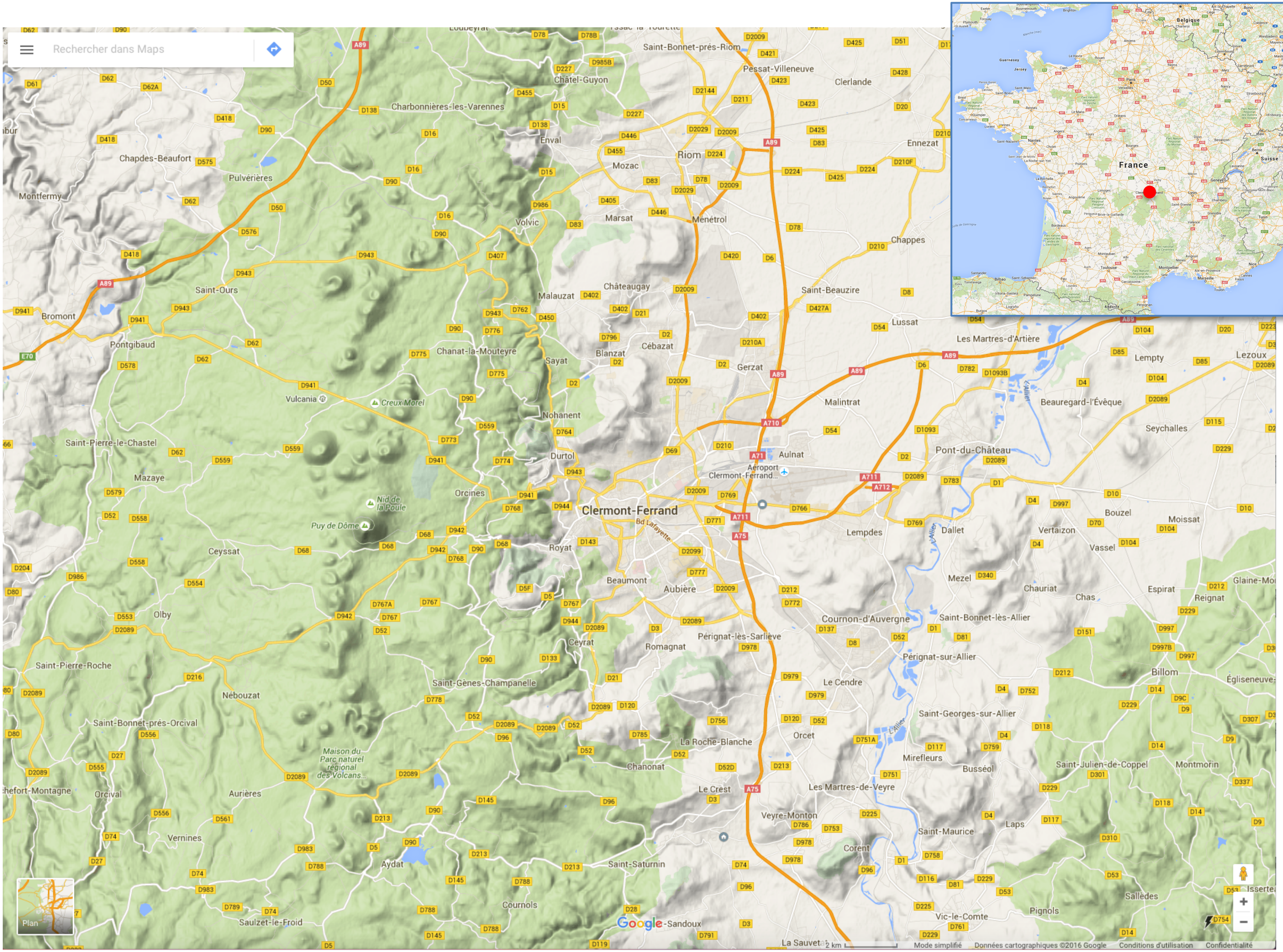
Structure and Dynamics of the Hexaploid Wheat Genome

Frédéric CHOULET

Genetics Diversity Ecophysiology of Cereals
INRA – U. Clermont-Ferrand, France



Rechercher dans Maps
Saint-Bonnet-près-Riom
Châtel-Guyon
Pessat-Villeneuve
Clerlande
Charbonnières-les-Varennes
Enval
Ennezat
Chapdes-Beaufort
Pulvérières
Riom
Mozac
Montfermy
Volvic
Marsat
Ménétrol
Chappes
Saint-Ours
Châteaugay
Malauzat
Saint-Beauzire
Bromont
Lussat
Pontgibaud
Cébazat
Blanzat
Les Martres-d'Artière
Chanat-la-Mouteyre
Sayat
Gerzat
Lempty
Lezoux
Vulcania
Creux Morel
Malintrat
Beauregard-l'Évêque
Nohanent
Seychalles
Saint-Pierre-le-Chastel
Durtol
Aulnat
Pont-du-Château
Mazaye
Aéroport Clermont-Ferrand...
Orcines
Nid de la Poule
Clermont-Ferrand
Bd Lafayette
Moissat
Bouzel
Puy de Dôme
Lempdes
Dallet
Vertaizon
Ceyssat
Royat
Vassel
Mezel
Beaumont
Aubière
Chauriat
Espirat
Glaine-Mo
Reignat
Chas
Olby
Cournon-d'Auvergne
Saint-Bonnet-lès-Allier
Ceyrat
Pérignat-lès-Sarliève
Romagnat
Pérignat-sur-Allier
Billom
Saint-Pierre-Roche
Églisneuve
Le Cendre
Saint-Genès-Champanelle
Nébouzat
Saint-Georges-sur-Allier
Saint-Bonnet-près-Orcival
Orcet
La Roche-Blanche
Maison du Parc naturel régional des Volcans...
Mirefleurs
Saint-Julien-de-Coppel
Chanonat
Busséol
Montmorin
Rochefort-Montagne
Orcival
Aurières
Le Crest
Les Martres-de-Veyre
Veyre-Monton
Laps
Vernines
Saint-Maurice
Corent
Aydat
Saint-Saturnin
Isserte
Salledes
Cournols
Pignols
Vic-le-Comte
Saulzet-le-Froid
Sandoux
La Sauvetat
Plan
Google
2 km
Mode simplifié
Données cartographiques ©2016 Google
Conditions d'utilisation
Confidentialité
France
Belgique
Suisse

GDEC
Genetics Diversity Ecophysiology of Cereals

- Structure, evolution of wheat genome
- Recombination
- Paleogenomics
- Grain composition
- Response to abiotic stress
- Resistance to pathogens
- Diversity, selection

Objectives

❑ Resources

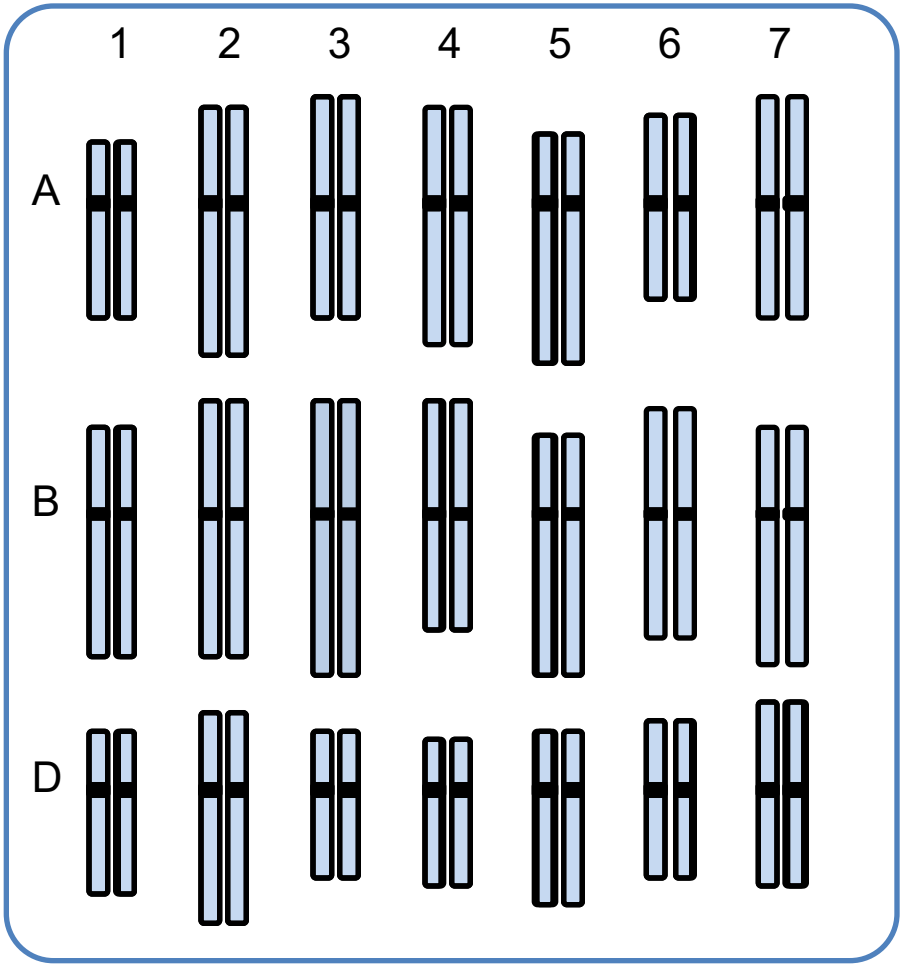
- Sequencing the wheat genome
- Markers
- Bioinformatics

❑ Research

- Structure / Expression / Evolution
  - TEs
  - Gene space
  - Duplications
  - Structural variations
  - Epigenome

❑ Why wheat?

- Important crop
- Complex genome



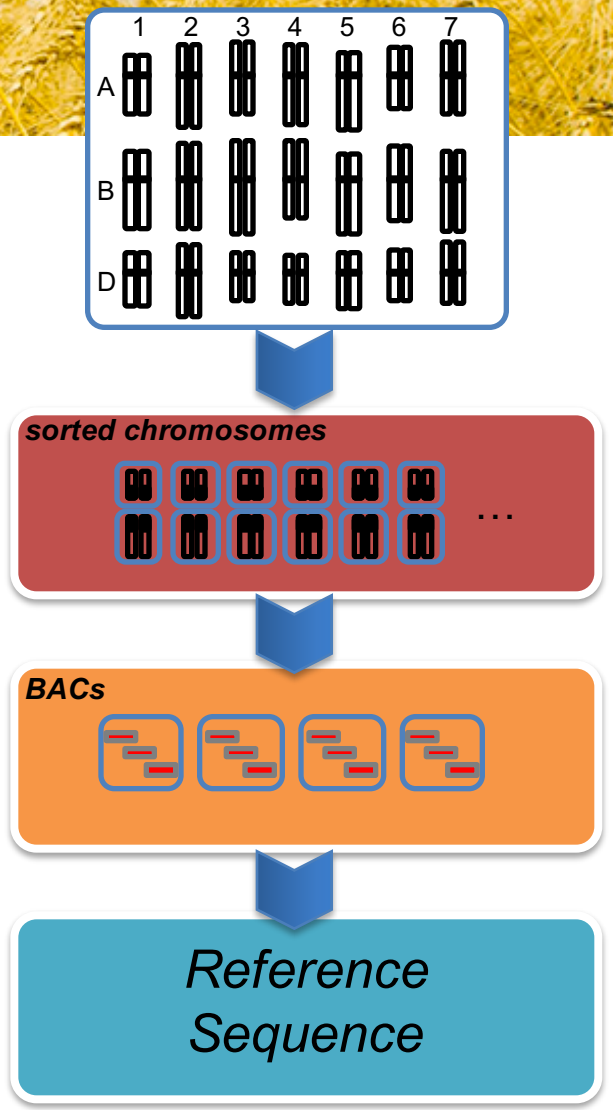
17 Gb

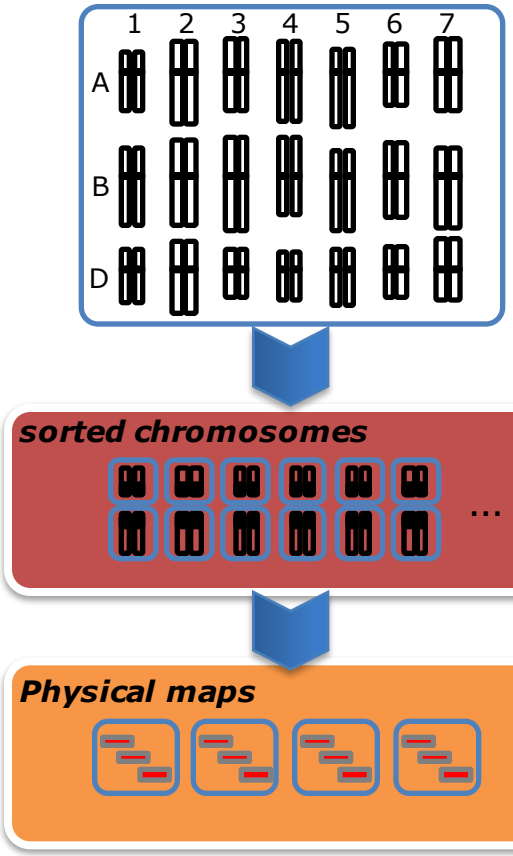
AABBDD

85% TEs

Generating a high quality genome sequence of bread wheat

- Launched in 2005
- Goal
  - ➢ *Produce a high quality ref seq of the bread wheat genome*
- Strategy
  - ➢ *Reduce the complexity*





