A strategy to assemble a reference sequence of the 1 Gb wheat chromosome 3B

Frédéric Choulet, Sébastien Theil, Josquin Daron, Natasha Marie Glover, Nicolas N. Guilhot, Philippe Leroy, Lise Pingault, Etienne Paux, Pierre Sourdille, Arnaud Couloux, et al.

TGAC/ESF-LESC workshop

Strategies for de novo assemblies of complex crop genomes

Norwich, Oct-9-2012

TGAC
The Genome Analysis Centre™

EUROPEAN SCIENCE FOUNDATION

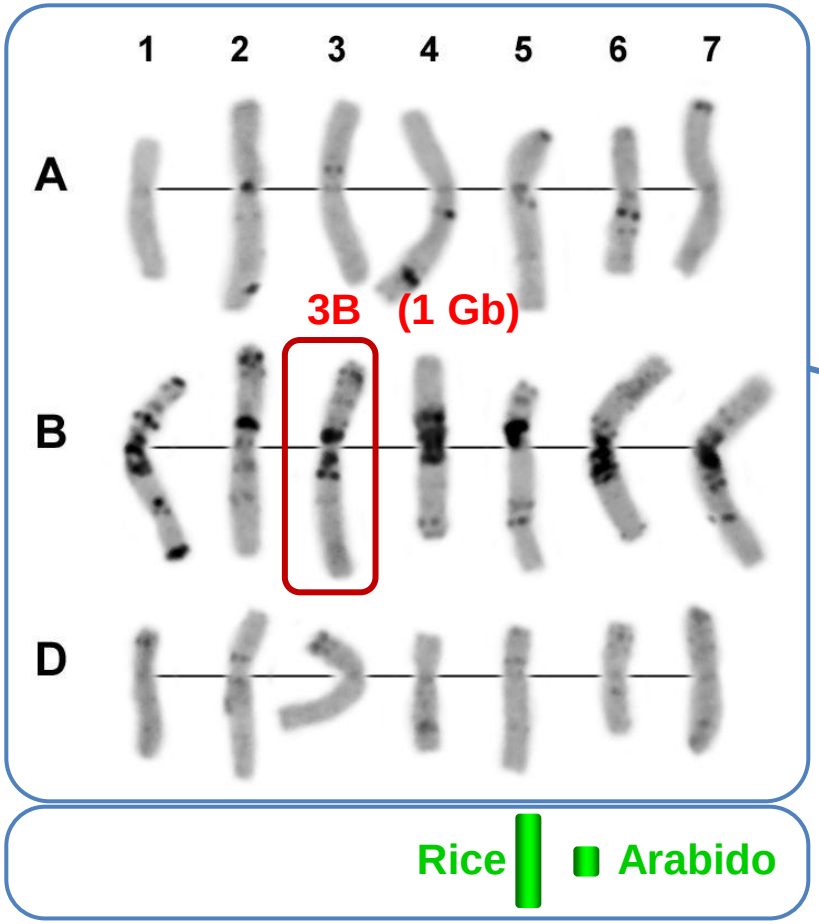
Sequencing and assembling wheat chromosome 3B

Frédéric CHOULET

UMR 1095
GDEC
INRA
Université Blaise Pascal

INRA

The bread wheat genome



- ✓ Allohexaploid (A, B & D)
- ✓ 17 Gb
- ✓ 85% of sequences derived from TEs

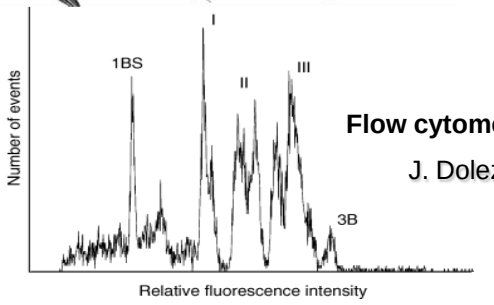
WGS

5x @ *www.cerealsdb.uk.net*

Chromosome-by-Chromosome

⇨ **Survey Seq. chr arms**

⇨ **BAC libraries**



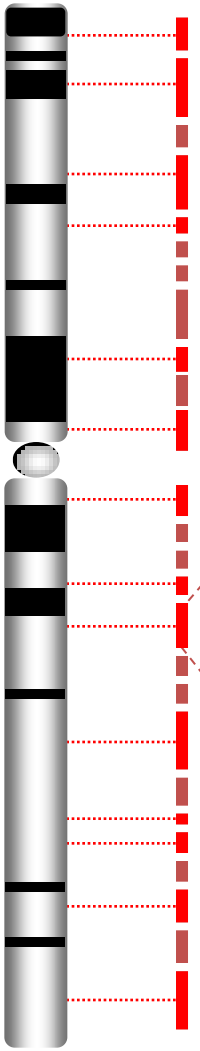
Flow cytometry

J. Dolezel (IEB)

An integrative map of chromosome 3B

Paux et al. Science 2008

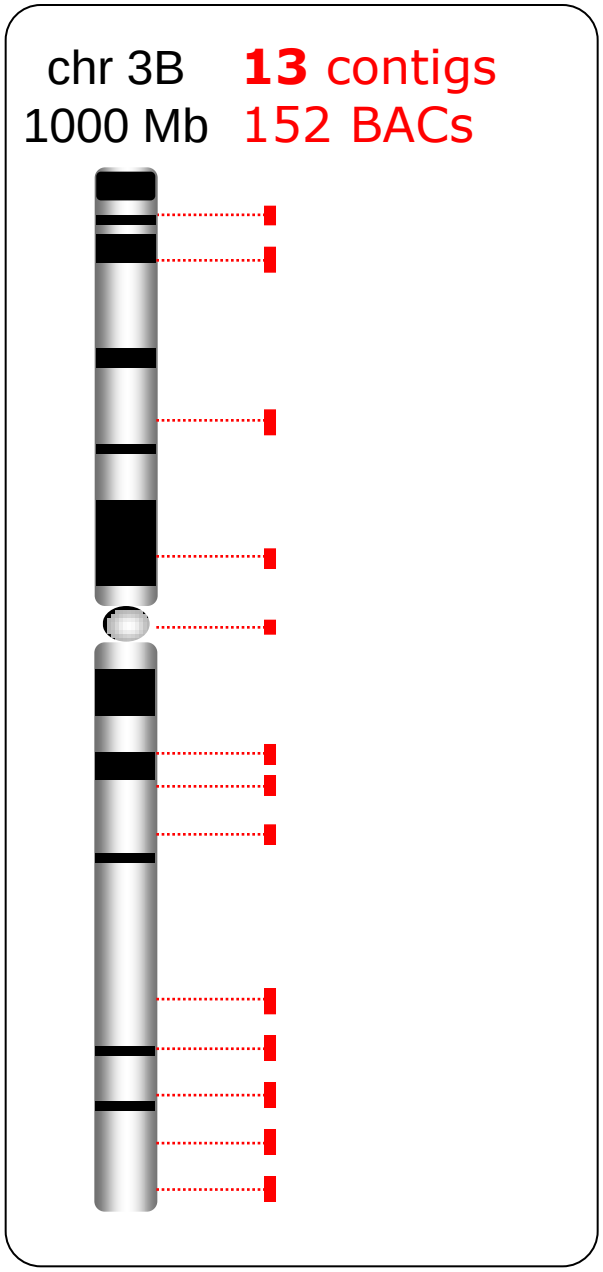
chr 3B
1000 Mb



Physical map

• Fingerprinted BACs	132,000 (19x)
• #contigs	1282
• #MTP BACs	8448
• Markers anchored	4463

