



HAL
open science

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.

► To cite this version:

Frédéric Choulet, Sébastien Theil, Josquin Daron, Natasha Marie Glover, Nicolas N. Guilhot, et al.. A strategy to assemble a reference sequence of the 1 Gb wheat chromosome 3B. TGAC/ESF-LESC workshop, Strategies for de novo assemblies of complex crop genomes, Genome Analysis Centre (TGAC). Norwich, GBR., Oct 2012, Norwich, United Kingdom. hal-02809032

HAL Id: hal-02809032

<https://hal.inrae.fr/hal-02809032>

Submitted on 6 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

TGAC/ESF-LESC workshop

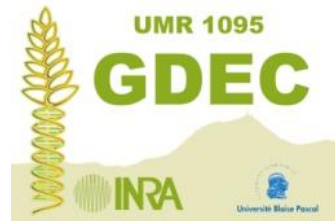
Strategies for de novo assemblies of complex crop genomes

Norwich, Oct-9-2012

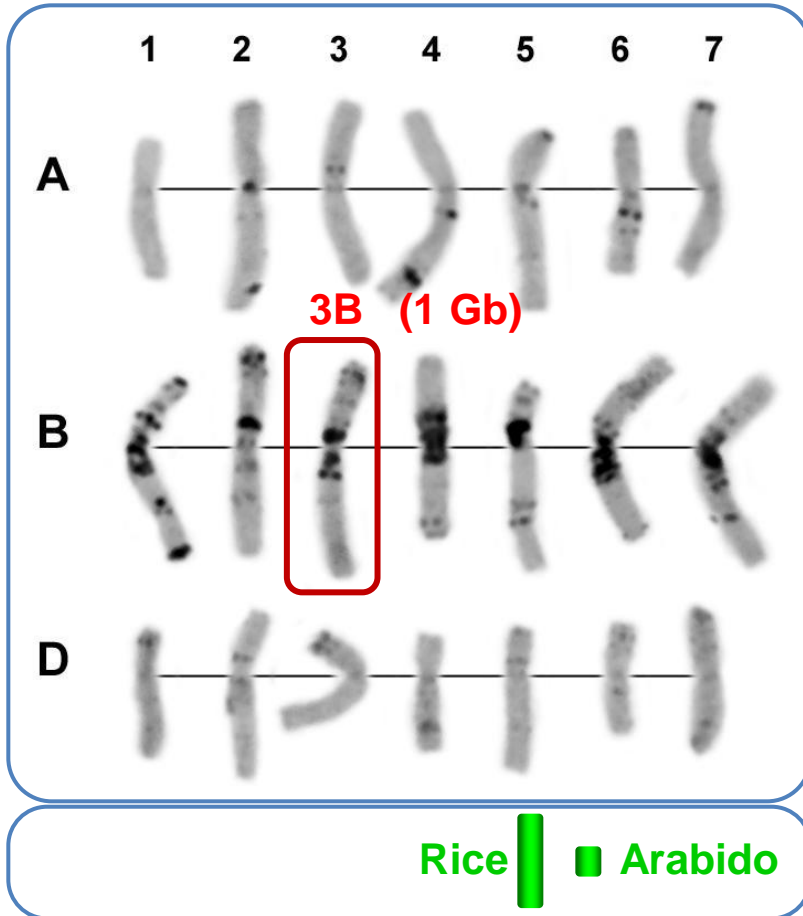


Sequencing and assembling wheat chromosome 3B

Frédéric CHOULET



The bread wheat genome



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

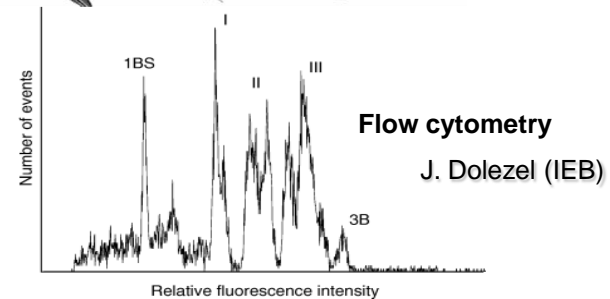
WGS

5x @ www.cerealsdb.uk.net

Chromosome-by-Chromosome

⇒ Survey Seq. chr arms

⇒ BAC libraries



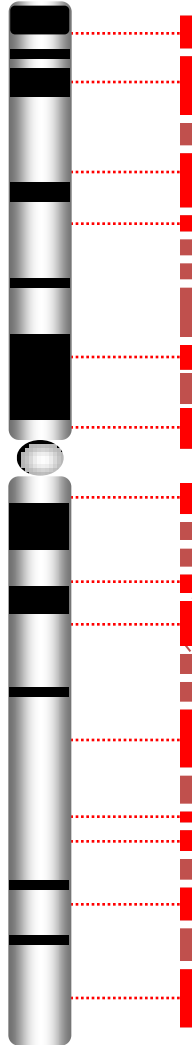
An integrative map of chromosome 3B

Paux et al. Science 2008

chr 3B
1000 Mb

Physical map

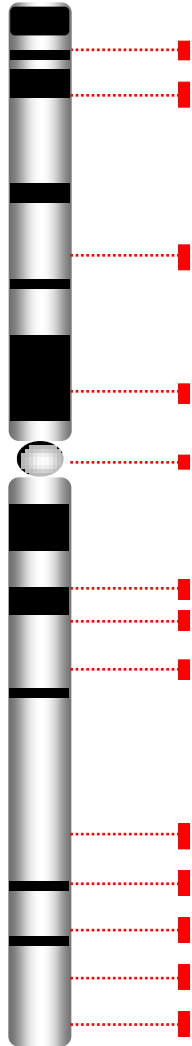
- Fingerprinted BACs 132,000 (19x)
- #contigs 1282
- #MTP BACs 8448
- Markers anchored 4463