WGS

- *Brenchley et al. Nature 2012*
- *Ling et al. Nature 2013*
- *Jia et al. Nature 2013*
- *Chapman et al. Genom Biol 2015*

Chr. Survey Seq (=CSS)

- *IWGSC Science 2014*

MTPseq

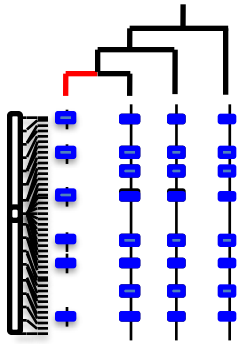
3B

- *Choulet et al. Science 2014*
- *Daron et al. Genom Biol 2015*
- *Pingault et al. Genom Biol 2015*
- *Glover et al. Genome Biol 2015*

❑ Chr Survey Seq (2014)

❑ Resources

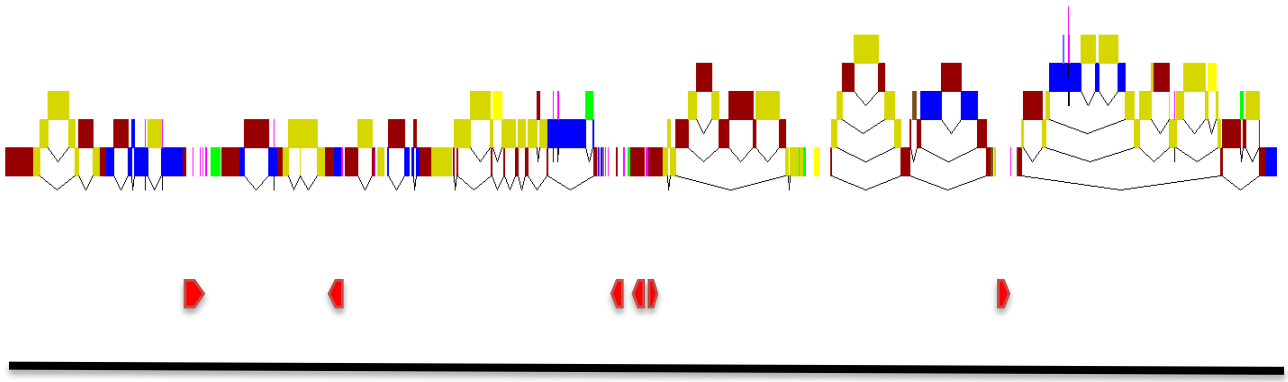
- 1 draft seq / chromosome arm
- **10** Gb - **10 M** contigs (N50: 2.4 kb)
- **99,000** genes
- ~60% genes "zipped"



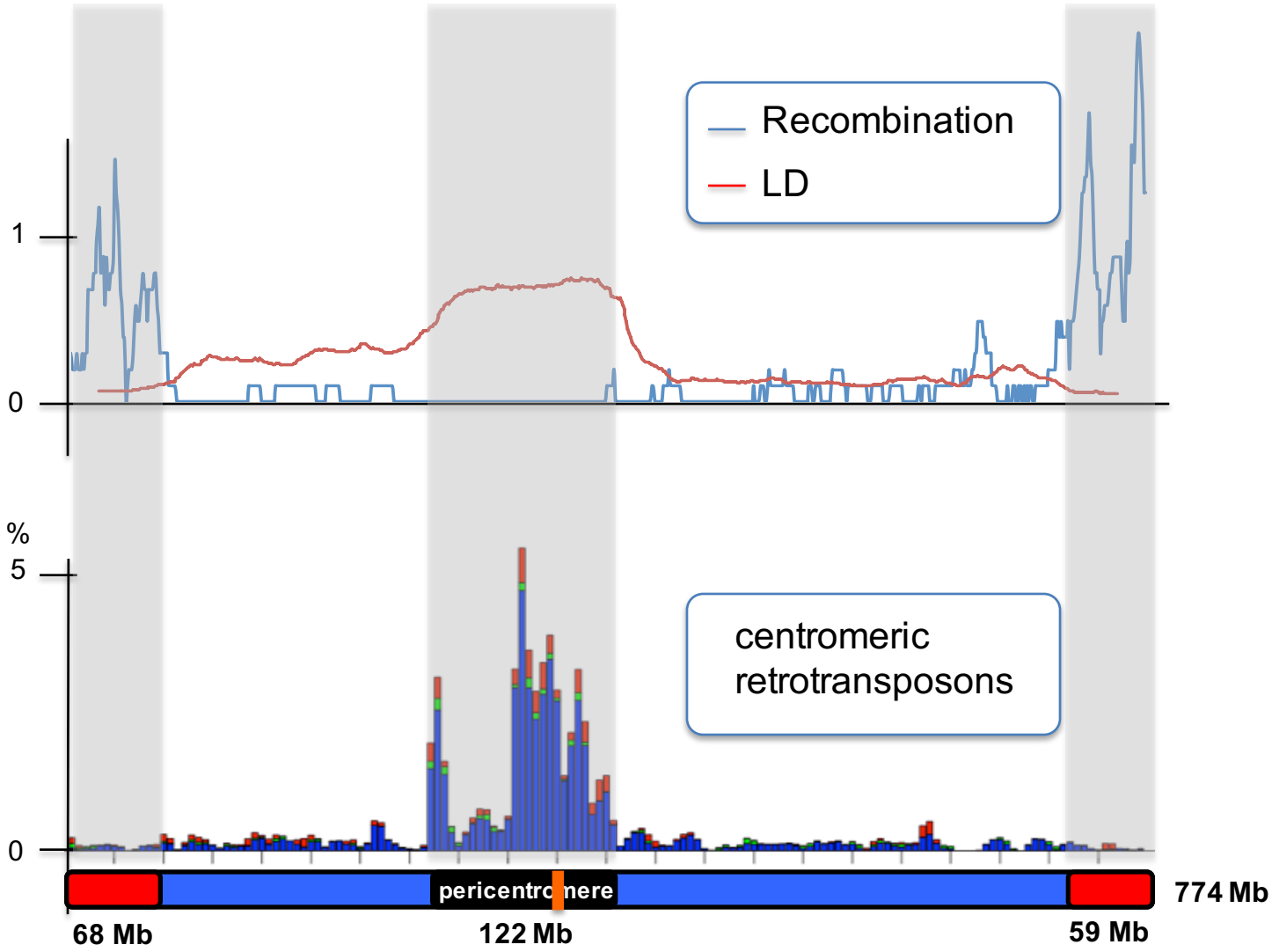
❑ Main Results

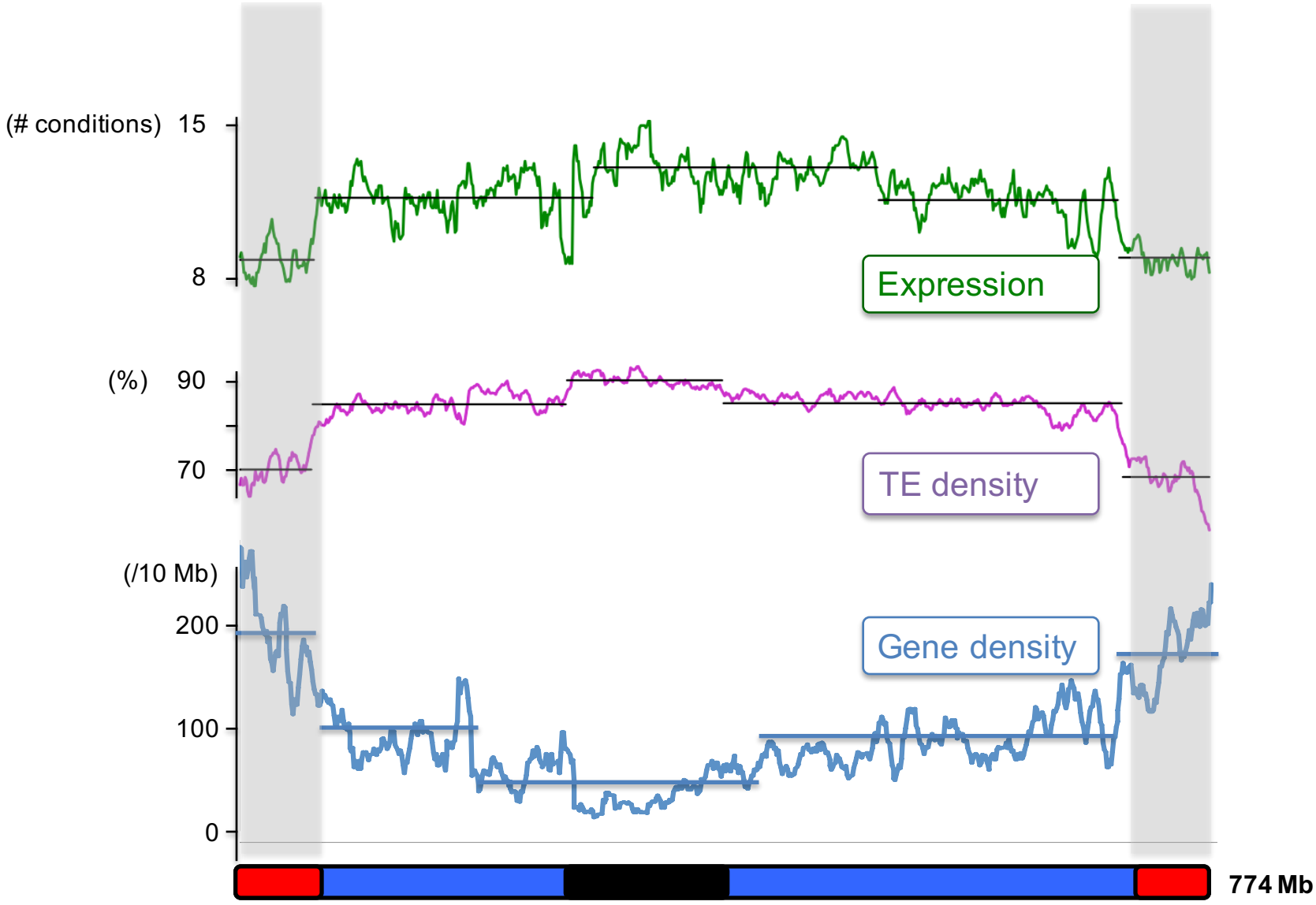
- Gene loss--
- SSD++
- Dominance--

❑ 3B pseudomolecule (3BSEQ project)

