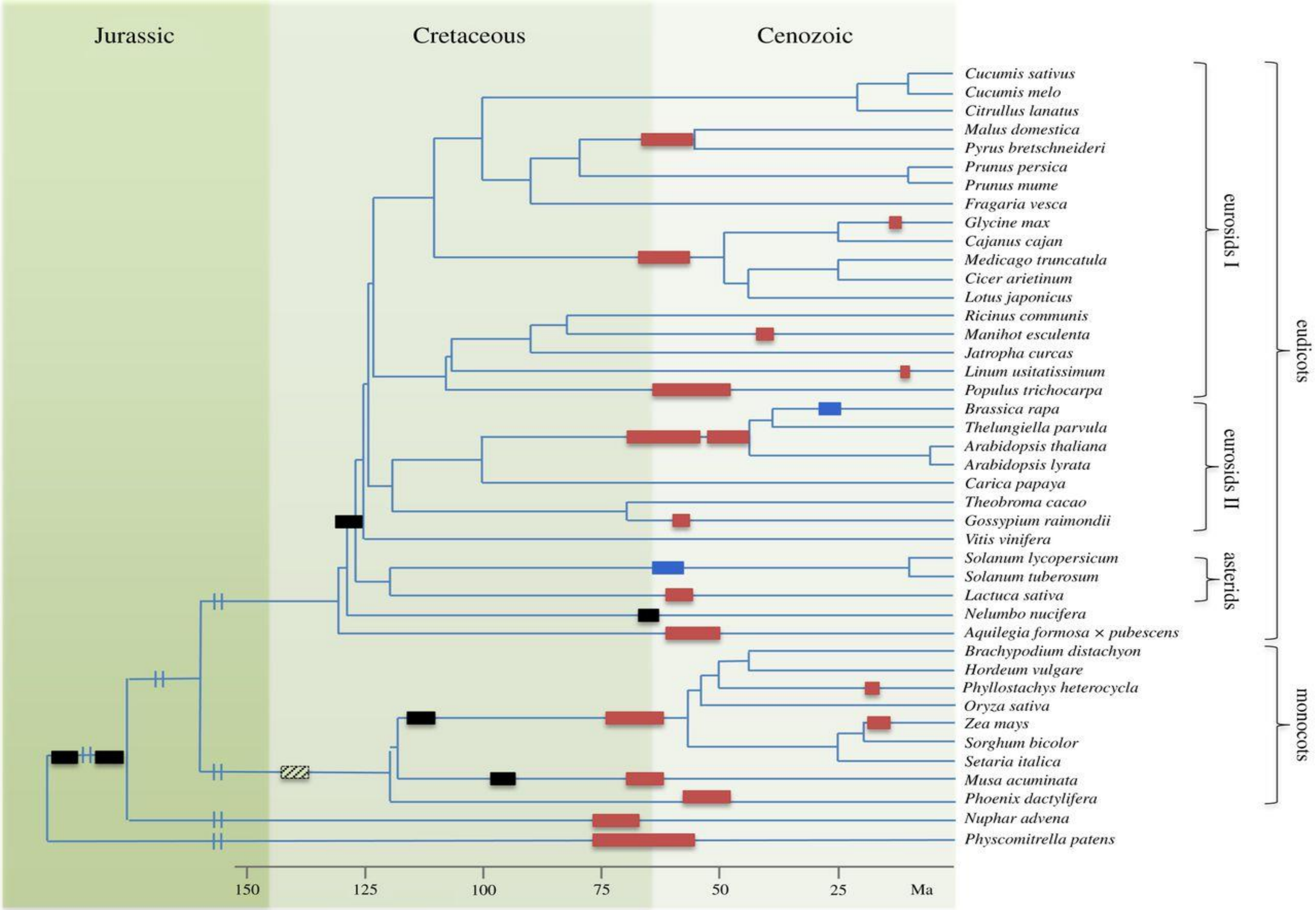
Phleum interspecific marker