3B Pilot (2007-2010)



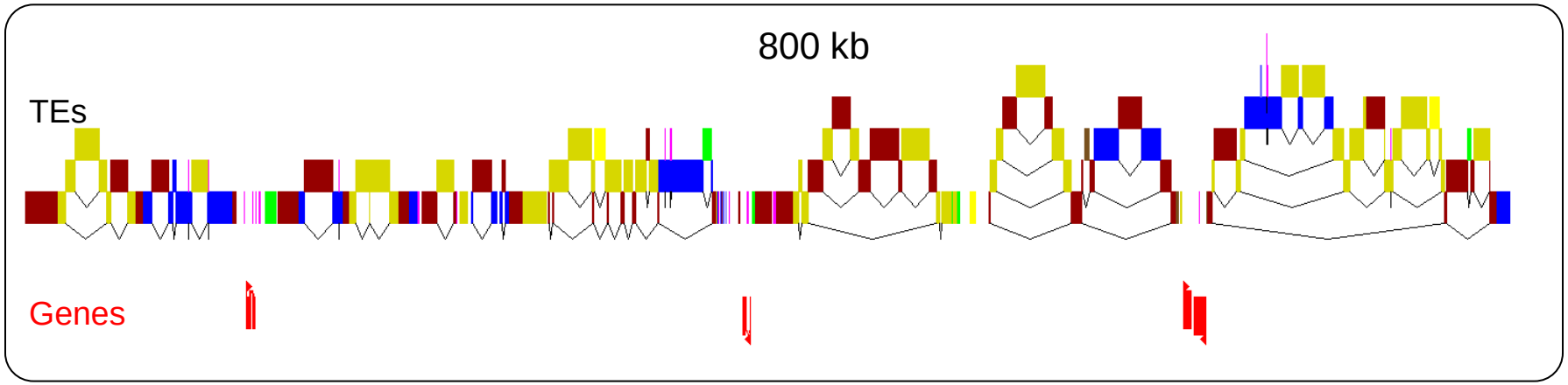
Choulet et al. Plant Cell 2010

18 Mb (**2%** of the 3B)
First set of large contiguous seq.

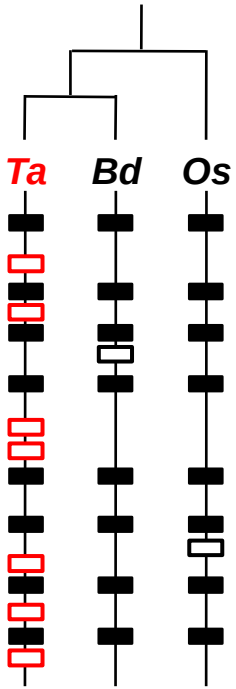
~3000 TEs

~200 protein-coding genes

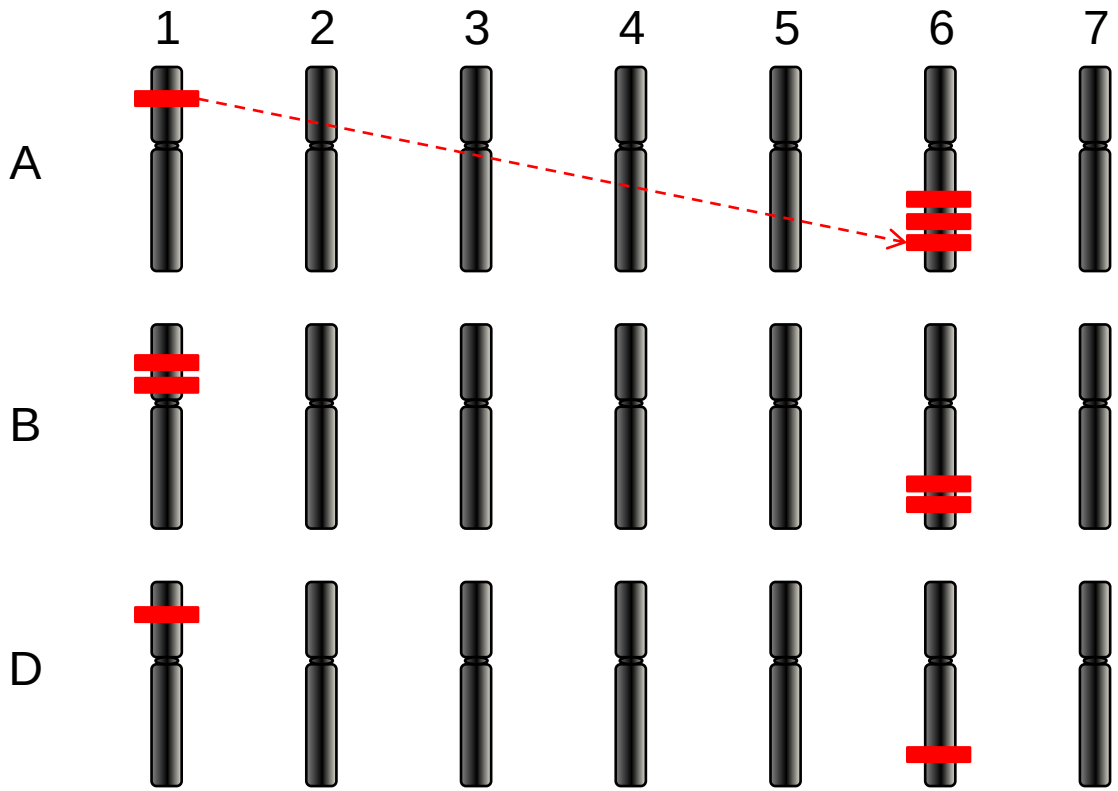
Main results



- 85% **TEs** highly **nested** and still **complete**
- 25% of **pseudogenes**
- 33% of **duplicates**
- No big **gene islands**
- Unexpected **high number of genes**
  (8400 *versus* 5500 expected based on synteny)
- 45% **non-syntenic** genes



Gene redundancy



3 **homoeologous** copies (ABD)

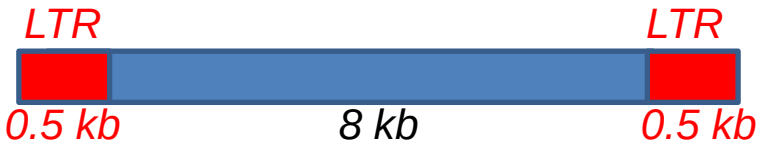
+ paralogs created by **interchromosomal duplication**