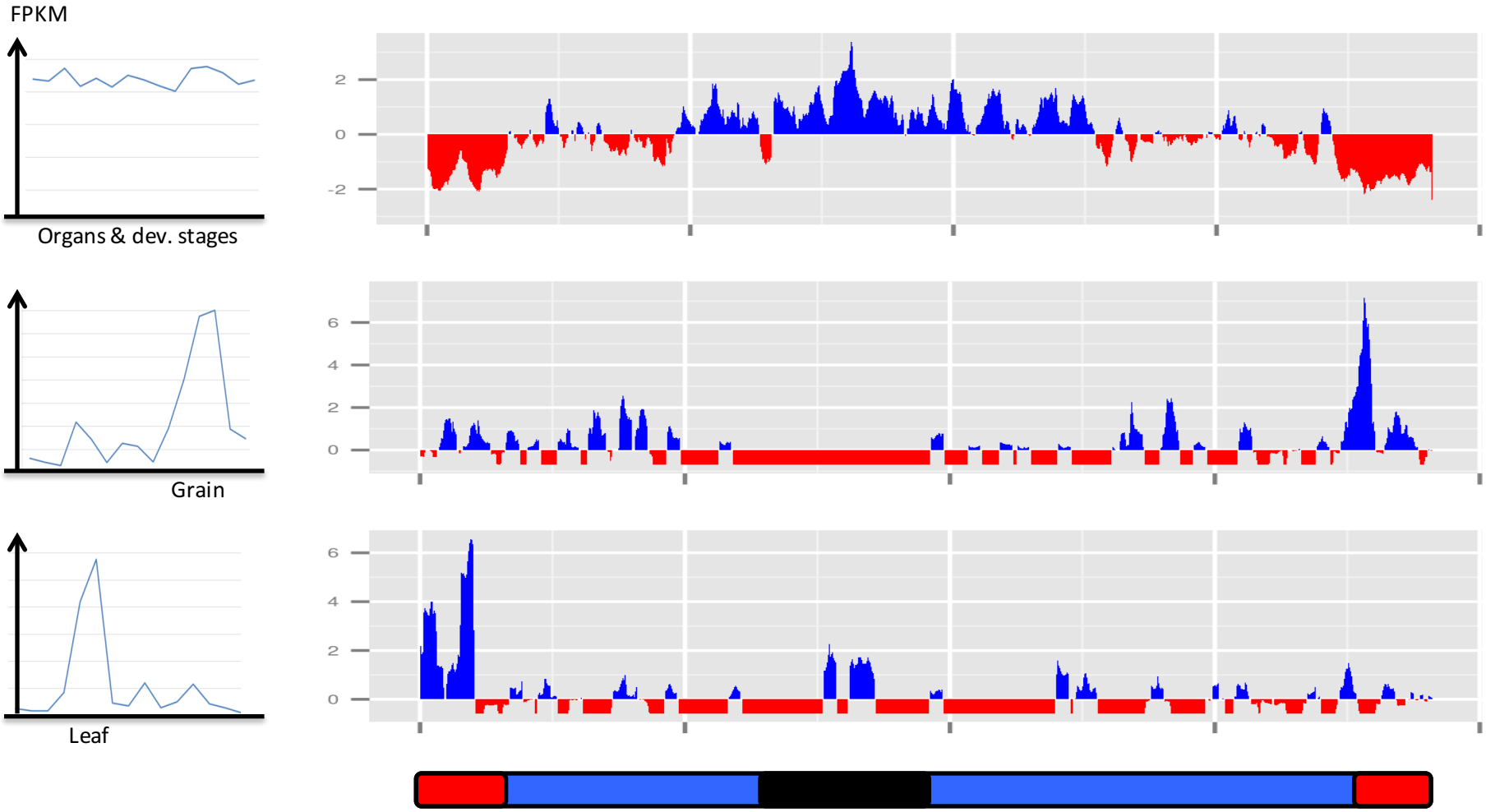


- protein coding genes 7,264
  (pseudogenes: 27%)
- transposable elements 252,879 (86%)


Recombination
LD
1
0
%
5
0
centromeric
retrotransposons
pericentromere
774 Mb
68 Mb
122 Mb
59 Mb

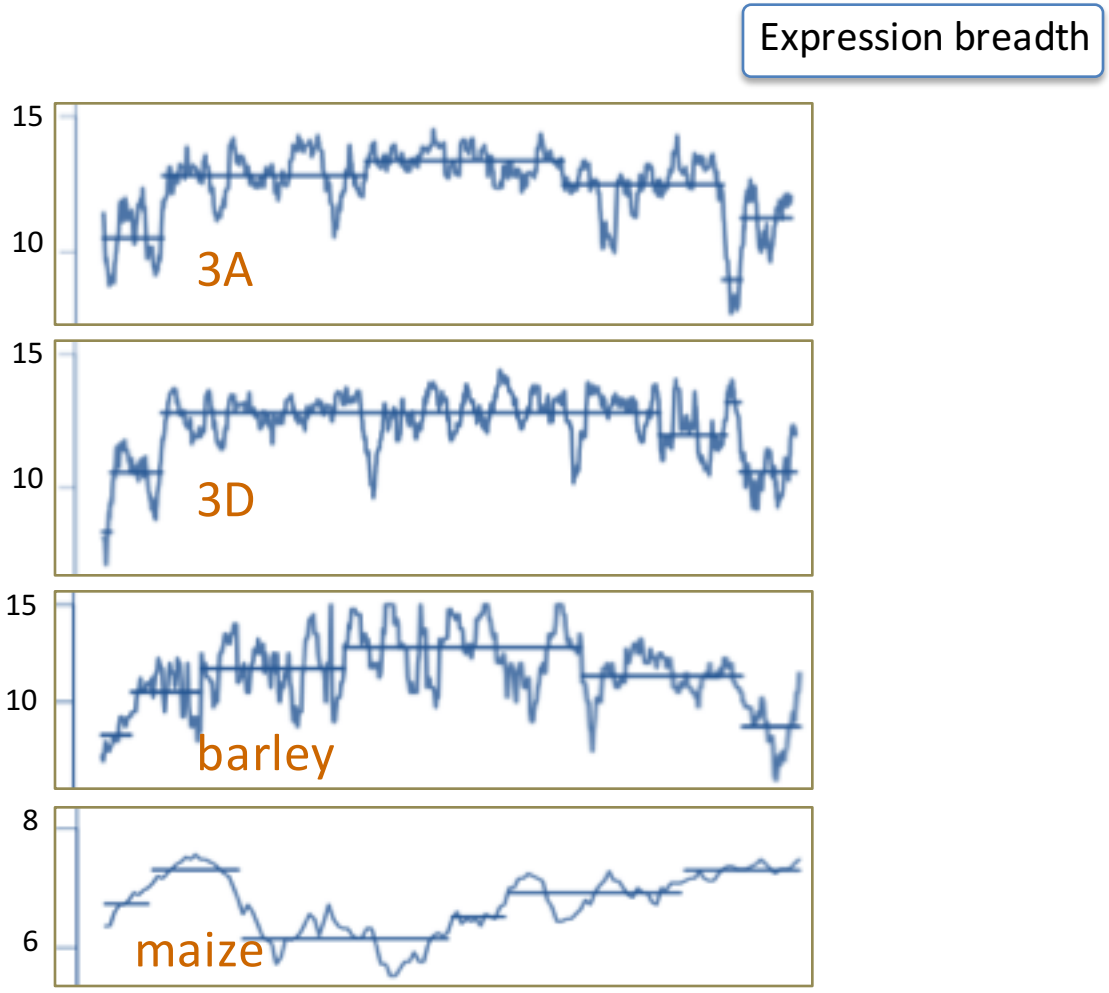


Choulet et al. Science 2014

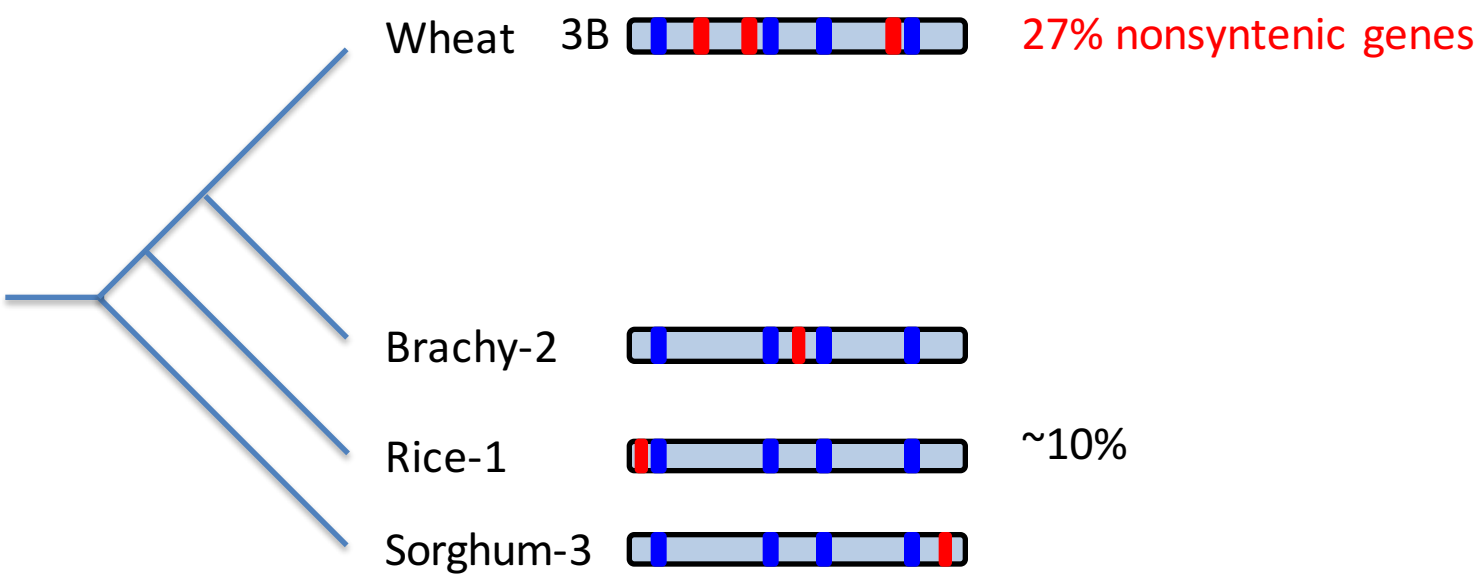


Pingault et al. Genom Biol 2015

❑ Chromosome partitioning: 3B-specific??



❑ Accelerated evolution in the *Triticeae*

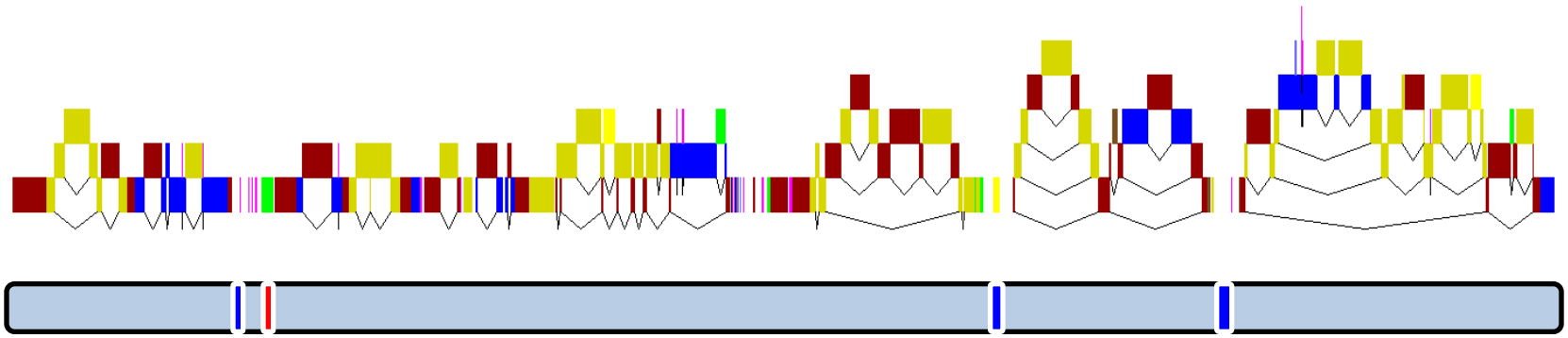


- inter-chromosomal gene duplications
- intra-chromosomal -

- More duplicated genes in the chr. extremties
- Enriched in adaptation functions