64L02~	TaaCsp3BF302C03~	TaaCsp3BF069M13~	TaaCsp3BF082I7	TaaCsp3BF082I7
77K16~	TaaCsp3BF381A10*	TaaCsp3BF064B20	TaaCsp3BF132M22	TaaCsp3BF132M22
83BF036B24	TaaCsp3BF352E22	TaaCsp3BF191I18~	TaaCsp3BF169J21~	TaaCsp3BF169J21~
87G16*	TaaCsp3BF384D02*	TaaCsp3BF218B22~	TaaCsp3BF366P23	TaaCsp3BF366P23
104I05	TaaCsp3BF261M19~	TaaCsp3BF237K01*	TaaCsp3BF102N04*	TaaCsp3BF102N04*
16J06~	TaaCsp3BF39K16*	TaaCsp3BF204F19	TaaCsp3BF337C12~	TaaCsp3BF300E
194	TaaCsp3BF091N08	TaaCsp3BF015F07~	TaaCsp3BF262L17~	TaaCsp3BF262L17~
	TaaCsp3BF039L10	TaaCsp3BF015F07*	TaaCsp3BF266N22*	TaaCsp3BF266N22*
0A08*	TaaCsp3BF215L11~	TaaCsp3BF225E15	TaaCsp3BF267P01~	TaaCsp3BF267P01~
	TaaCsp3BF102B17	TaaCsp3BF027H08	TaaCsp3BF199J18	TaaCsp3BF199J18
	TaaCsp3BF047F16~	TaaCsp3BF073P12	TaaCsp3BF112G08	TaaCsp3BF112G08
	TaaCsp3BF168N19	TaaCsp3BF098I09	TaaCsp3BF112G08~	TaaCsp3BF112G08~
	TaaCsp3BF236D13	TaaCsp3BF073P12*	TaaCsp3BF001H17~	TaaCsp3BF001H17~
	TaaCsp3BF353D21*	TaaCsp3BF007N19~	TaaCsp3BF376D15*	TaaCsp3BF376D15*
N15	TaaCsp3BF380N03	TaaCsp3BF007N19	TaaCsp3BF001G18*	TaaCsp3BF001G18*
9	TaaCsp3BF058F18~	TaaCsp3BF103E15*	TaaCsp3BF324H04	TaaCsp3BF324H04
17*	TaaCsp3BF179K19	TaaCsp3BF354O03	TaaCsp3BF295B04	TaaCsp3BF295B04
	TaaCsp3BF112E04*	TaaCsp3BF006E14	TaaCsp3BF233F19	TaaCsp3BF233F19
	TaaCsp3BF032G11	TaaCsp3BF106E15	TaaCsp3BF375P20~	TaaCsp3BF375P20~
	TaaCsp3BF341K09	TaaCsp3BF225G08	TaaCsp3BF002E09	TaaCsp3BF002E09
	TaaCsp3BF066L16	TaaCsp3BF033P14	TaaCsp3BF295B06*	TaaCsp3BF295B06*
	TaaCsp3BF081E14	TaaCsp3BF188M11	TaaCsp3BF179G14	TaaCsp3BF179G14
TaaCsp3BF215N04~	TaaCsp3BF112P10~	TaaCsp3BF386I11~	TaaCsp3BF005E04	TaaCsp3BF005E04
	TaaCsp3BF065O02	TaaCsp3BF233G17	TaaCsp3BF228K24*	TaaCsp3BF228K24*
TaaCsp3BF346A09*	TaaCsp3BF346B19	TaaCsp3BF199K07	TaaCsp3BF028J09	TaaCsp3BF028J09
TaaCsp3BF079D10	TaaCsp3BF353I07	TaaCsp3BF061H10	TaaCsp3BF087I07	TaaCsp3BF087I07
F162A22	TaaCsp3BF323H20~	TaaCsp3BF315O08	TaaCsp3BF016A17	TaaCsp3BF016A17
aCsp3BF079D10	TaaCsp3BF381B09~	TaaCsp3BF237J16	TaaCsp3BF153F13~	TaaCsp3BF153F13~
B3G20~	TaaCsp3BF112K02	TaaCsp3BF128G13~	TaaCsp3BF041E21	TaaCsp3BF041E21

3B Pilot (2007-2010)

chr 3B **13 contigs**
1000 Mb **152 BACs**



Choulet et al. Plant Cell 2010

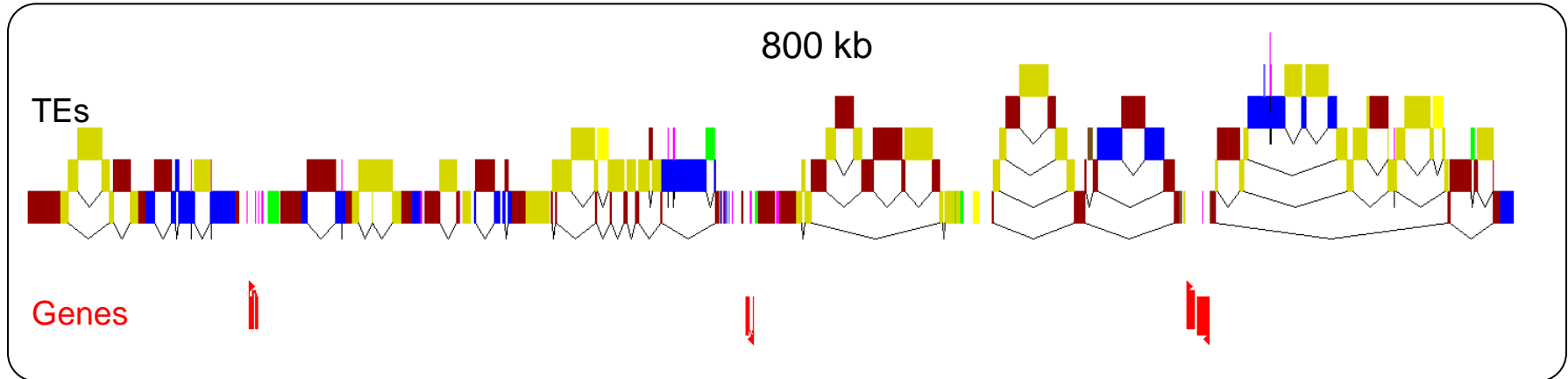
18 Mb (2% of the 3B)
First set of large contiguous seq.