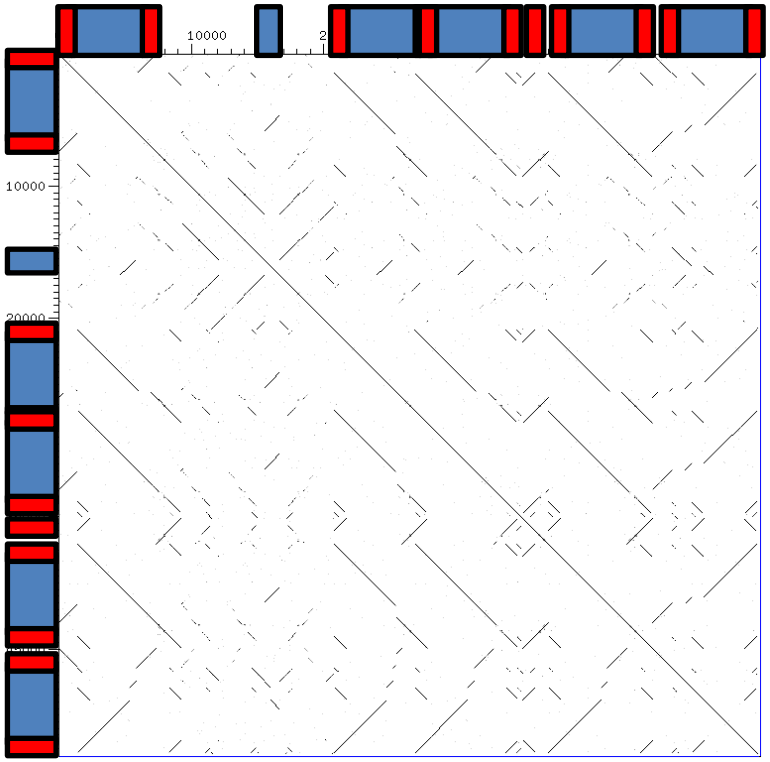
+ paralogs created by **tandem duplication**

/!\ assembly /!\

Repeats

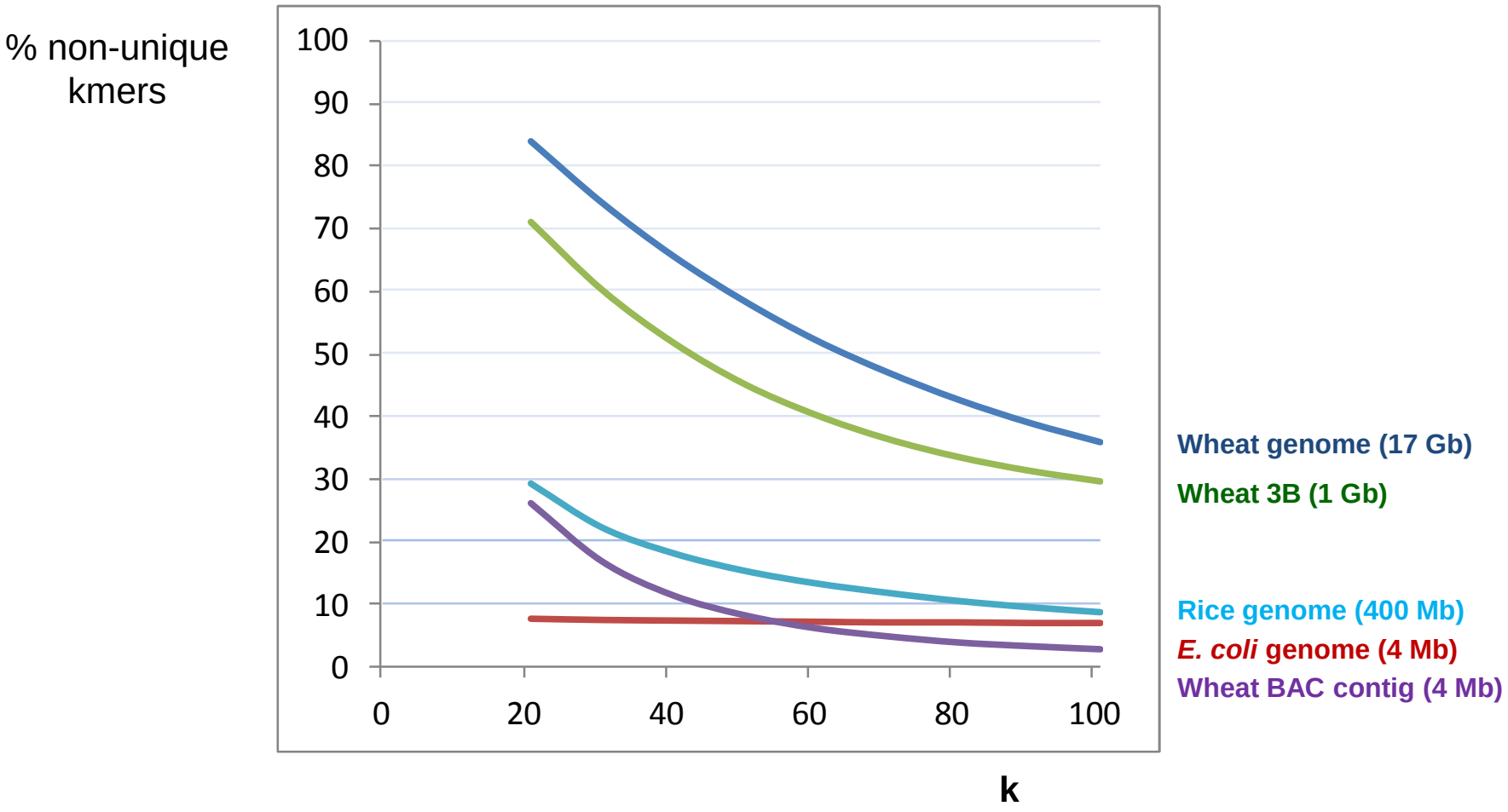
- Genes are repeated
- 85% TEs != 85% repeats
- TEs are made of repeats...





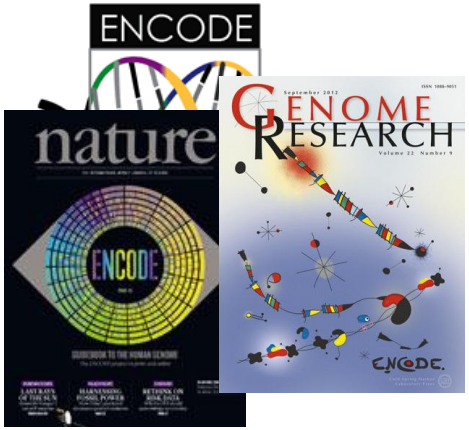
50 kb

kmers


% non-unique kmers
100
90
80
70
60
50
40
30
20
10
0
0
20
40
60
80
100
k
Wheat genome (17 Gb)
Wheat 3B (1 Gb)
Rice genome (400 Mb)
E. coli genome (4 Mb)
Wheat BAC contig (4 Mb)

A reference sequence? What for?