Festuca interspecific marker

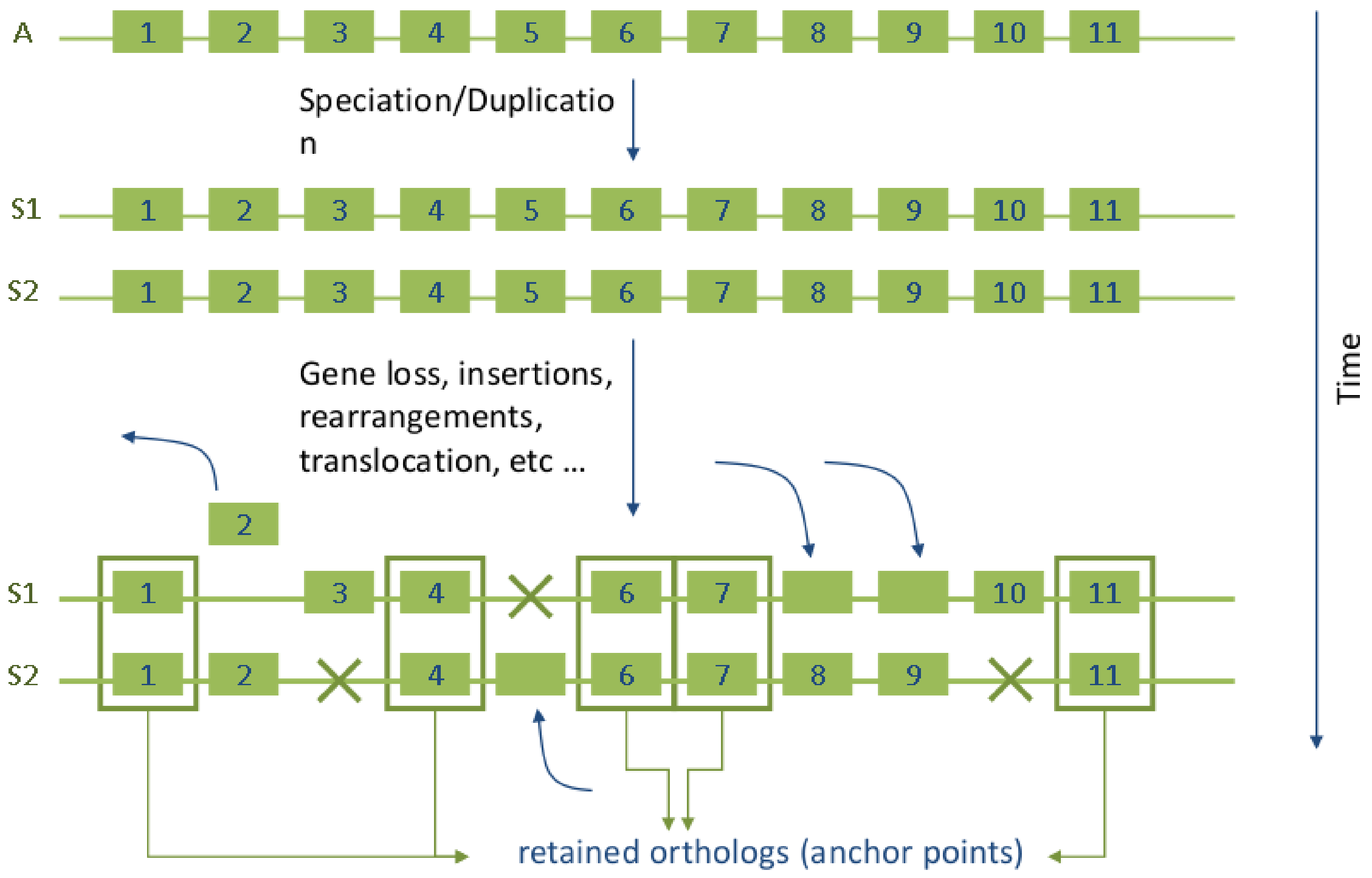