~3000 TEs

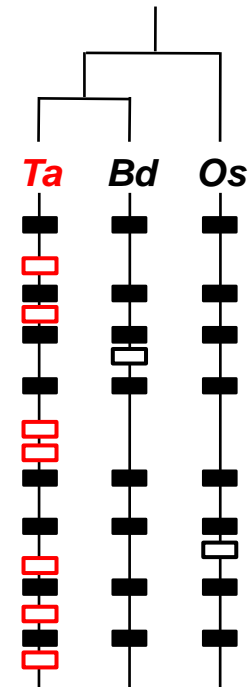
~200 protein-coding genes



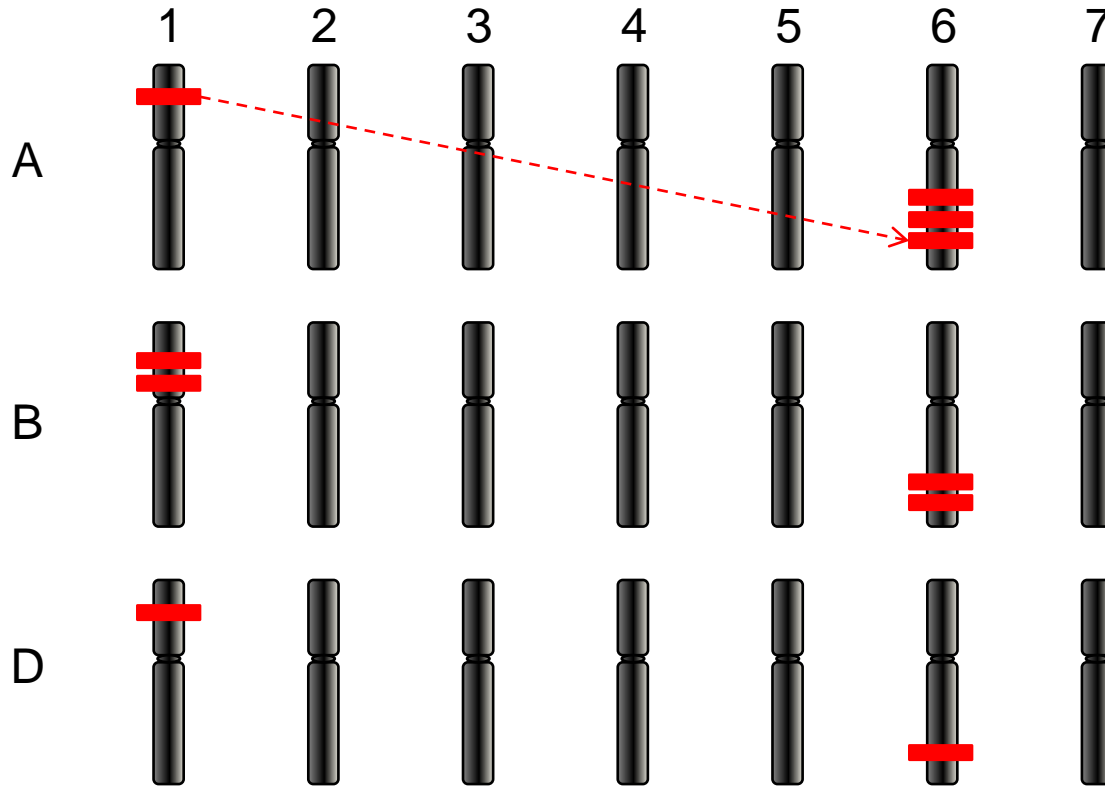
Main results



- ❑ 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)

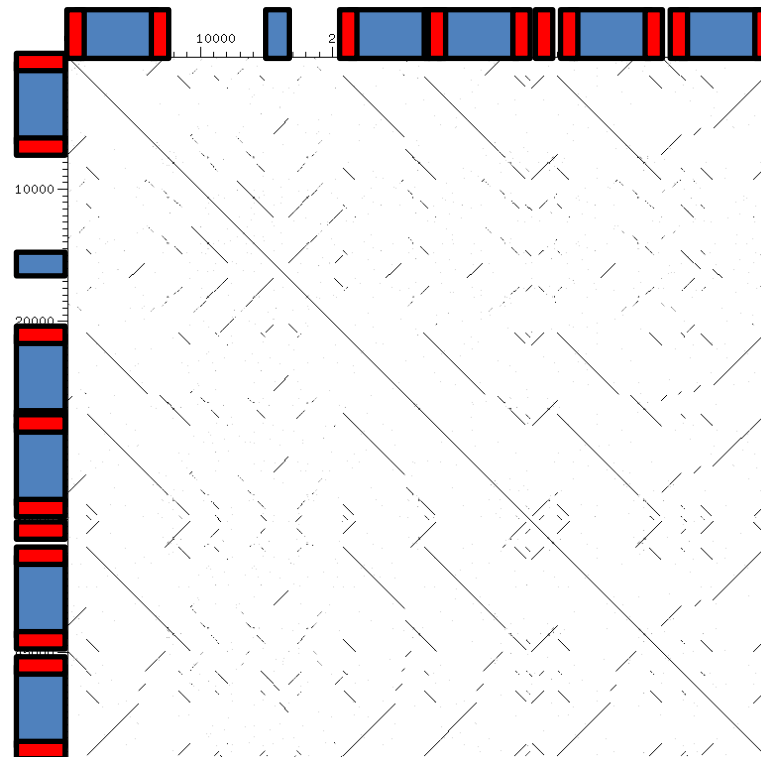
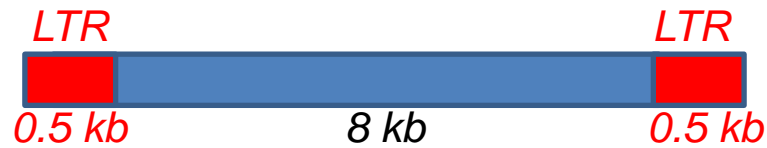
+ paralogs created by **interchromosomal duplication**