- Distinguish homoeologous + paralogous copies of gene families
- Solving the assembly of tandem duplicates (e.g. NBS-LRR involved in resistance)
- Study CNVs
- Study genome structure organization & impact on gene expression
- Identify genes under QTLs
- Distinguish genes *vs* pseudogenes
  pseudogenes in cv. Chinese Spring can be functional in others
- Assess what is behind the gene catalog!!!



Agence Nationale de la Recherche
ANR

FranceAgriMer

UMR 1095
GDEC
INRA
Université Blaise Pascal

GENOSCOPE
Centre National de Séquençage

URGI

Céréales Vallée

Sequencing the 3B

❑ Whole 3B Shotgun (Illumina)

Flow sorted DNA of 3B

contaminations

↓

Multiple Displacement Amplification

bias

↓

Illumina HiSeq2000

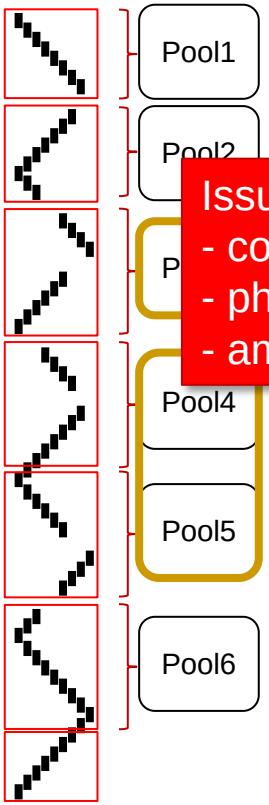
845 M reads (2x108) PE 0.5kb

→ 82X

assembly

❑ MTP Sequencing (Roche/454)

8448 BACs



928 pools of **8-12** BACs

928 libraries: LPE **8 kb**

Issues:
- cost & time
- physical map is incomplete
- amount of work to assemble

~150 runs GSFLX

↓

→ 40x coverage

❑ BAC-Ends of the MTP

Sequencing the 3B

❑ Whole 3B Shotgun (Illumina)

De novo assembly with ABYSS (k=71)

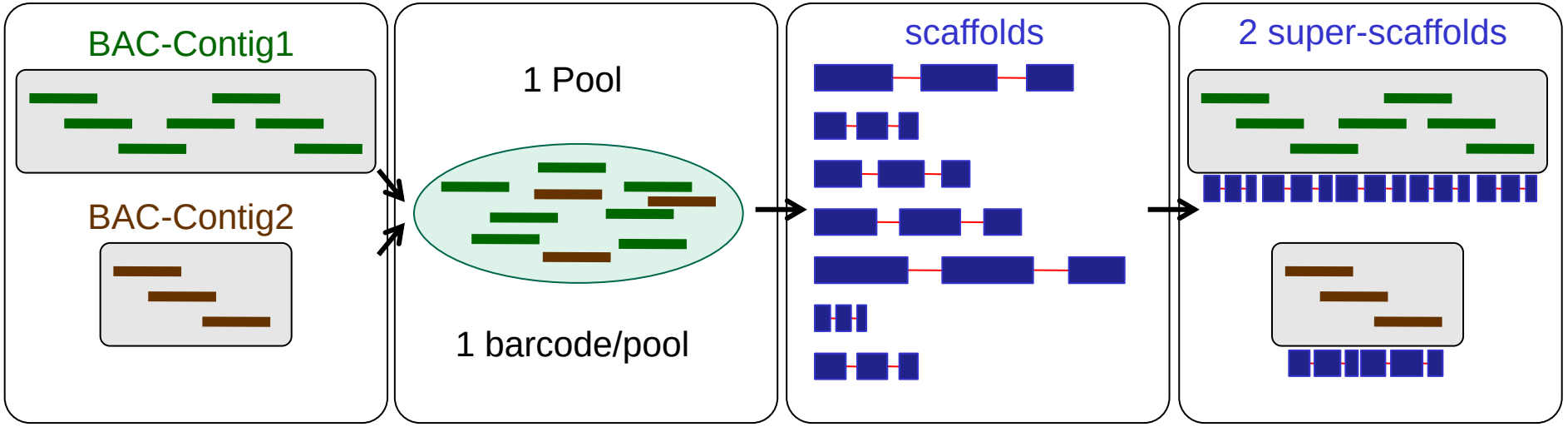
TGAC The Genome Analysis Centre™

J. Wright
M. Caccamo
J. Rogers

	Contigs >200bp	After Masking TEs
# Contigs	546,922	60,319
# bp	639 Mb	125 Mb
N50	2.7 kb	6.7 kb
Max ctg size	48 kb	48 kb