Glover et al. Genom Biol 2015

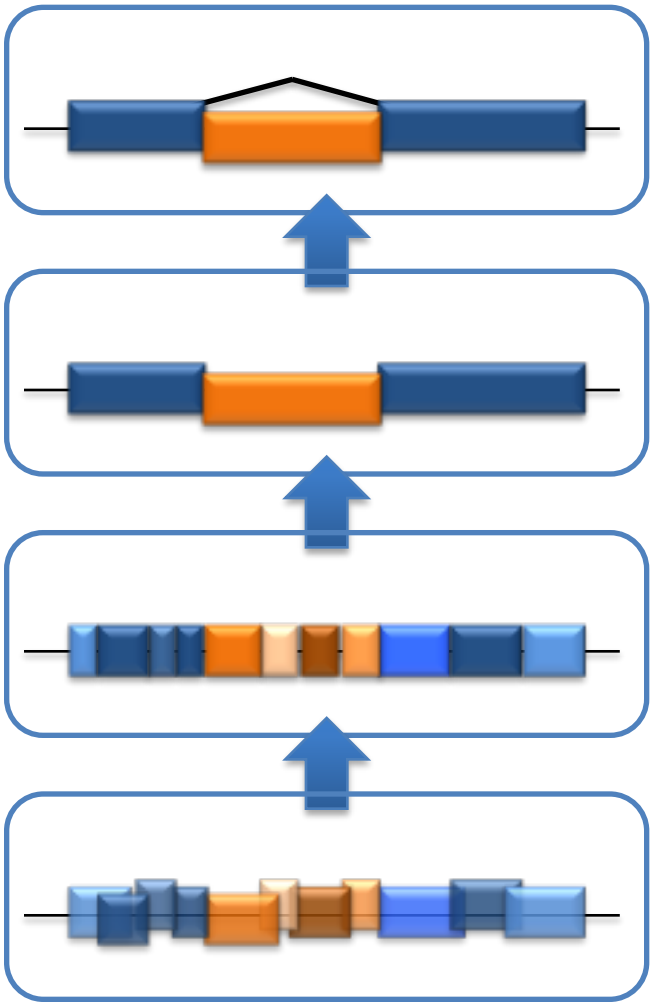
❑ TEs



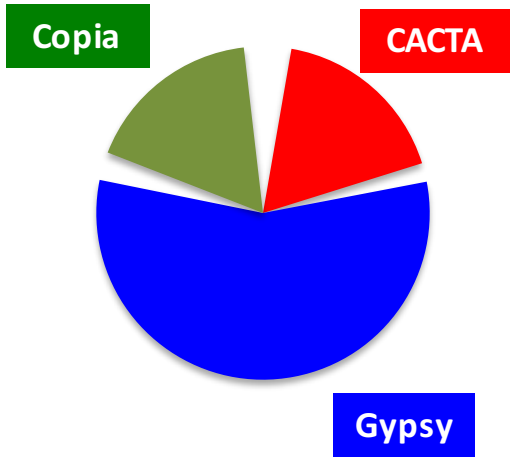
- Annotation challenge
- Impact on genome biology

Daron et al. Genom Biol 2015

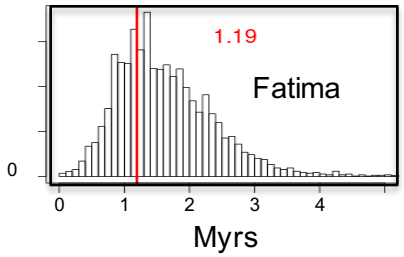
❑ New tools: **CLARITE** and **ClariTeRep**

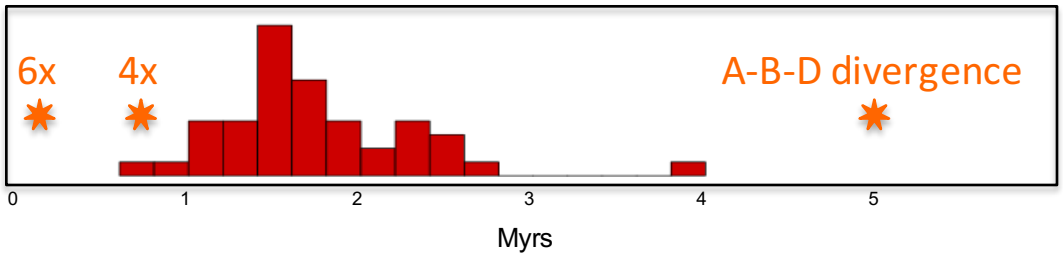


3B → ***252 000*** TEs

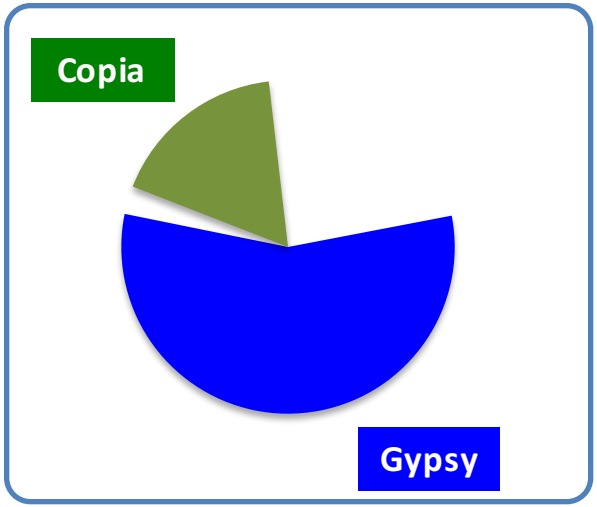


❑ Insertion dynamics

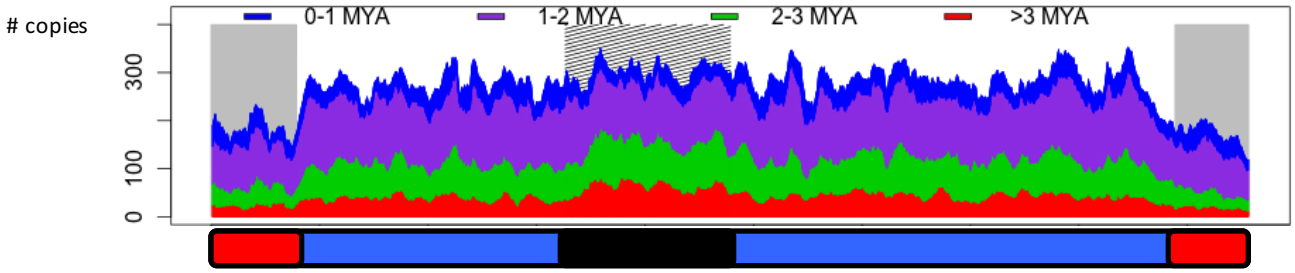


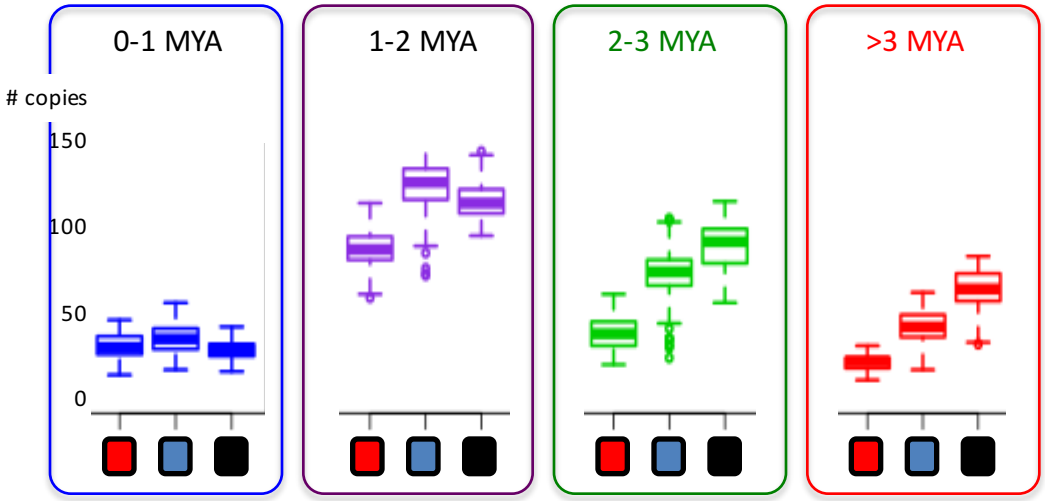


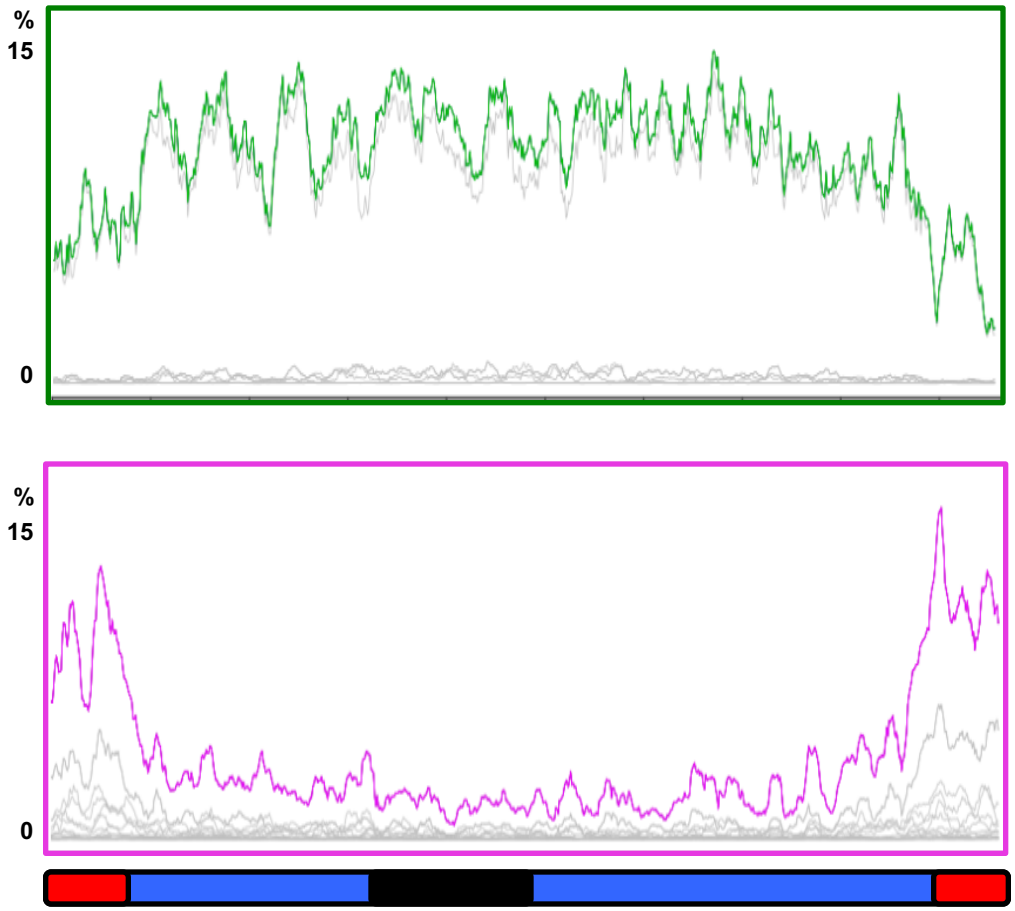
- silenced until polyploidization…
- … but shared betw A-B-D