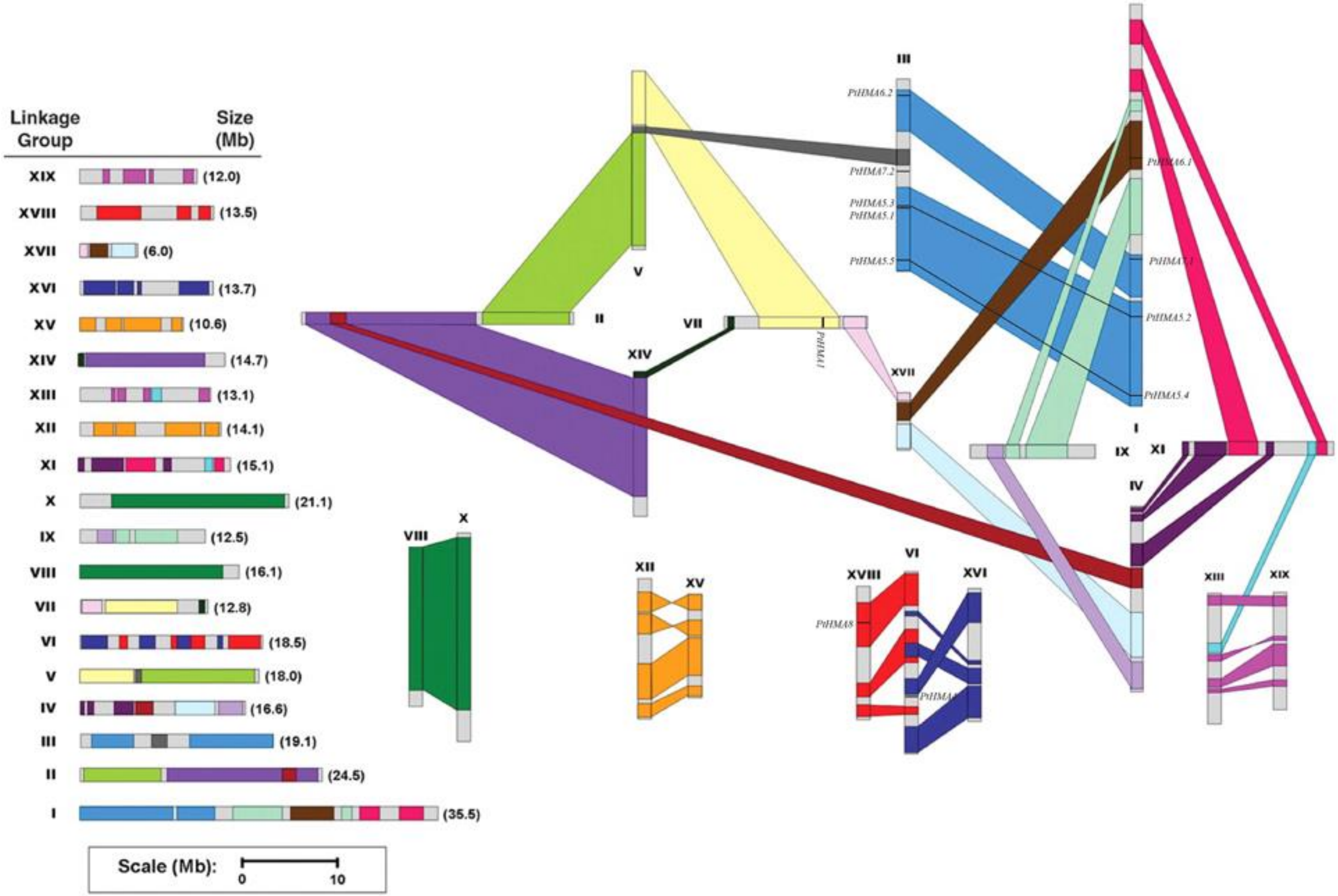
and now it's time for something completely different