+ paralogs created by **tandem duplication**

/!\ assembly /!\

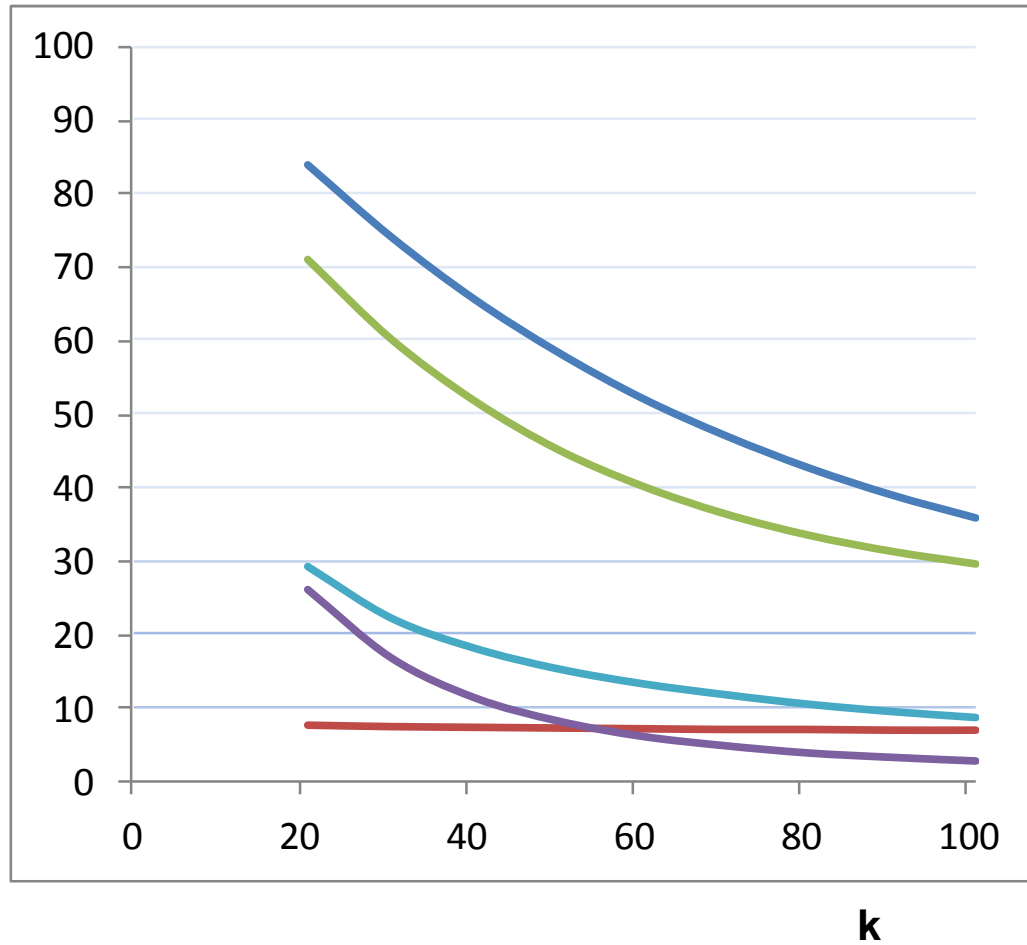
Repeats

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



kmers

% non-unique
kmers



Wheat genome (17 Gb)

Wheat 3B (1 Gb)

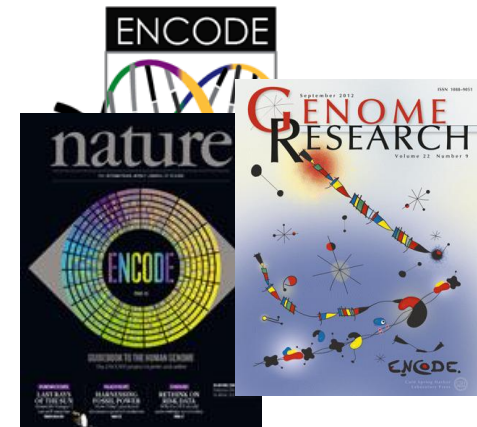
Rice genome (400 Mb)

E. coli genome (4 Mb)

Wheat BAC contig (4 Mb)

A reference sequence? What for?

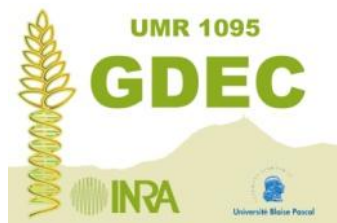
- ❑ Distinguish **homoeologous** + **paralogous copies** of gene families
- ❑ Solving the assembly of **tandem duplicates** (e.g. NBS-LRR involved in resistance)
- ❑ Study **CNVs**
- ❑ Study genome structure organization & impact on **gene expression**
- ❑ Identify genes under **QTLs**
- ❑ Distinguish **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



Sequencing the 3B

□ Whole 3B Shotgun (Illumina)