Sequencing the 3B

❑ **MTP Sequencing (Roche/454)**



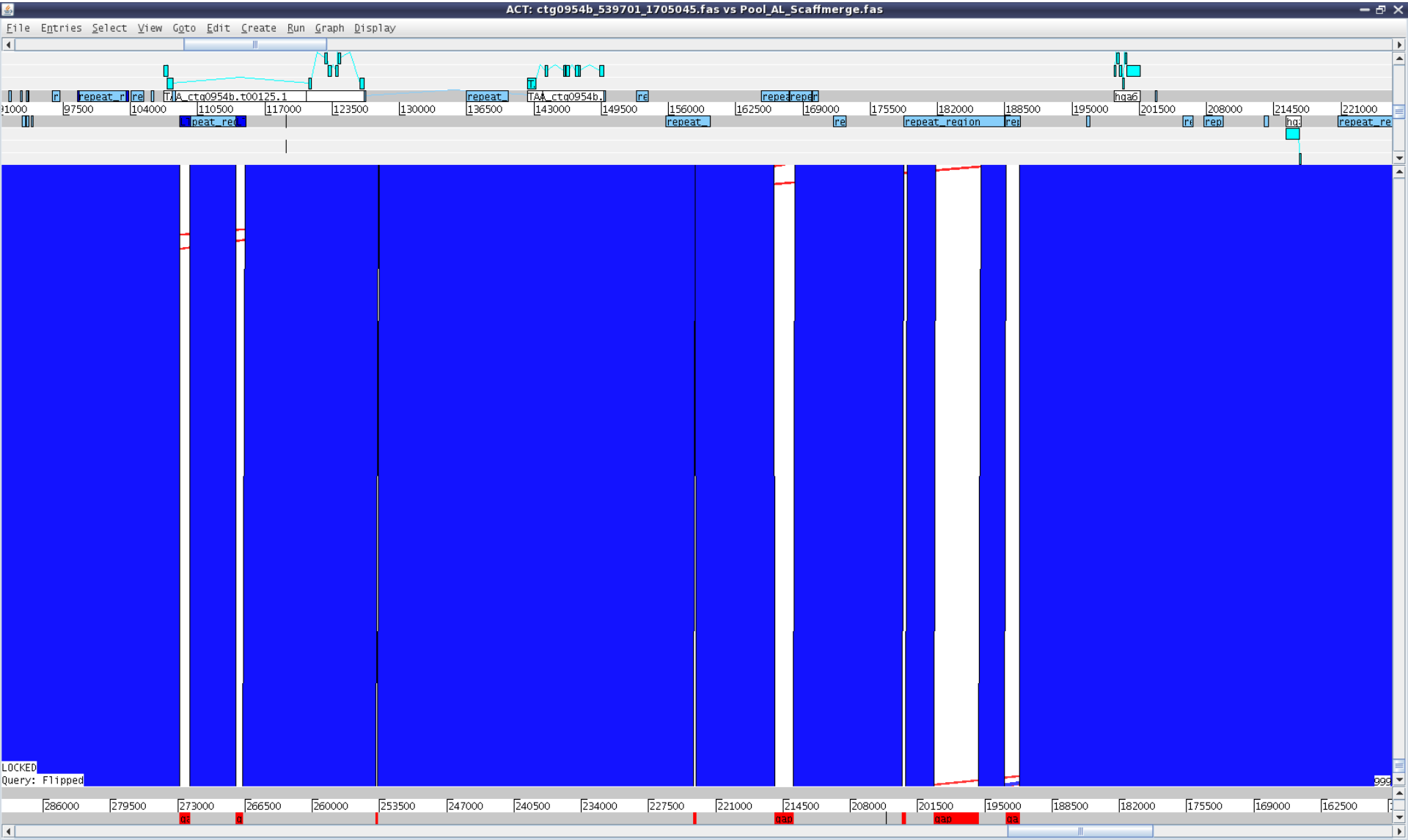
Physical map
x**1282 ctg**

Pooling
x**928 pools**

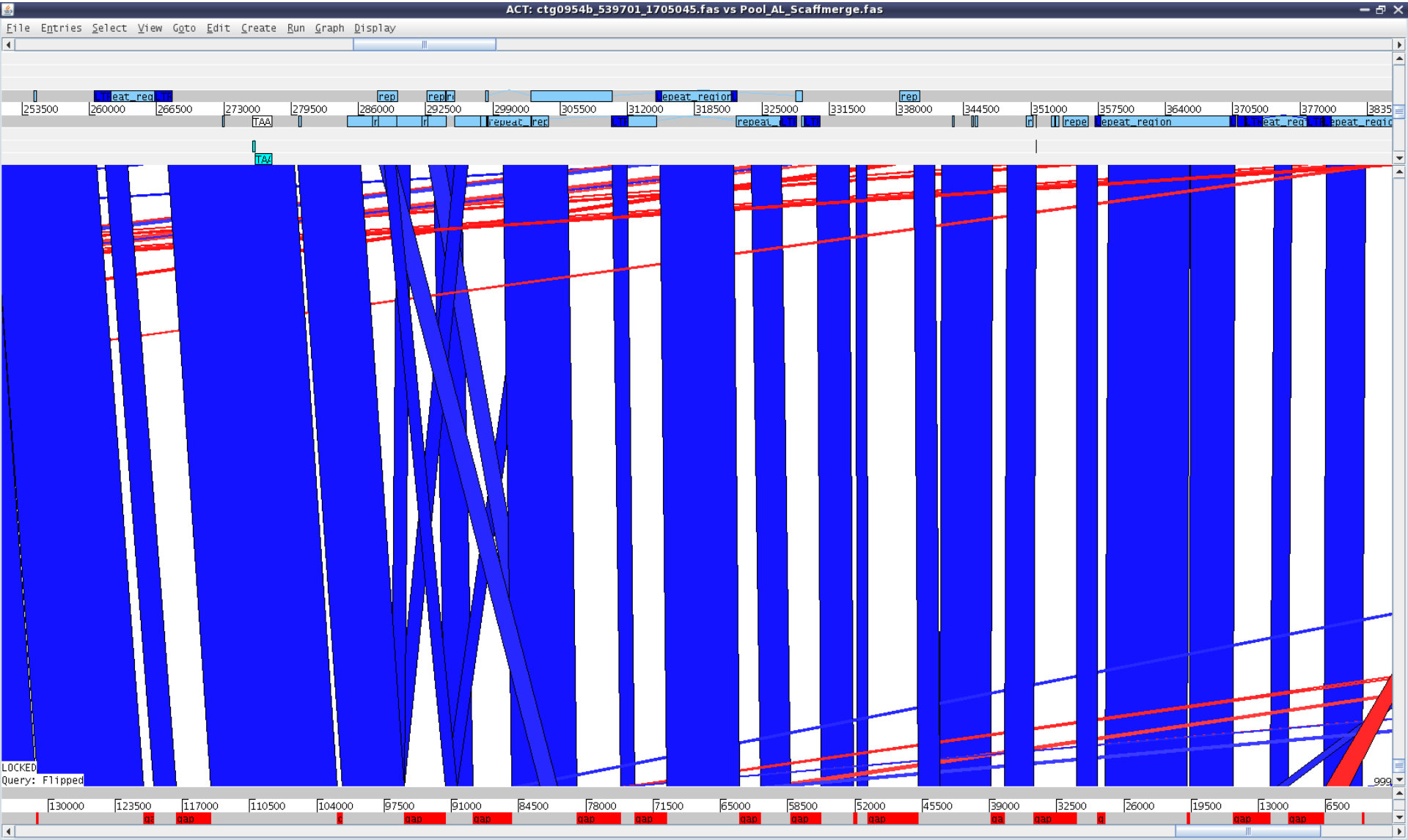
Sequencing +
Assembling

Assigning scaffolds
to physical contigs

Mapping on reference sequence



Mapping on reference sequence



What is missing in the 454-MTP-seq?

- How many sequences present on **Illumina-shotgun contigs** are missing in **MTP scaffolds**?

  - *mask TEs in Illumina contigs*
  - *similarity search against 454 assembly*

✓Match	x%
✓ Absent	x%

➔ Gaps in the physical map (5-10%)
➔ Contamination when flow sorting of 3B chromosomes (5-10%)

What is missing in the Illumina assembly?