Genome duplications are wide spread in plants



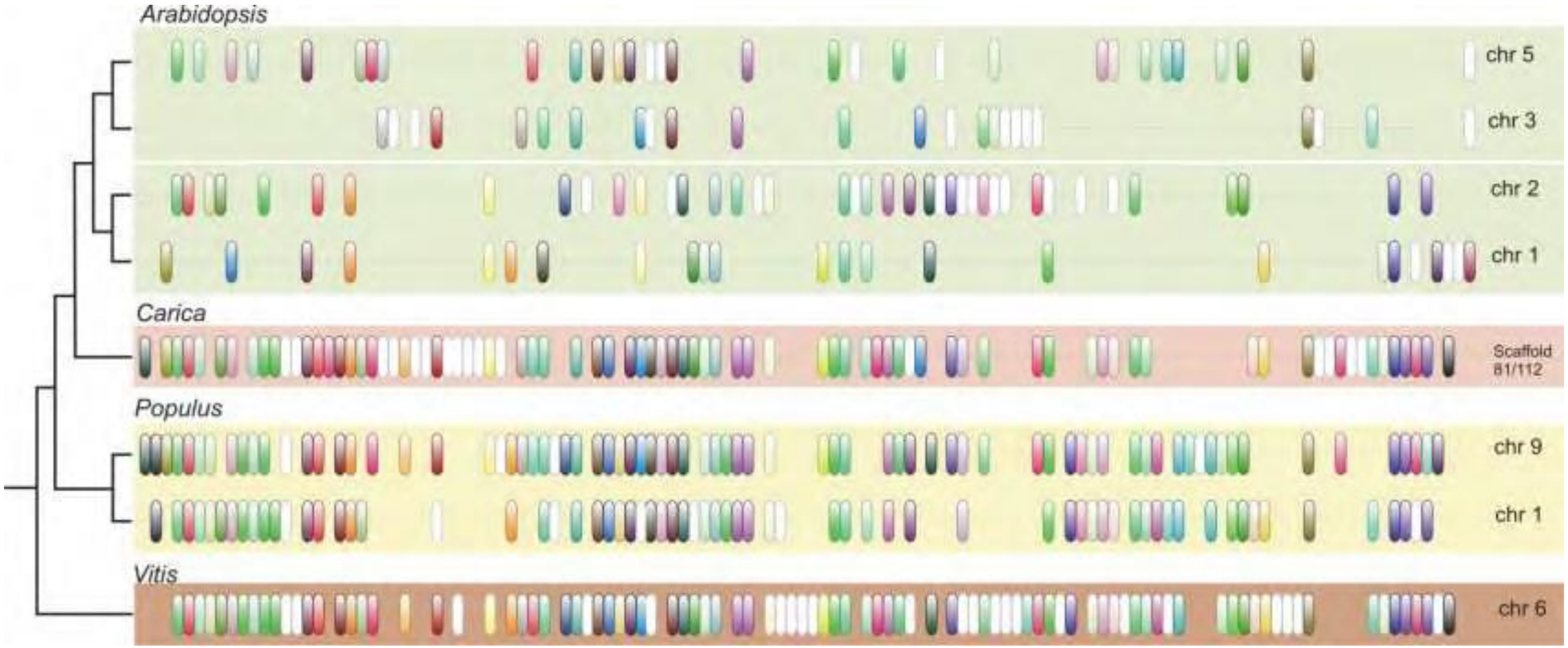
Detecting colinearity and large-scale gene duplications