Flow sorted DNA of 3B

contaminations



Multiple Displacement Amplification

bias



Illumina HiSeq2000

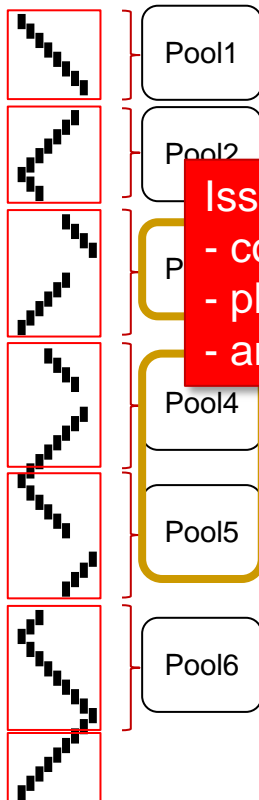
845 M reads (2x108) PE 0.5kb

→ 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)



J. Wright
M. Caccamo
J. Rogers

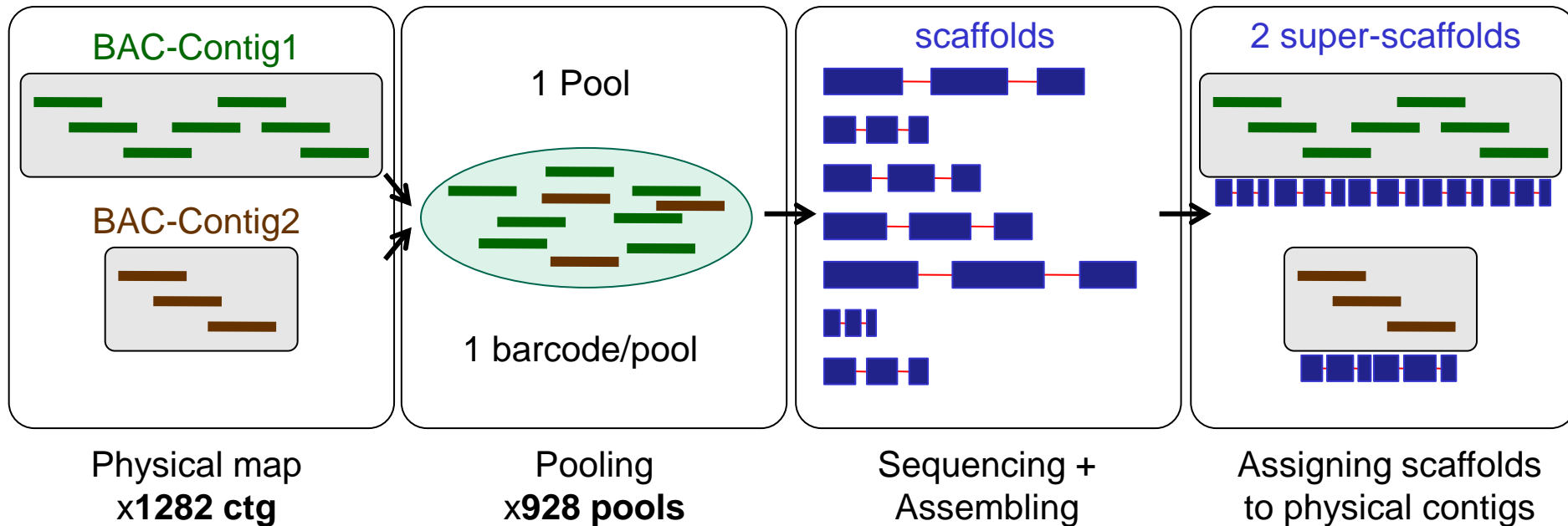
De novo assembly with **ABYSS** (k=71)