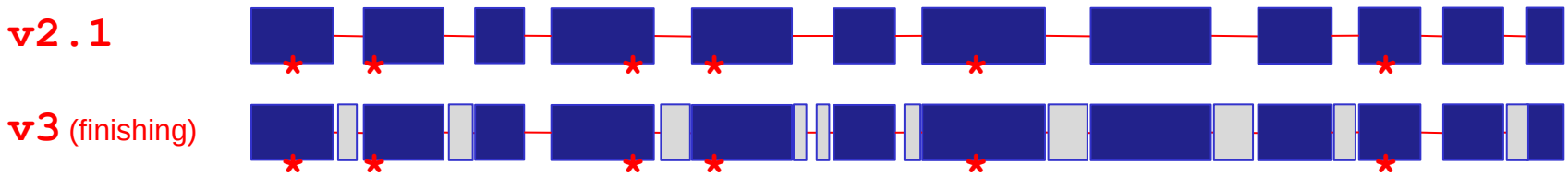
- How many sequences present on **MTP scaffolds** are missing in **Illumina-shotgun contigs**?

  - *extract ~33000 coding exons annotated on 454-scaffolds*
  - *similarity search against Illumina contigs>200 bp*

✓Full	x%
✓Partial	x%
✓ Absent	x%

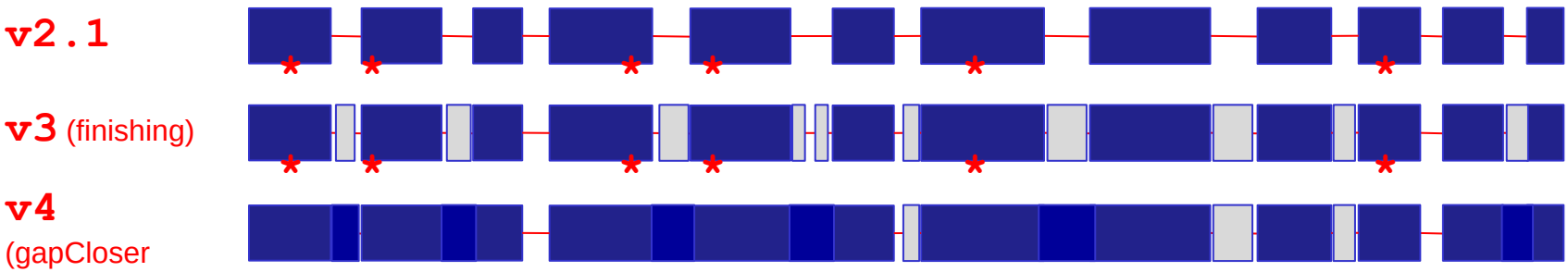
➔ Despite 82X Illumina coverage, x% of coding exons are absent/partial

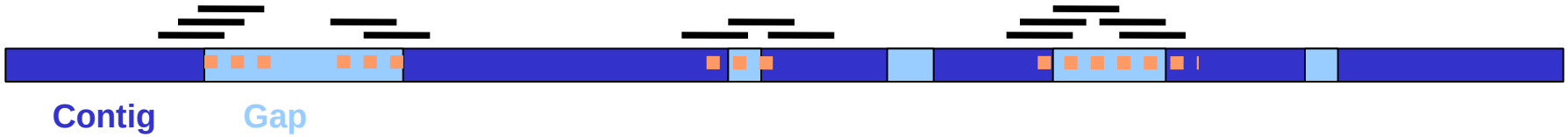
Improve scaffolding by manual curation



- **Curation** of the scaffolding (for the entire project [16000 scaffolds])
  - *Automated parsing of Newbler outputs to flag regions showing ambiguous / nonambiguous pairs*
  - *Decision taken by a curator*
    - ***!!! Time consuming !!!***

Improve 454 assembly with Ill reads





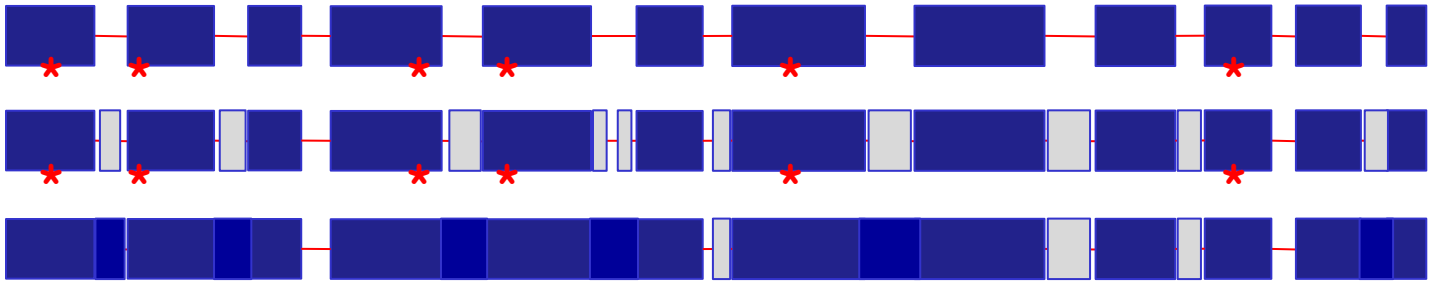
- ❑ **close gaps in the scaffolds**
  With **gapCloser** (JM. Aury et al.)

Gaps:
18% → x%

- ❑ **correct homopolymer errors and mismatches**

xxxxxx bases corrected
(0.1% of low copy regions)