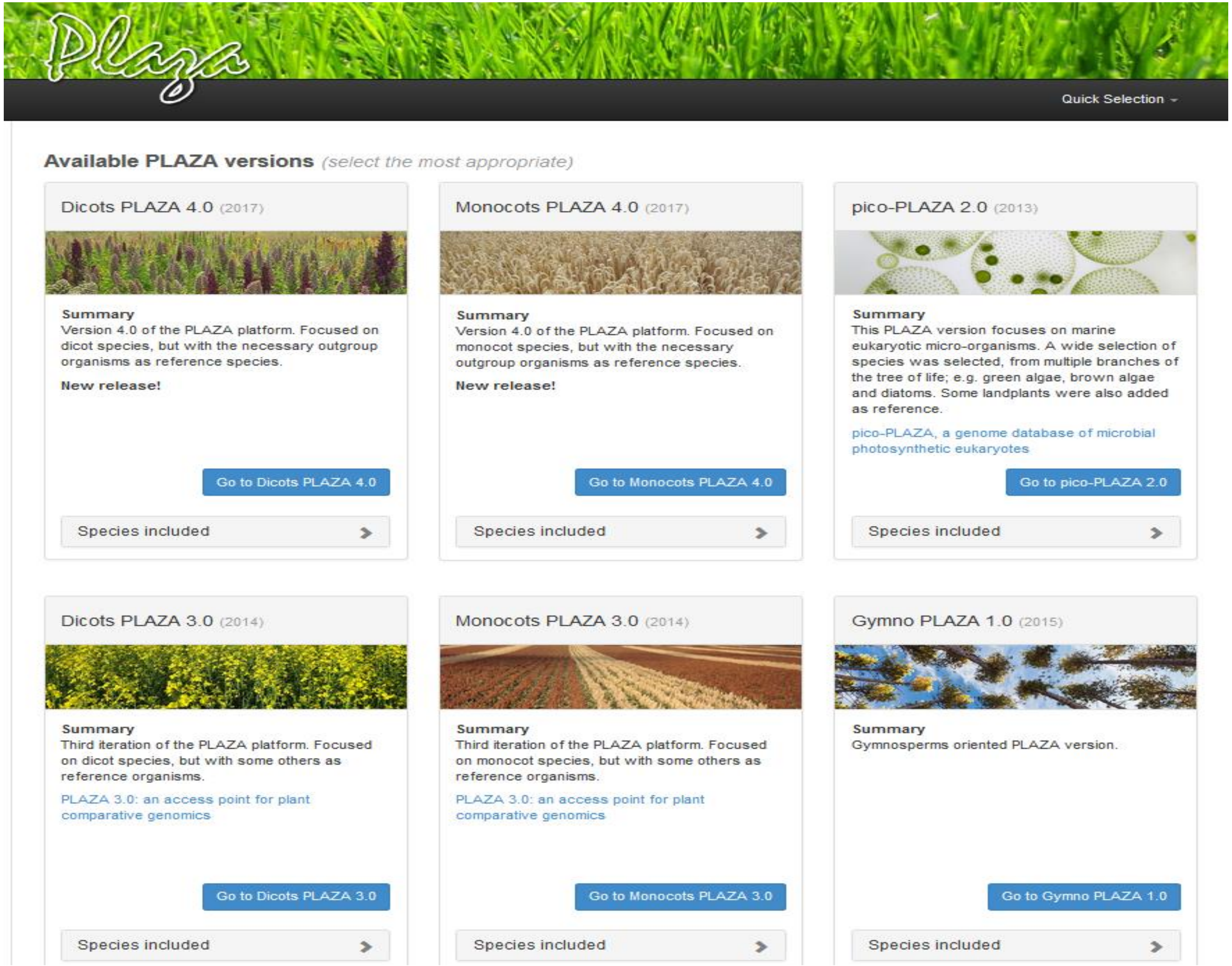

Linkage Group
Size (Mb)
XIX (12.0)
XVIII (13.5)
XVII (6.0)
XVI (13.7)
XV (10.6)
XIV (14.7)
XIII (13.1)
XII (14.1)
XI (15.1)
X (21.1)
IX (12.5)
VIII (16.1)
VII (12.8)
VI (18.5)
V (18.0)
IV (16.6)
III (19.1)
II (24.5)
I (35.5)
Scale (Mb): 0 10
V
II
VII
XIV
III
PtHMA6.2
PtHMA7.2
PtHMA5.3
PtHMA5.1
PtHMA5.5
PtHMA1
XVII
PtHMA6.1
PtHMA7.1
PtHMA5.2
PtHMA5.4
I
IX
XI
IV
VIII
X
XII
XV
XVIII
VI
XVI
PtHMA8
PtHMA
XIII
XIX

Gene loss and colinearity/synteny

WGDs add to the complexity of plant genomes (both in numbers of genes and genome structure) …



PLAZA: a resource to analysing plant genomes



Available PLAZA versions *(select the most appropriate)*

Dicots PLAZA 4.0 (2017)

Summary
Version 4.0 of the PLAZA platform. Focused on dicot species, but with the necessary outgroup organisms as reference species.

New release!

Monocots PLAZA 4.0 (2017)

Summary
Version 4.0 of the PLAZA platform. Focused on monocot species, but with the necessary outgroup organisms as reference species.

New release!

pico-PLAZA 2.0 (2013)

Summary
This PLAZA version focuses on marine eukaryotic micro-organisms. A wide selection of species was selected, from multiple branches of the tree of life; e.g. green algae, brown algae and diatoms. Some landplants were also added as reference.

pico-PLAZA, a genome database of microbial photosynthetic eukaryotes

Dicots PLAZA 3.0 (2014)

Summary
Third iteration of the PLAZA platform. Focused on dicot species, but with some others as reference organisms.

PLAZA 3.0: an access point for plant comparative genomics

Monocots PLAZA 3.0 (2014)

Summary
Third iteration of the PLAZA platform. Focused on monocot species, but with some others as reference organisms.

PLAZA 3.0: an access point for plant comparative genomics

Gymno PLAZA 1.0 (2015)

Summary
Gymnosperms oriented PLAZA version.