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

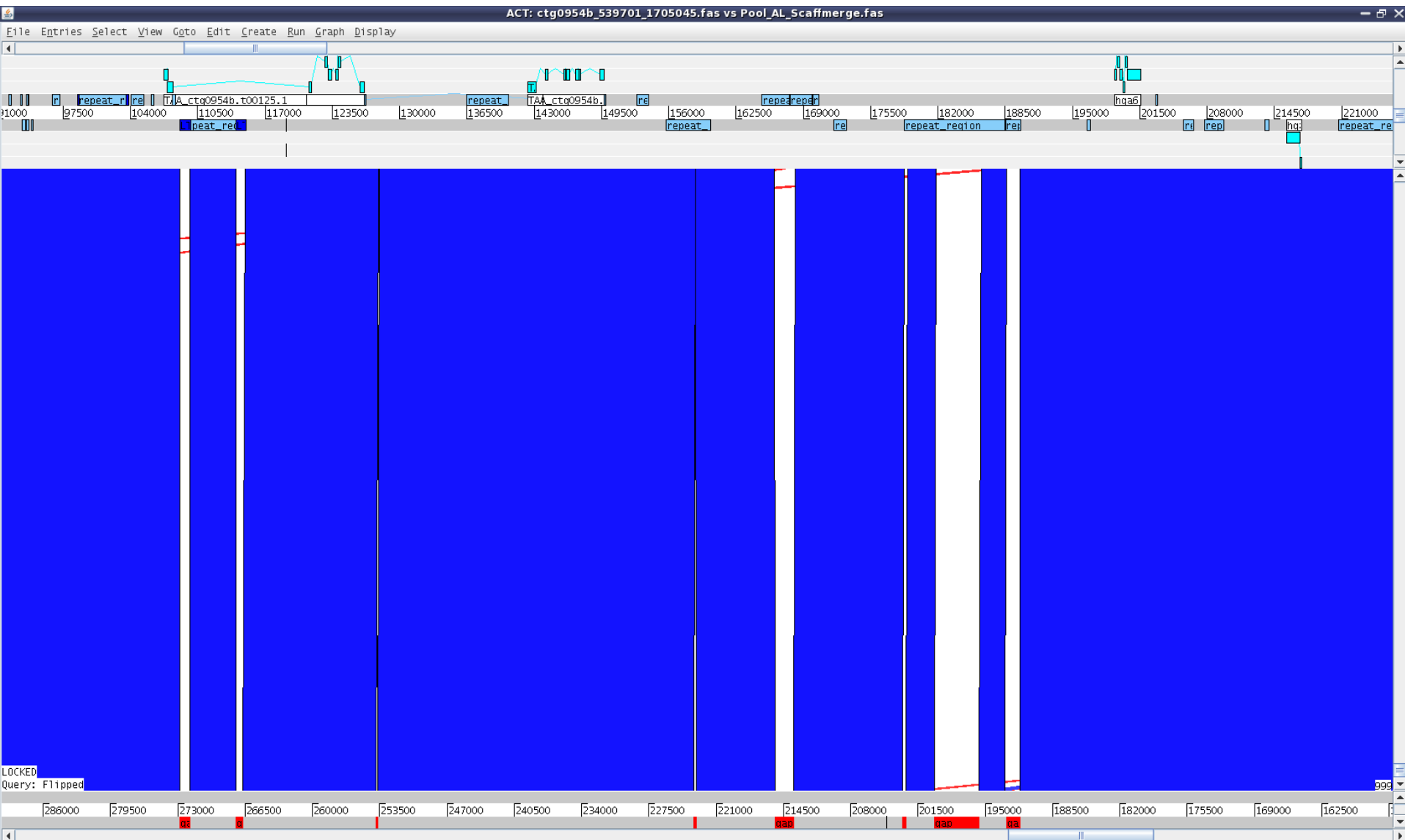


Sequencing the 3B

□ MTP Sequencing (Roche/454)



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**?

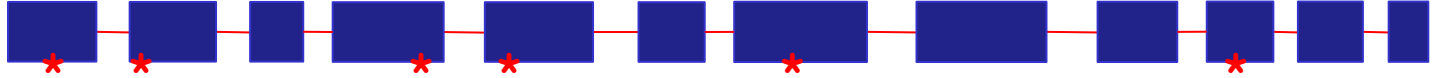
- *extract ~33000 coding exons annotated on 454-scaffolds*
- *similarity search against Illumina contigs >200 bp*

✓ Full	x%
✓ Partial	x%
✓ Absent	x%

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

Improve scaffolding by manual curation

v2.1



v3 (finishing)



□ **Curation** of the scaffolding (for the entire project [16000 scaffolds])

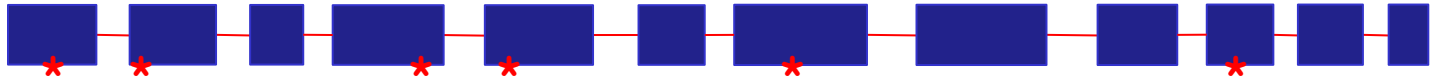
- *Automated parsing of Newbler outputs to flag regions showing ambiguous / nonambiguous pairs*
- *Decision taken by a curator*

!!! Time consuming !!!

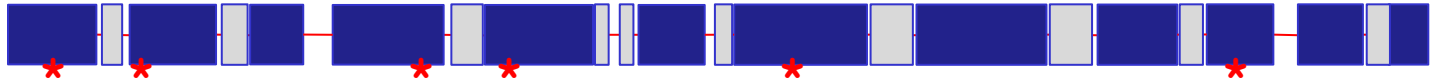


Improve 454 assembly with I1 reads

v2.1

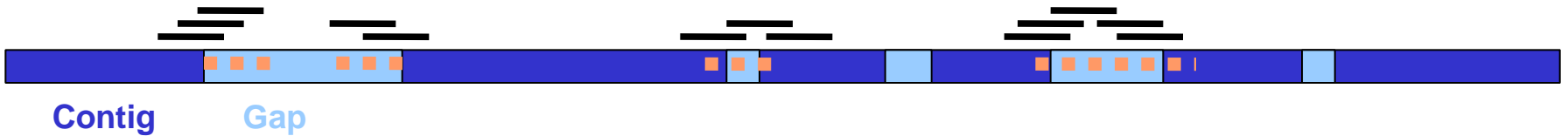


v3 (finishing)



v4

(gapCloser
+ ssrCorrection)



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

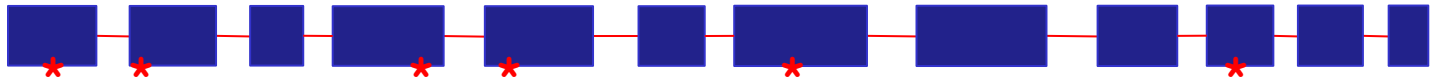
→ Gaps:
18% → x%

- correct homopolymer errors and mismatches

→ xxxxxx bases
corrected
(0.1% of low copy
regions)

Improve scaffolding

v2.1



v3 (finishing)



v4

(gapCloser
+ ssrCorrection)