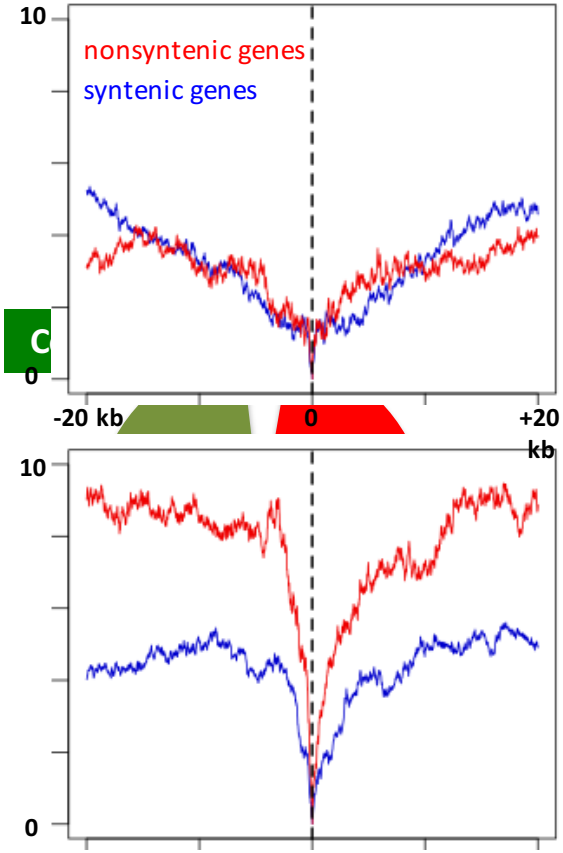


❑ Evolutionary forces driving TE distribution






%
15
0
%
15
0

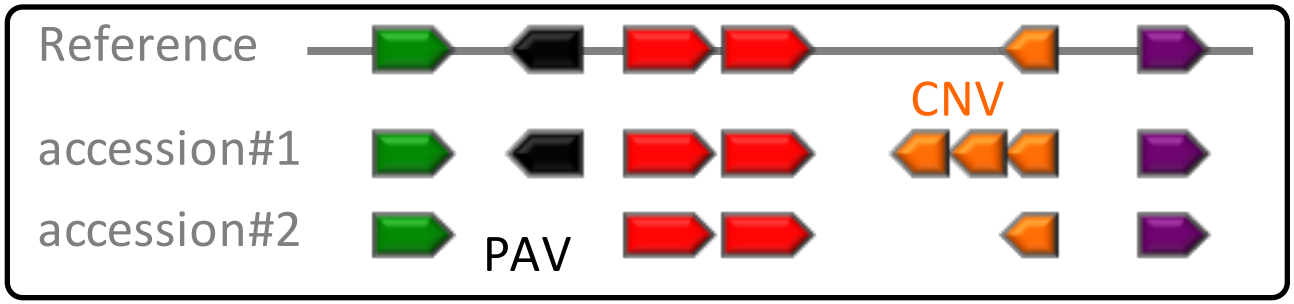

10
nonsyntenic genes
syntenic genes
C
0
-20 kb
0
+20 kb
10
0

Structural Variations

- CNVs
- PAVs

❑ Structural Variations (SVs)

- Small indels
- Inversions, translocations
- **Duplications & Deletions --> CNVs & PAVs**

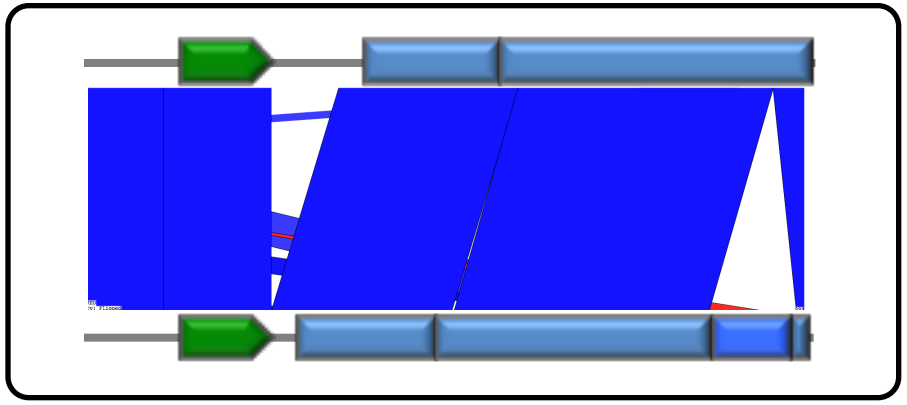


Main questions:

- Extent of SVs among the *Triticeae,* hexaploid wheat acc.?
- Impact of polyploidization?
- Relationships betw SVs and chr. organization?
- Impact on phenotypes?

❑ Structural Variations (SVs) – **Detection**

- Aligning orthologous sequenced loci

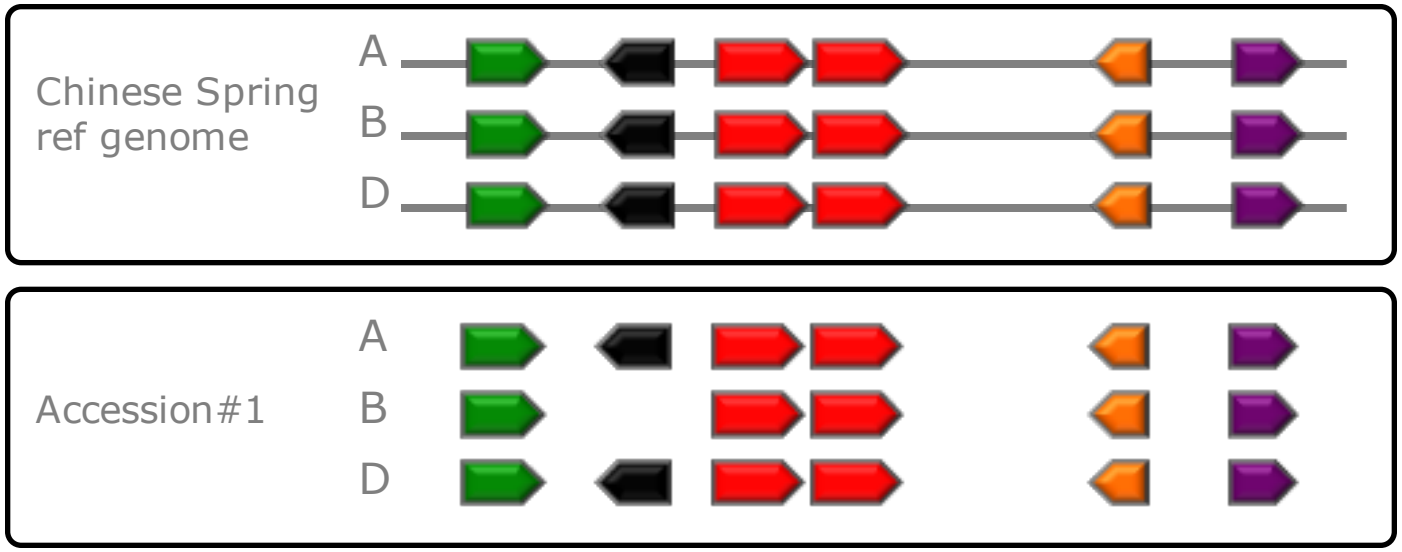


ACT view

- Using resequencing data (short read-based)
  - not properly mapped paired-reads
  - split reads
  - depth of coverage

→ ***Limitations*** *in polyploid TE-rich genomes*

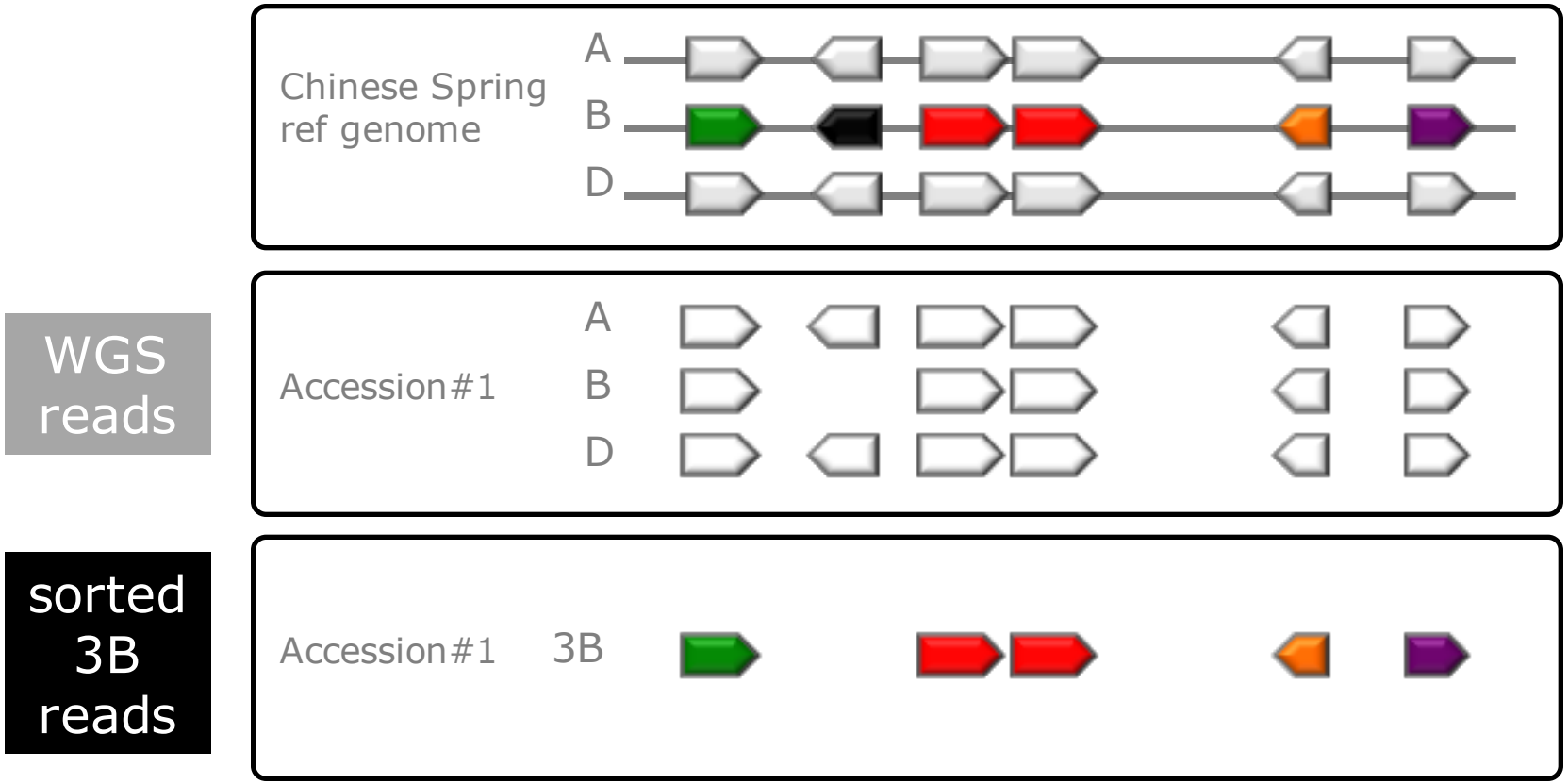
❑ Resequencing data



WGS reads

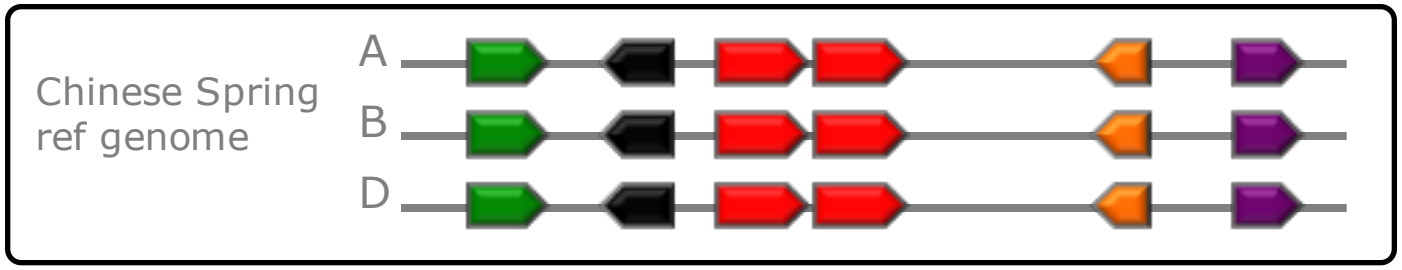
- **limitations:** homeologs and paralogs (repeated genes)

❑ Resequencing data



- **advantage:** diploid context
- **limitations:** 3B DNA amplified before seq.