http://bioinformatics.psb.ugent.be/plaza/

Exploiting cross-species genome information

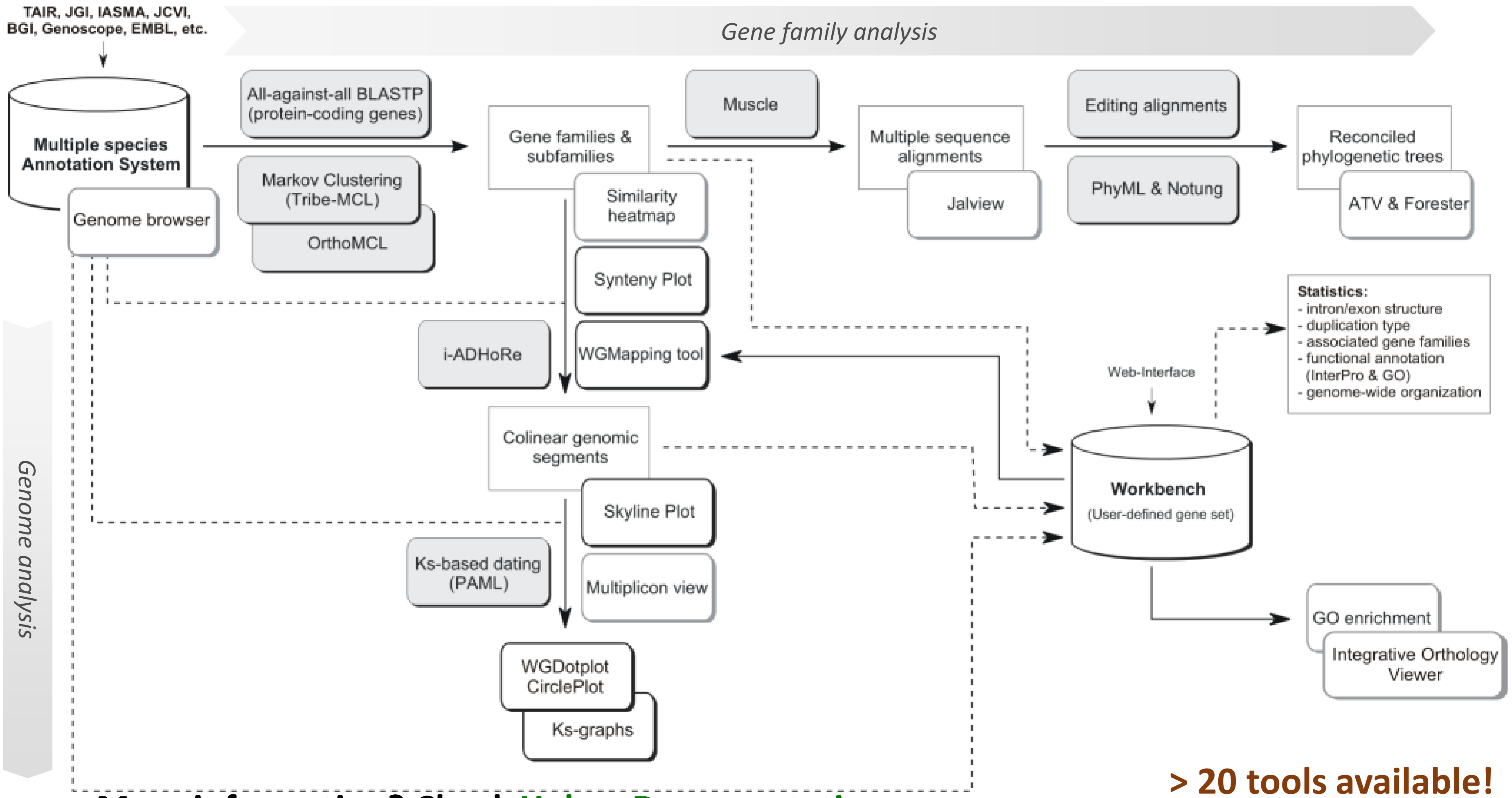
- **Centralized infrastructure**

- Detailed **gene catalog** per species
  - Structural annotation (gene models, UTRs)
  - Functional annotation (experimental, sequence-based)

- **Intuitive** & advanced data mining **tools for non-expert** users
    - Gene function
    - Genome organization
    - Pathway evolution
    - Data manipulation

- **Computational resources**

PLAZA : a versatile toolbox for plant genomics



> 20 tools available!