□ V2.1

□ V4

#scaff

16136 scaff

xxxx scaff

x1/3

Cumul. size

1040 Mb

xxx Mb

Gaps

18 %

x %

N50

275 kb

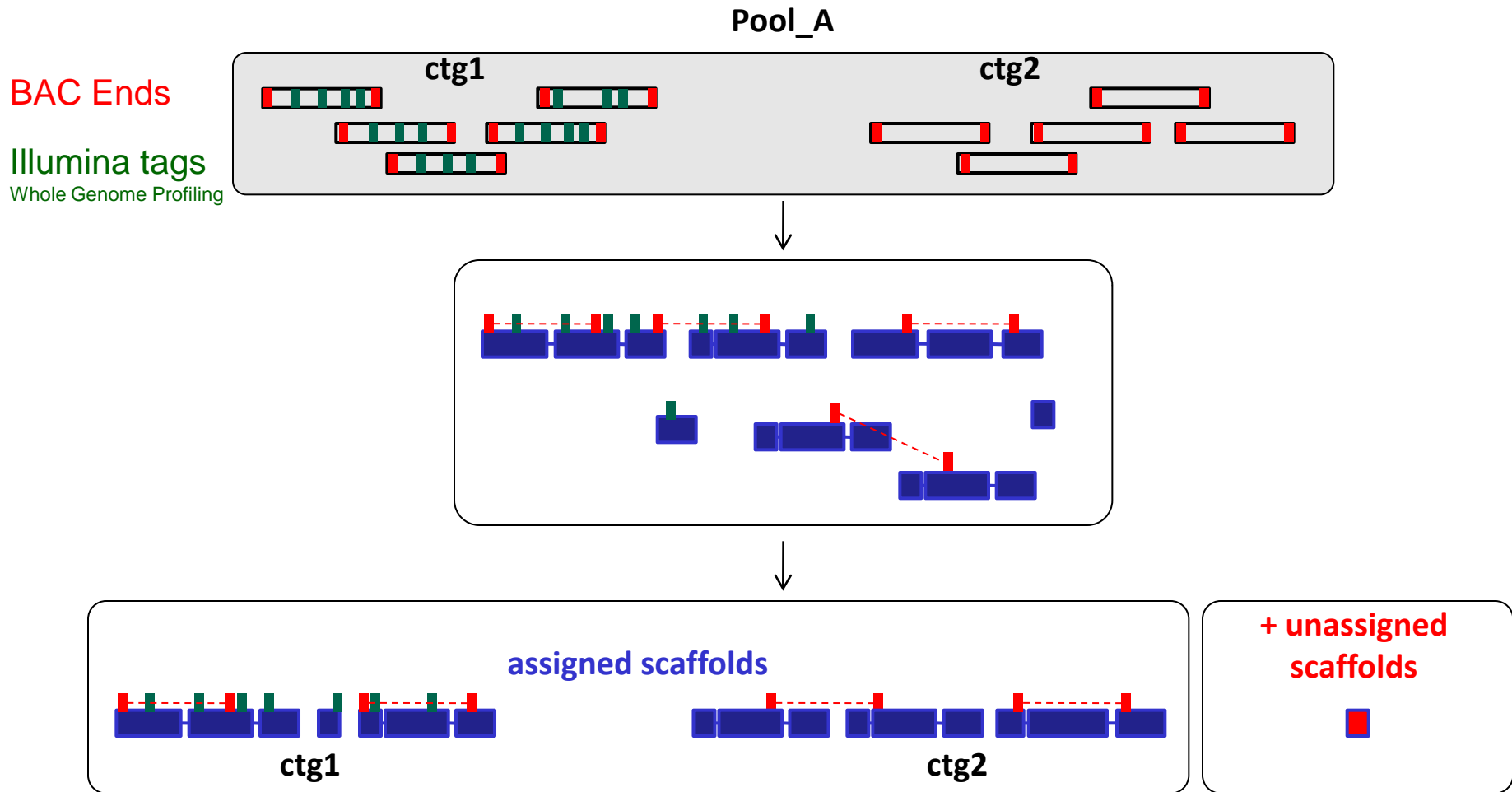
xxx kb

+70%



Route towards a pseudomolecule

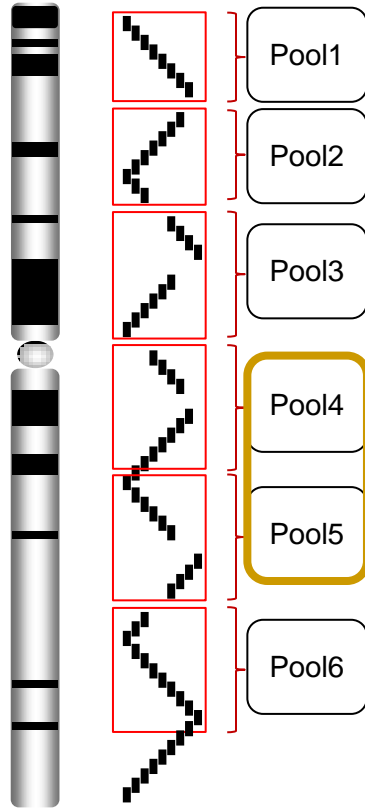
Assignment scaffolds ⇔ BAC-contigs



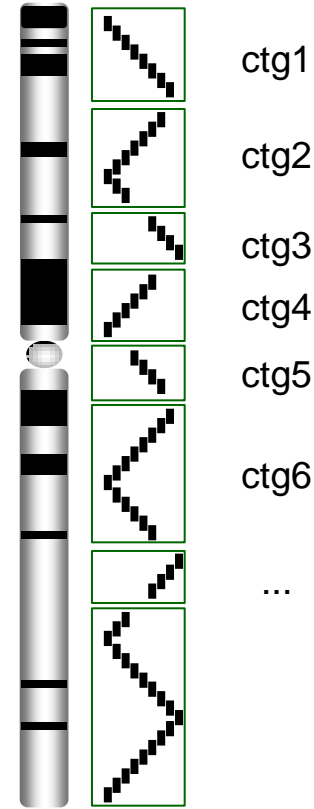
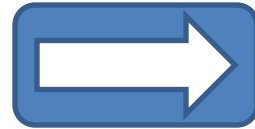
Assignment scaffolds ⇔ BAC-contigs