❑ Resequencing data



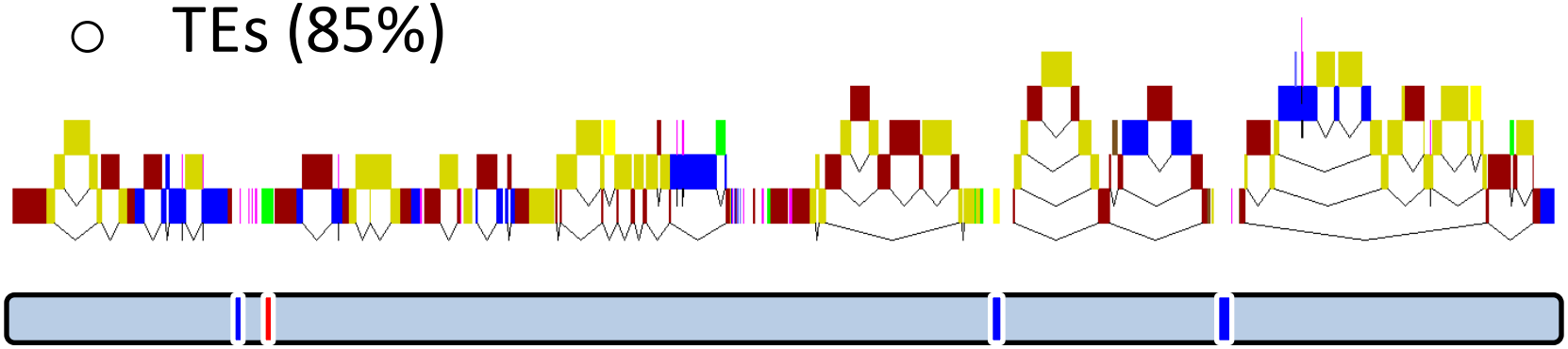
sorted 3B reads

45 accessions

22 hexaploids: *T. aestivum*
6 hexaploids: *T. macha, spelta*
17 tetraploids: *T. durum, dicoccoides, dicoccum, carthlicum*

- Illumina 2x100bp
- Depth: **~40x**

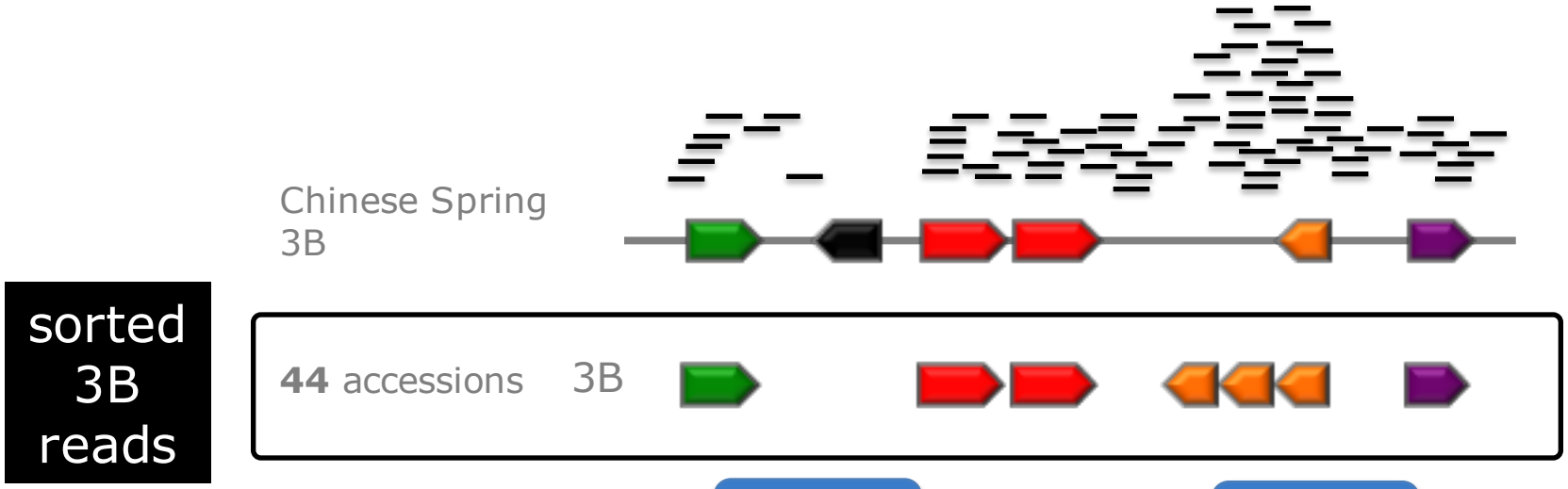
- TEs (85%)



- Genes (2%)

❑ SVs in genes – Methodology

- BWA & samtools



PAVs

<**10%** of the gene length covered by reads

CNVs

depth of coverage-based approach

Normalization:

- GC%
- sequencing depth
- gene length
- $\log_2(cov^{acc}/cov^{Chinese\ Sp})$

❑ SVs in genes – Results

PAVs

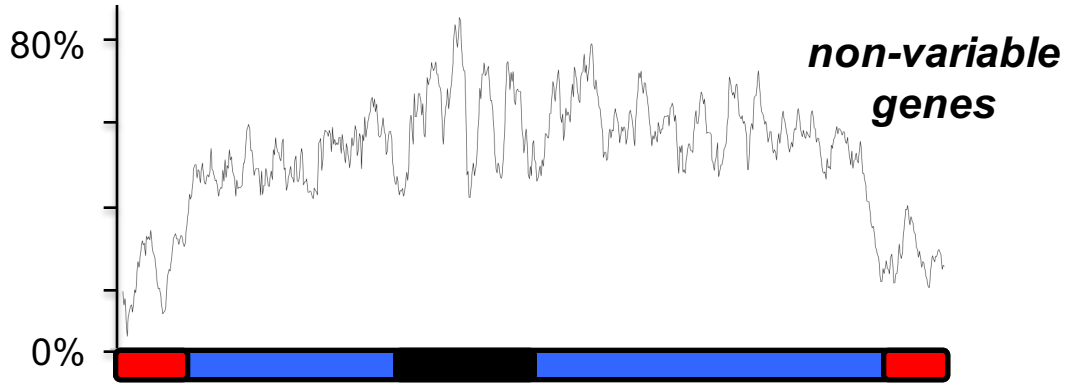
XX% genes deleted in 1+ accessions [XX..XX]

CNVs

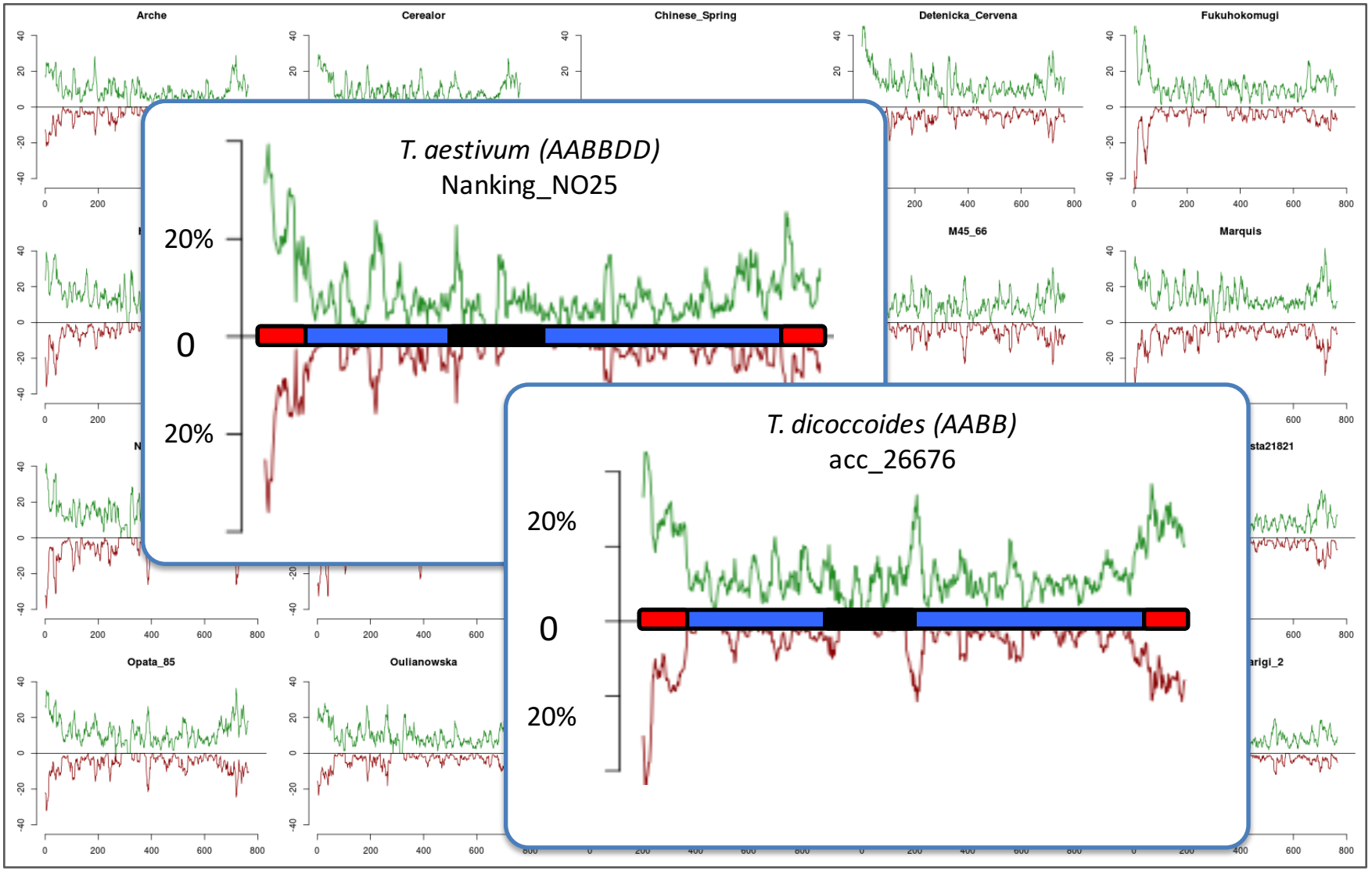
down^{CNVs}: X% of the 3B genes (on average per acc.)

up^{CNVs}: X% "

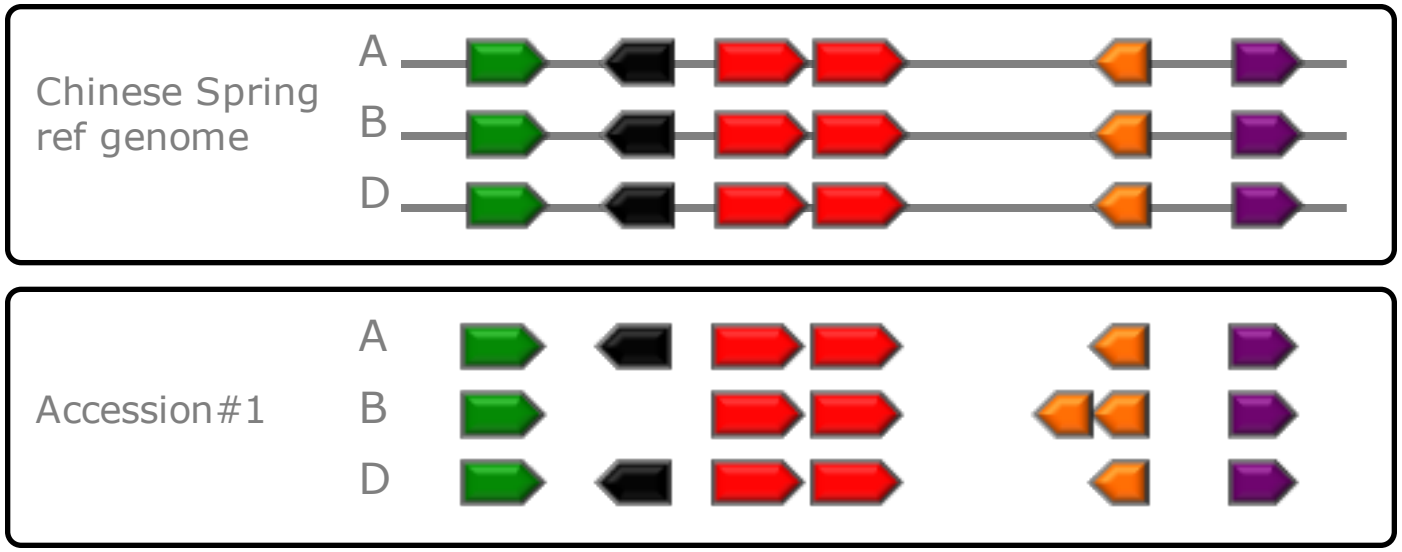
XXXX (**XX%**) genes with no variation among 45 accessions