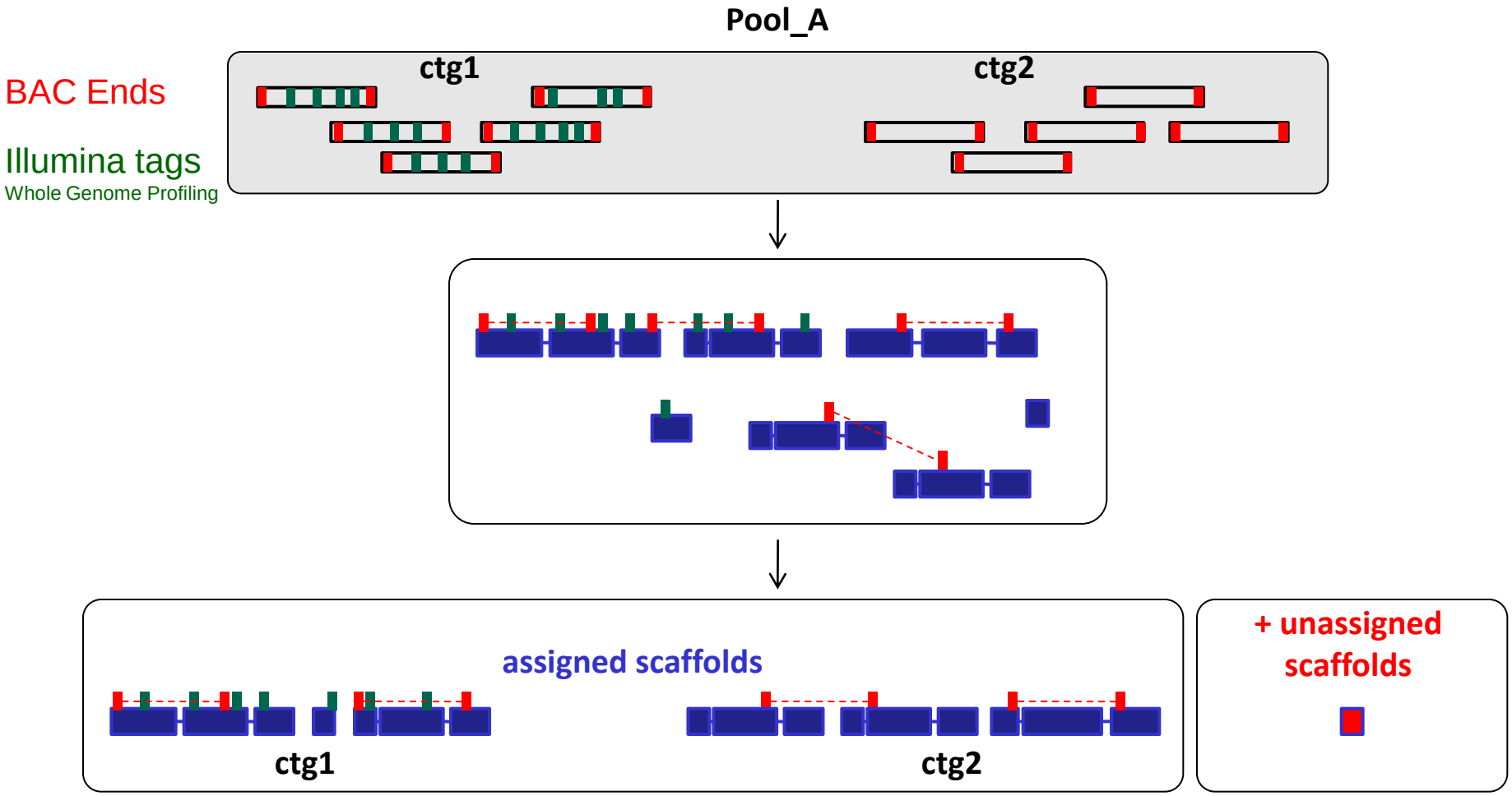
Improve scaffolding



	❑ V2.1	❑ V4	
#scaff	16136 scaff	xxxx scaff	x1/3
Cumul. size	1040 Mb	xxx Mb	
Gaps	18 %	x %	
N50	275 kb	xxx kb	+70%

➡ Route towards a pseudomolecule

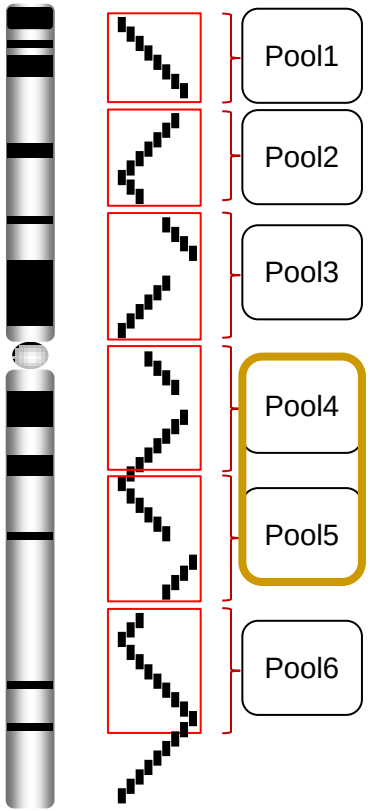
Assignation scaffolds ⇔ BAC-contigs

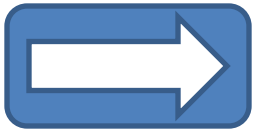


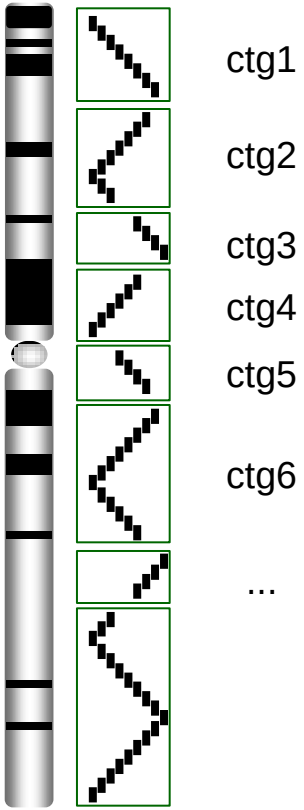
Assignation scaffolds ⇔ BAC-contigs

Assigned: xxx Mb

Unassigned: xx Mb





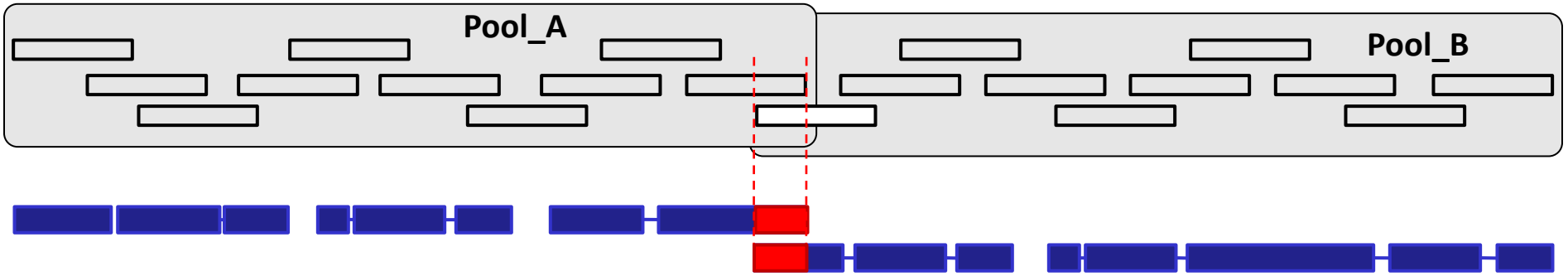


xxxx scaffolds

x scaff / BAC-contig

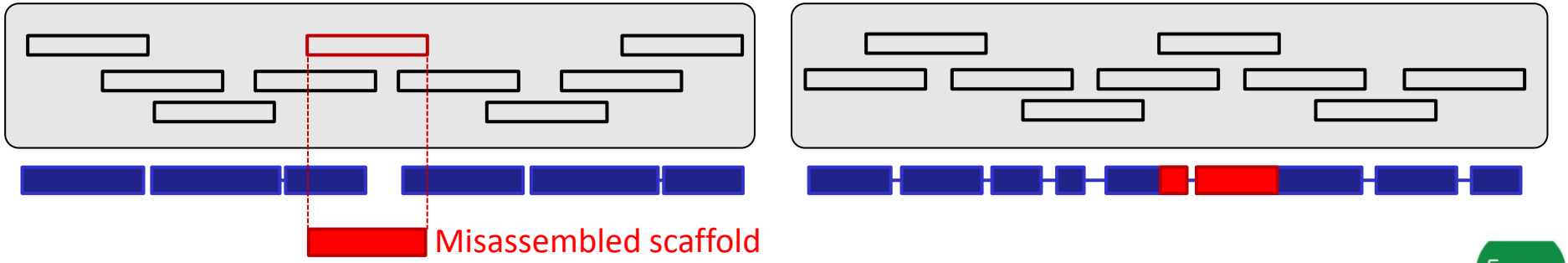
Detect & assemble overlapping scaffolds

❑ Overlapping BAC-contigs



- *Expected overlap*: **525** pools
- *Unexpected overlap*: ???