More information? Check Help – Documentation

- **Data content & Construction**
- **Tutorial & FAQ**

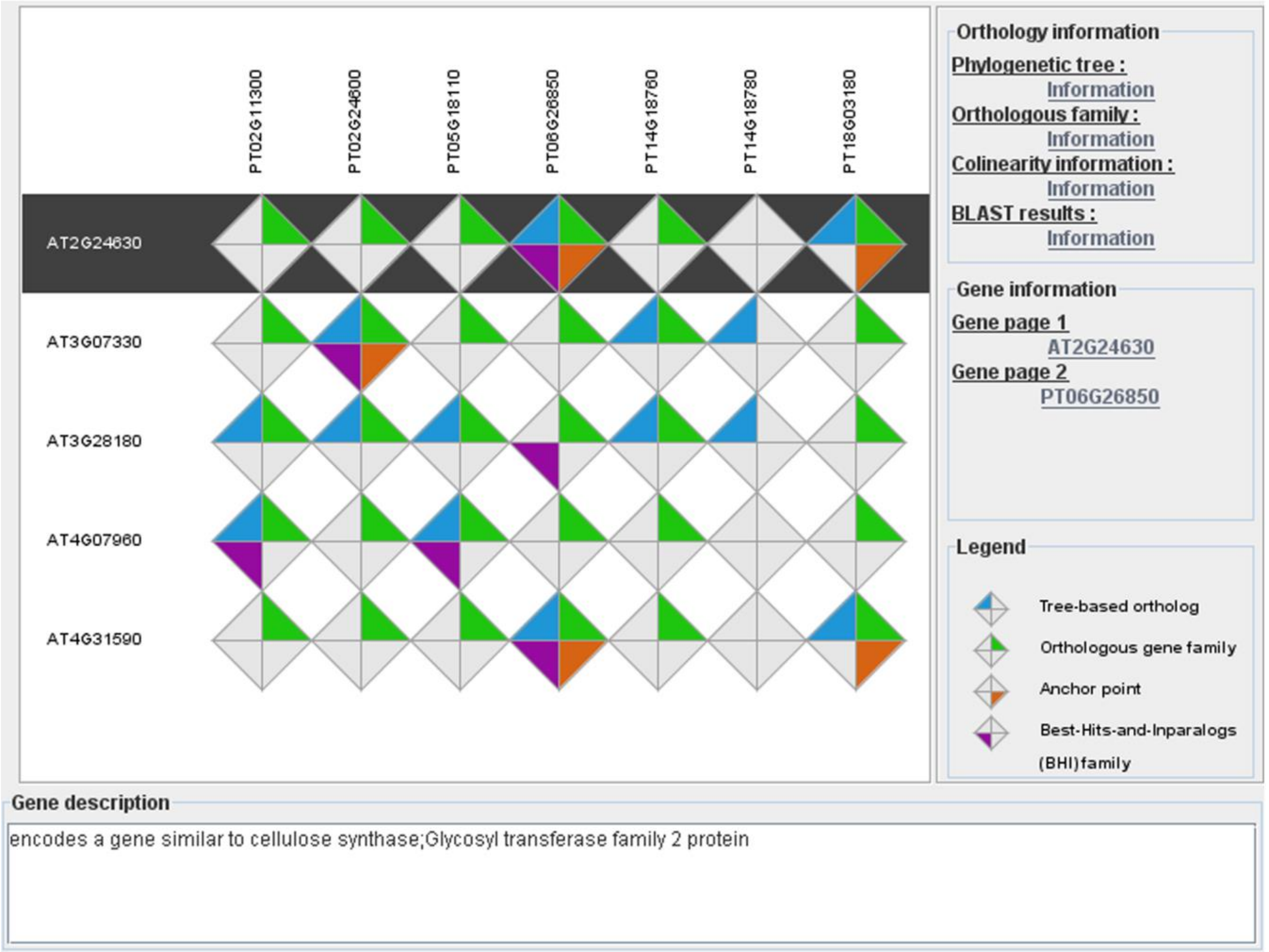
Detection of orthologous plant genes

- Meaning…
  - Orthology = genes derived from a common ancestor in different species
  - Functionally conserved homologs = genes in different species having similar/equal functions

- Due to gene duplication events , complex many-to-many gene orthology is frequently observed

- Functional homologs in different species share …
  - similar expression?
  - regulation?
  - protein-protein interactions?

Integrative Orthology Viewer - an ensemble of different gene orthology prediction approaches



encodes a gene similar to cellulose synthase;Glycosyl transferase family 2 protein

- *Tree-based orthologs (TROG)* inferred using tree reconciliation
- *Orthologous gene families (ORTHO)* inferred using OrthoMCL
- *Anchor points* refer to gene-based colinearity between species
- *Best hit families (BHIF)* inferred from Blast hits including inparalogs

Thanks to:

- ILVO (www.ilvo.vlaanderen.be)
    - Tom Ruttink
    - Isabel Roldan-Ruiz

- Ugent-VIB bioinformatics (bioinformatics.psb.ugent.be)
    - Klaas Vandepoele
    - Michiel Van Bel

    - Pierre Rouze
    - Yves Van de Peer