❑ SVs in genes – Results



❑ SVs in genes – Methodology

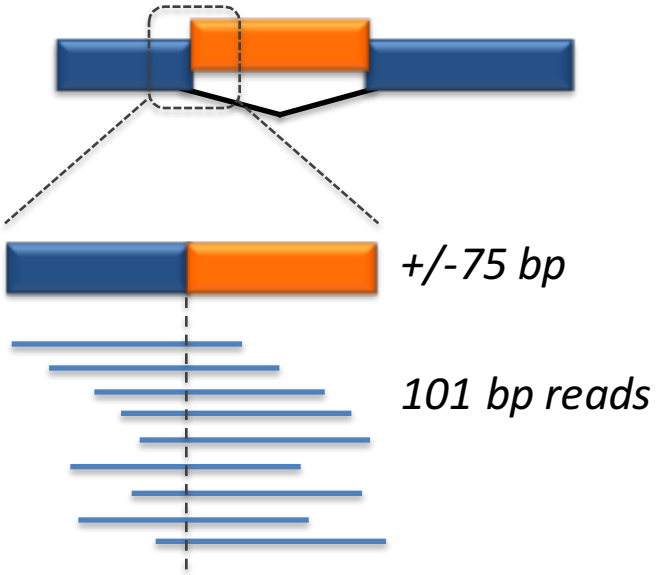


WGS reads

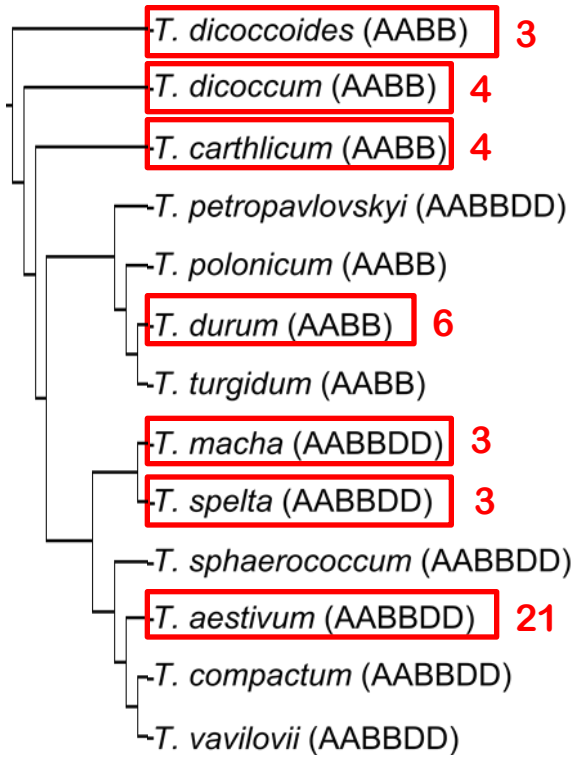
=> develop a fine-tuned strategy?... in progress

❑ SVs in TEs – Methodology

TE junction-based approach (**PAVs** only)



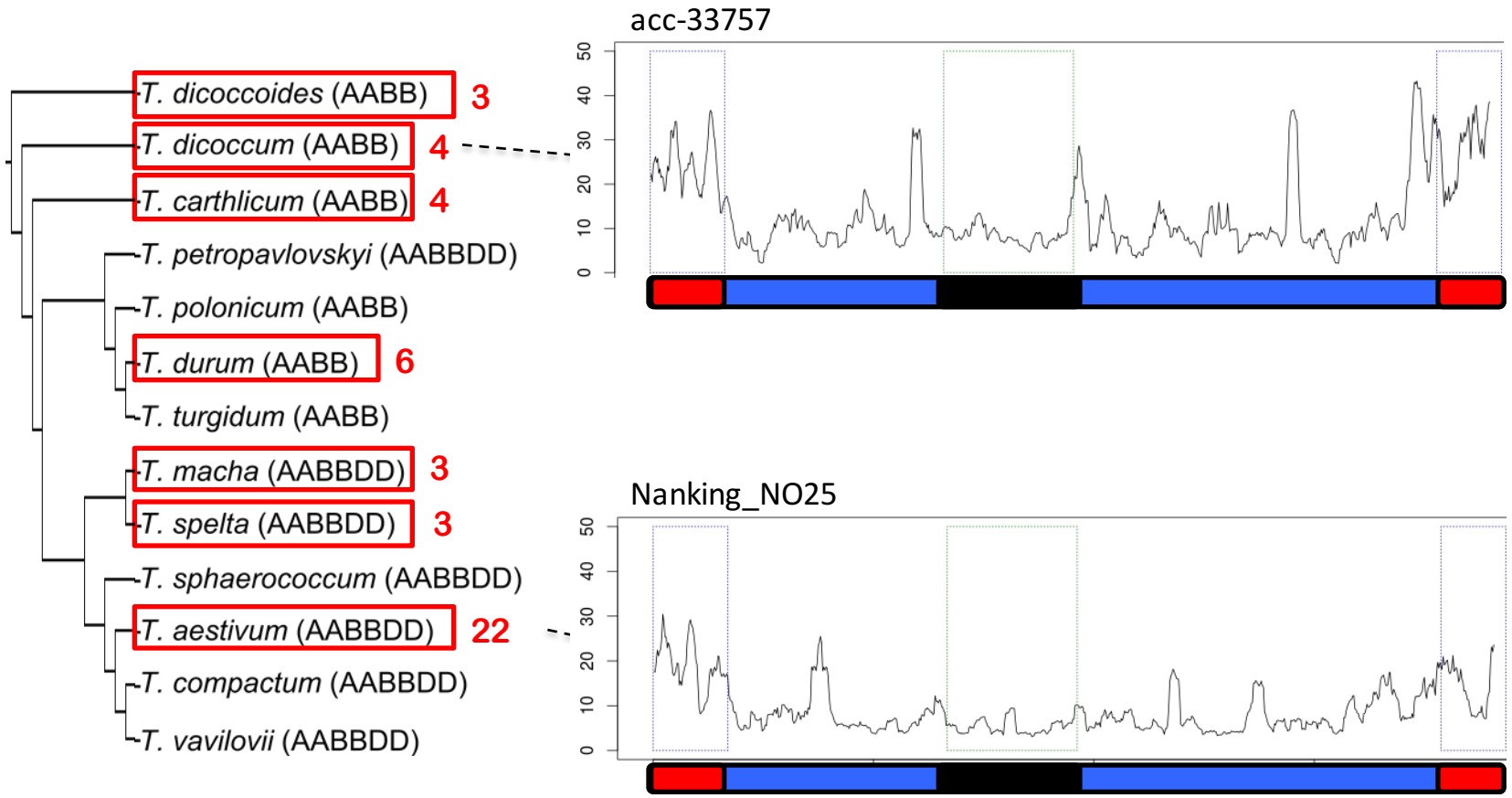
❑ SVs in TEs – Results



3B ➔ ~**500k** loci to study TE-PAVs along chr3B

- Polymorphic loci among 45 accessions: XX%
- Average per accession: XX% [XX% .. XX%]

❑ SVs in TEs – Results



❑ Conclusions

- Gene PAVs limited (on 3B)
- CNVs: **XX%** of the 3B genes (versus ~10% in barley *[Munoz-Amatriain et al. 2013]* 14 wild+cultiv. genotypes)
- High level of **TE**-related SVs
- Importance of **chr. extremities** in the diversity of *Triticeae* (also observed in barley)

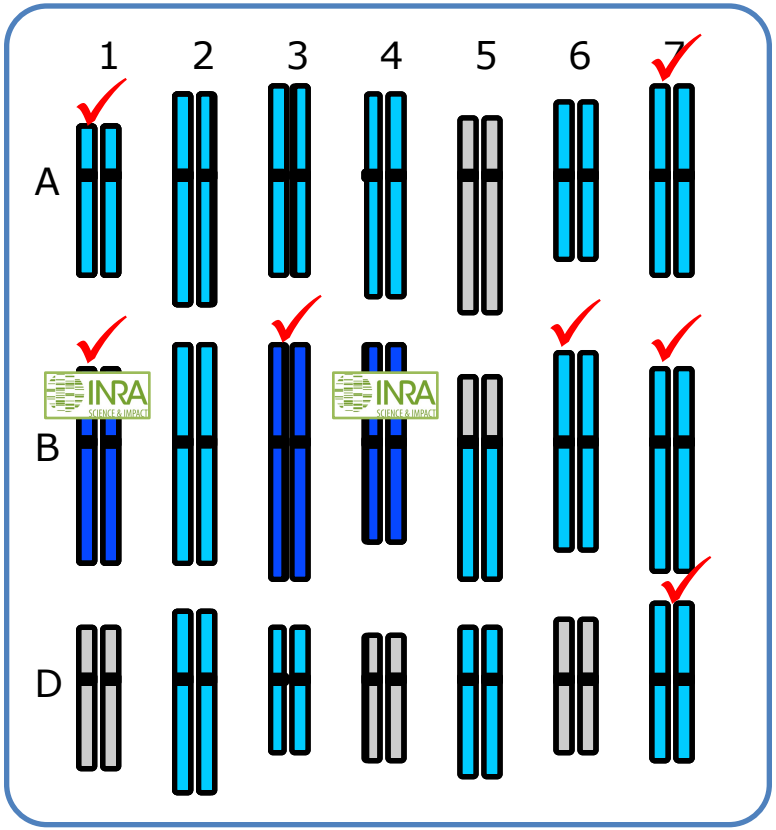
❑ Next

- Validate the approach/thresholds used for CNV calling
- SVs at the whole genome level
  - ->use 3B-*vs*-3B results as QC
  - ->increase sample size through exon capture
- GO term enrichment
- Relationships betw TE SVs / gene expression
- Unmapped reads (pangenome)

Wheat genome seq initiatives in 2016:

- MTPseq 1A, 1B, 6B, 7A, 7B, 7D
- **IWGSC Whole Genome Assembly (NRGene)**
- TGAC WGS (several varieties)
- U Maryland WGS Pacbio+Ill
- UC Davis *Ae. tauschii* WGS+MTPseq
- BGI *T. urartu* WGS+MTPseq
- Wild Emmer Wheat (NRGene)

chr-by-chr approach, status in 2015…



IWGSC-Whole Genome Assembly

Partners:

N. Stein

C. Pozniak

J. Poland

A. Diestelfeld

A. Scharpe

G. Ronen

M. Thompson

K. Eversole, J. Rogers

F. Choulet

Strategy:

WGS Illumina 180x - 3 MP lib

DeNovoMAGIC™ 2.0

Timeline:

- Aug 2015 -> start
- Sept
- Oct
- Nov -> Sequencing done
- Dec -> Assembly v0.1

  QC *(M. Mascher, F. Choulet)*
- Jan. 2016 -> Assembly v0.2

IWGSC-Whole Genome Assembly

				N50	
2013	T.ur	BGI	T.ura-BGI	64 kb /	19000 scaff
2013	Ae.t	BGI	Ae.ta-BGI	58 kb /	19000 scaff
2014	CS	IWGSC	CSS	2 kb /	>1M scaff
2014	CS(3B)	GDEC/CNS	3B-pseudo	892 kb /	296 scaff
2015	Synth	IPK/JGI	Syn-JGI	21 kb /	120000 scaff
2015-Jul	WEW	NRGene	WEW-NRGene	7000 kb /	414 scaff
2015-Dec	CS	IWGSC	IWGSC-WGA	**7394 kb /**	**547 scaff**

IWGSC-Whole Genome Assembly

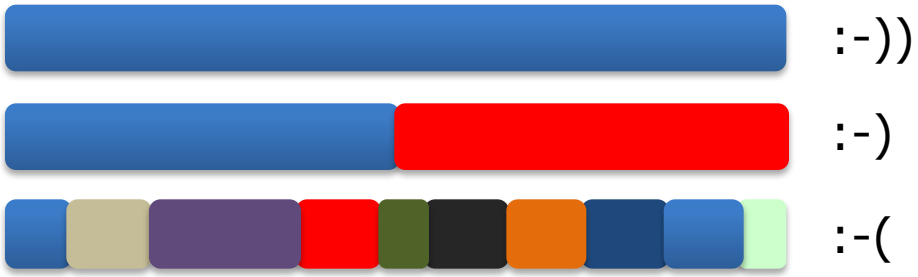
QC :: Completeness?

- exons: **98.7%** match with 100%id-100%ov
- ISBPs (4.2M): **97.7%** match with 100%id-100%ov
- WGPtags-4B (0.9M): **96.2%** match with 100%id-100%ov

➔ Completeness+++

IWGSC-Whole Genome Assembly

QC :: Chimeras?