❑ Misassembled BAC clones



Using *scaffAssembler* (S. Theil)

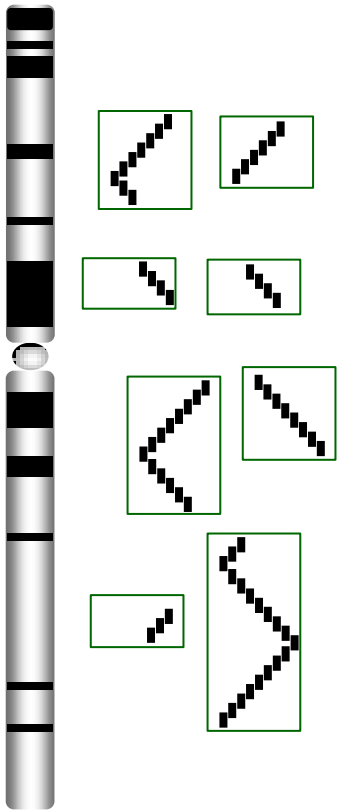
- BLAT --extendThroughNs
- Parse with 99%id & >10kb overlap at extremities
- Merge

1. Merge scaffolds between pools with expected overlap

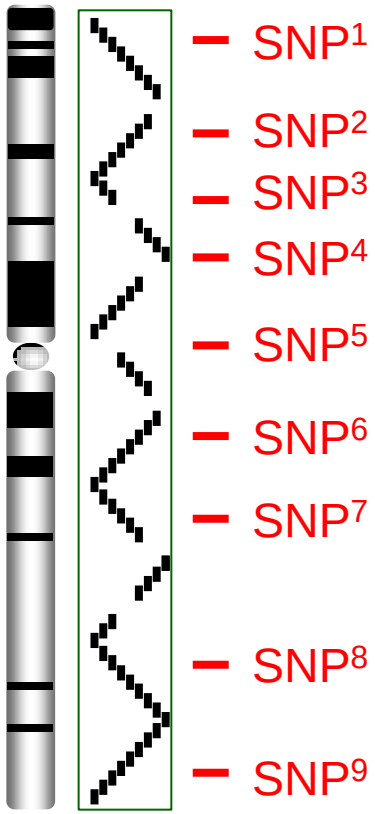
2. Search for overlap between BAC-contigs anchored in the same region

	❑ V4	❑ V4.2	❑ V4.3
#scaff	xxxx scaff	xxxx scaff	xxx scaff
Cumul. size	xxx Mb	xxx Mb	xxx Mb
N50	xxx kb	xxx kb (xxx scaff)	xxx kb (xxx scaff)

Order/Orientate scaffolds with markers



Our objective



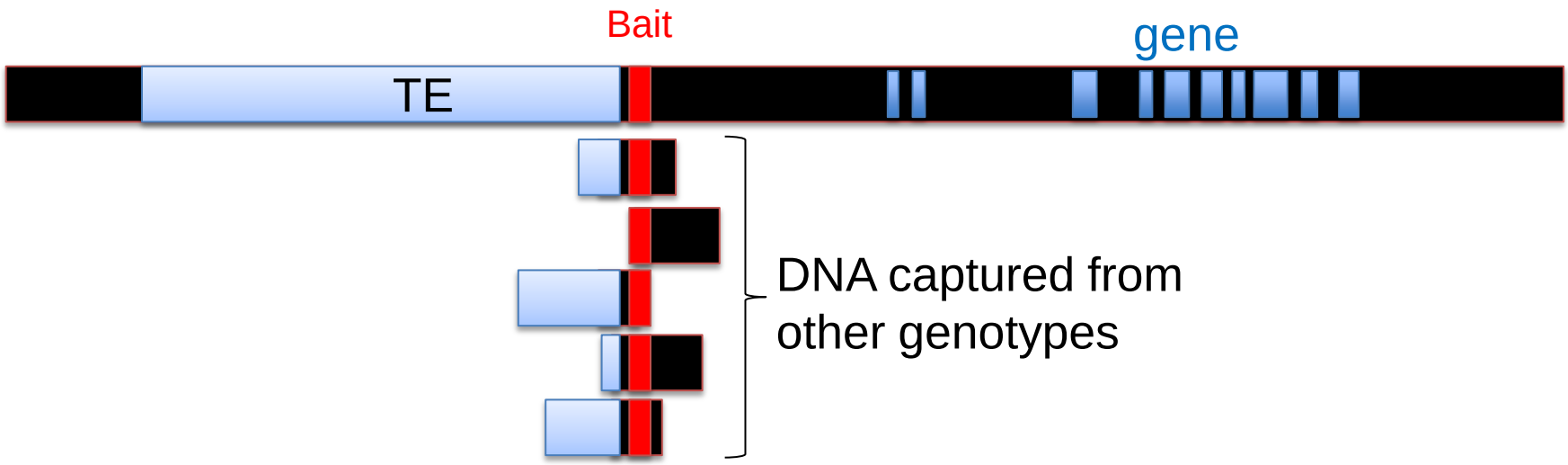
❑ Unordered BAC-contigs assembled into xxxx scaffolds

❑ 1 pseudomolecule

Order/Orientate scaffolds

- SNP discovery with *SureSelect*® Seq. Capture

(*E. Paux, N. Cubizolles, E. Rey*)



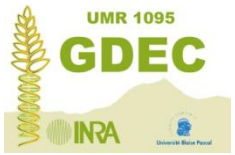
- xxxxx SNPs found on scaffolds accounting for xxx Mb
  - xxxx chr-wide SNPs chosen for genotyping

Biological analyses undertaken

- ❑ Annotation with TriAnnot
- ❑ Comparative genomics
- ❑ TE dynamics
- ❑ RNASeq data analyses
- ❑ CNVs
- ❑ Recombination/LD studies

+ 15 map-based cloning projects on 3B

Acknowledgements



Catherine Feuillet
Etienne Paux
Pierre Sourdille
Philippe Leroy
Nicolas Guilhot
Lise Pingault
Josquin Daron
Natasha Glover
Sébastien Theil

Patrick Wincker
Adriana Alberti
Valérie Barbe
Sophie Mangenot
Jean-Marc Aury
Arnaud Couloux

Hadi Quenesville
Michael Alaux
Claire Viseux

Hélène Berges
Arnaud Bellec

Scientific Advisory Board:
J. Rogers, P. Schnable, D. Ware,
S. Rounsley, K. Eversole