- alignment to CSS genes and ISBPs
- alignment to MTPseq (3B, 1B)
- alignment to genetic map (POPseq, CsRe)
- alignment to physical map (WGPtags)
- alignment to HiC map

➔ Chimeric scaffolds found (<200) -> corrected in v0.2/v0.3

IWGSC-Whole Genome Assembly

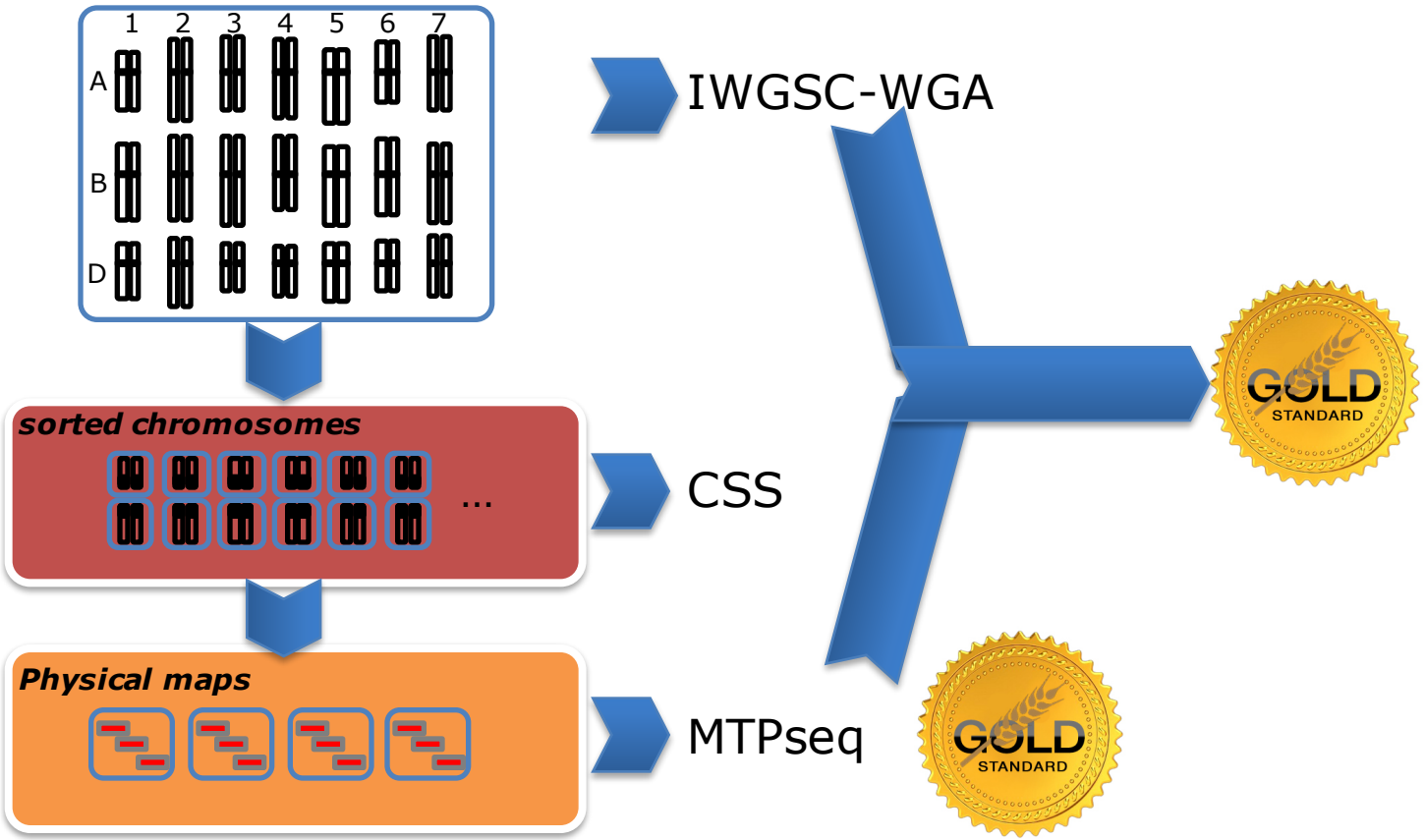
IWGSC-WGA v0.2 metrics:

#Scaff≥2kb:	37,872
Size:	14.532 Gb
Gaps:	1.8%
L50:	7.058 Mb / 566 scaff
L90:	1.261 Mb / 2,363 scaff
max:	45.794 Mb

21 pseudomolecules constructed with HiC data (IPK)

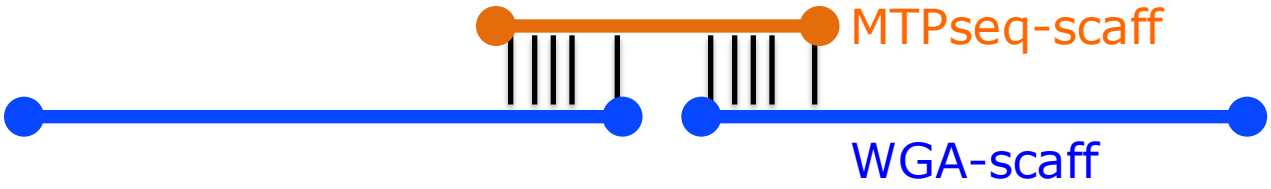
→ 14.0 Gb (**96%**) IWGSC-WGA ordered

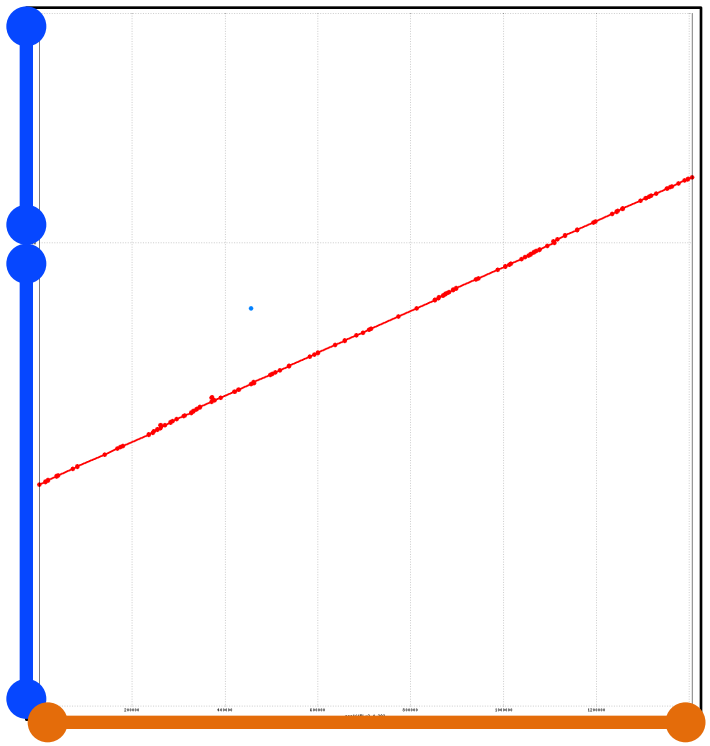
IWGSC roadmap update



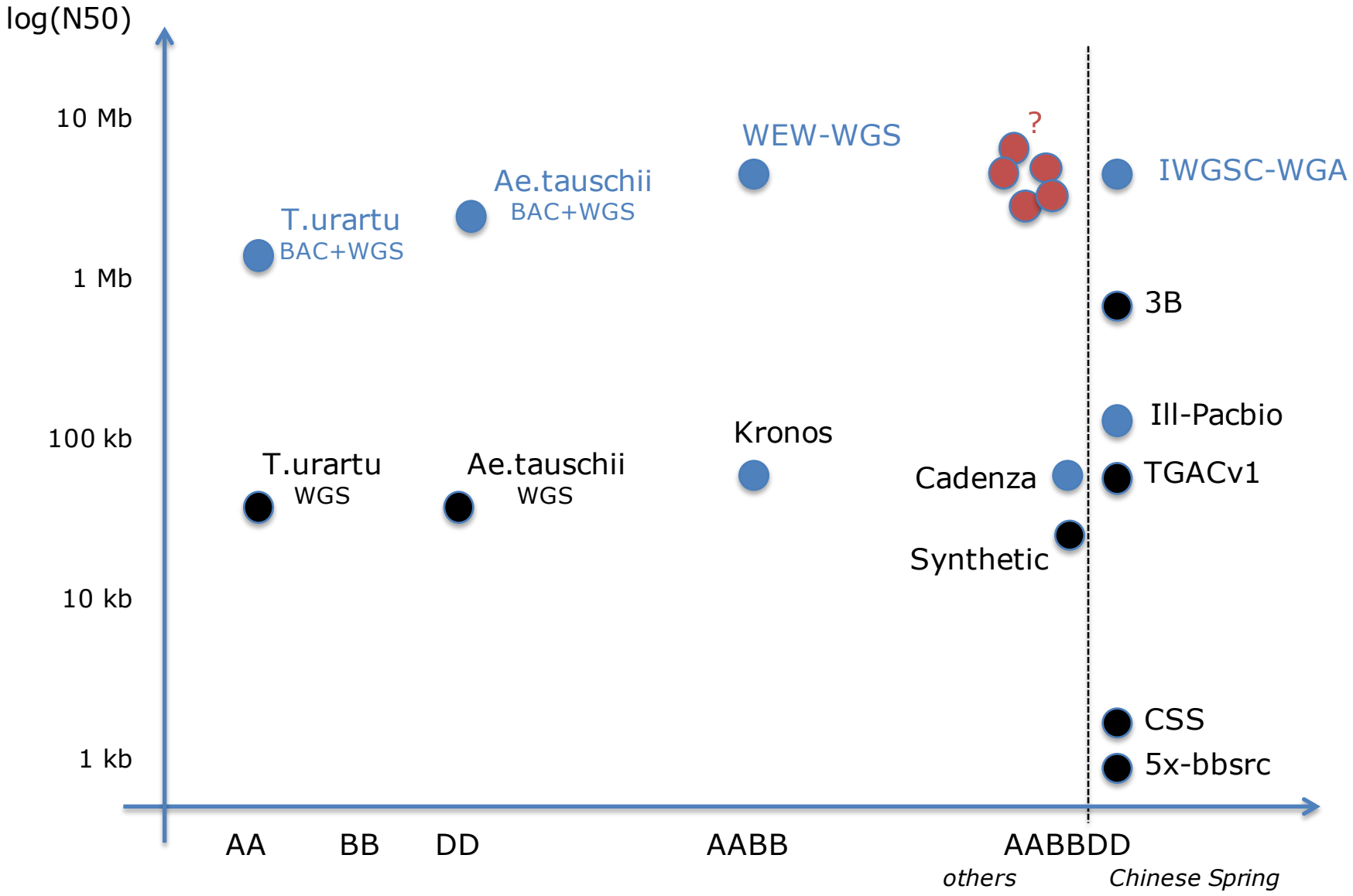
- Integration WGA ⇔ MTPseq in progress…

Strategy based on comparing ISBPs (speed++ specificity++)





207 WGA-scaff joined by 1B-MTP-scaff
representing 593 Mb


log(N50)
10 Mb
1 Mb
100 kb
10 kb
1 kb
T.urartu
BAC+WGS
Ae.tauschii
BAC+WGS
WEW-WGS
?
IWGSC-WGA
3B
Ill-Pacbio
TGACv1
CSS
5x-bbsrc
T.urartu
WGS
Ae.tauschii
WGS
Kronos
Cadenza
Synthetic
AA
BB
DD
AABB
AABBDD
others
Chinese Spring

Thanks

INRA GDEC

Etienne Paux
Hélène Rimbert
Ambre-A. Josselin
Romain De Oliveira
Jonathan Kitt
Benoit Darrier
Nicolas Guilhot
Philippe Leroy
Pierre Sourdille
François Balfourier
Charles Poncet
Josquin Daron
Lise Pingault
Natasha Glover
Sébastien Theil
Aurélie Evrard
Emeric Dynomant
Aurélien Bernard

CEA-Génoscope

P. Wincker, V. Barbe et al.

INRA CNRGV

H. Bergès et al.

INRA URGI

H. Quesneville, M. Alaux et al.

IEB

J. Dolezel et al.

IWGSC

K. Eversole, J. Rogers

IWGSC-WGA working team

IPK, U Sask, KSU, U TelAviv,
GIFS, Illumina, NRGene, …

Frédéric CHOULET