Assigned: xxx Mb

Unassigned: xx Mb



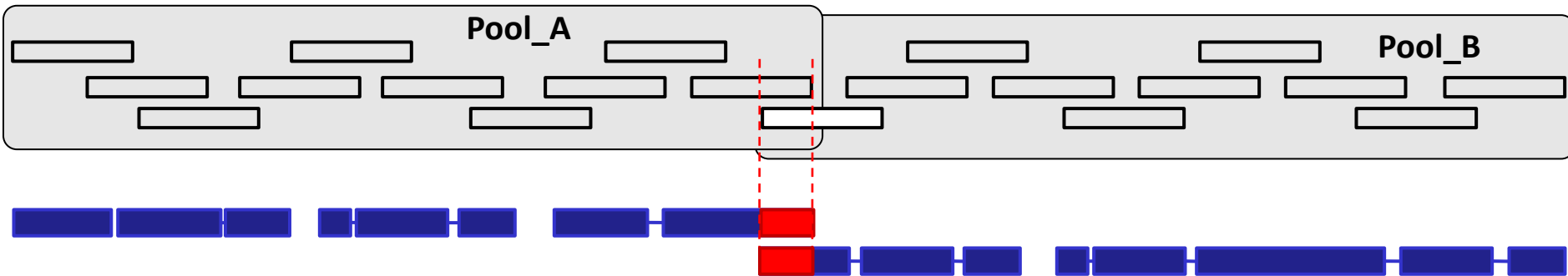
xxxx scaffolds



x scaff / BAC-contig

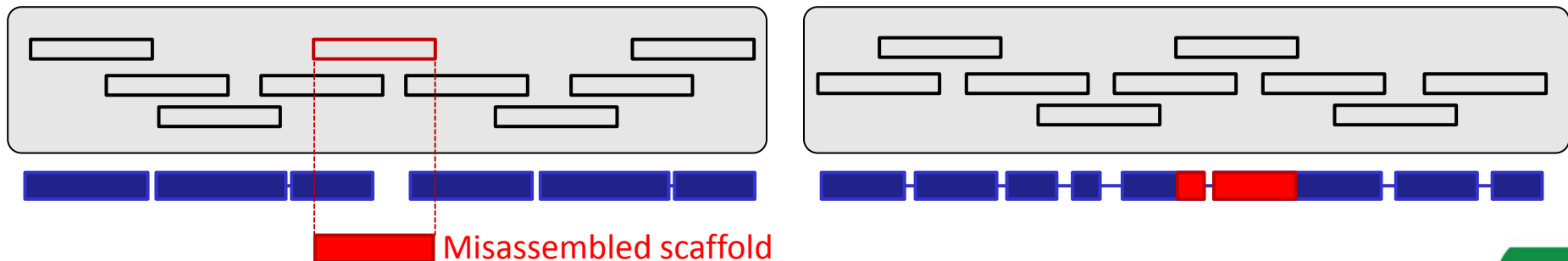
Detect & assemble overlapping scaffolds

❑ 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

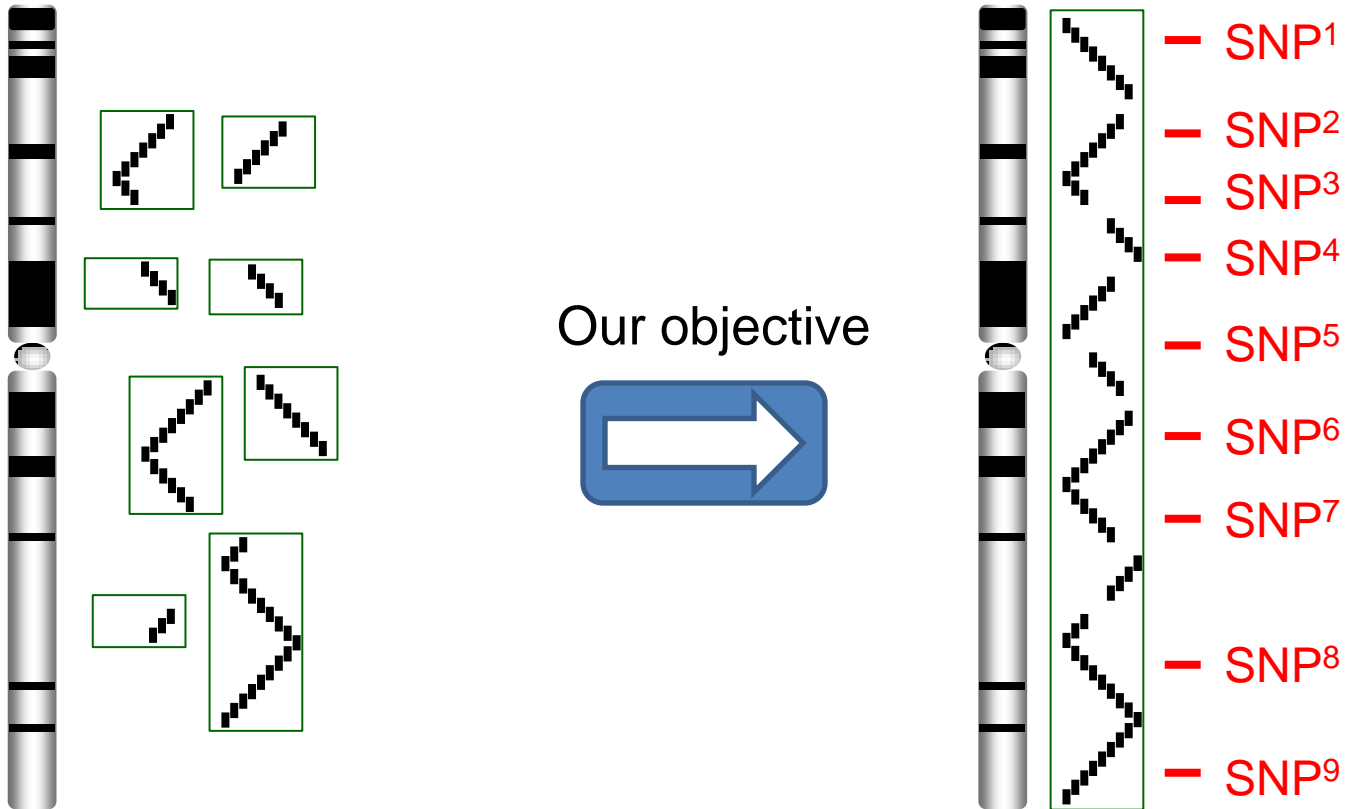
N50

xxx kb

xxx kb (xxx scaff)

xxx kb (xxx scaff)

Order/Orientate scaffolds with markers



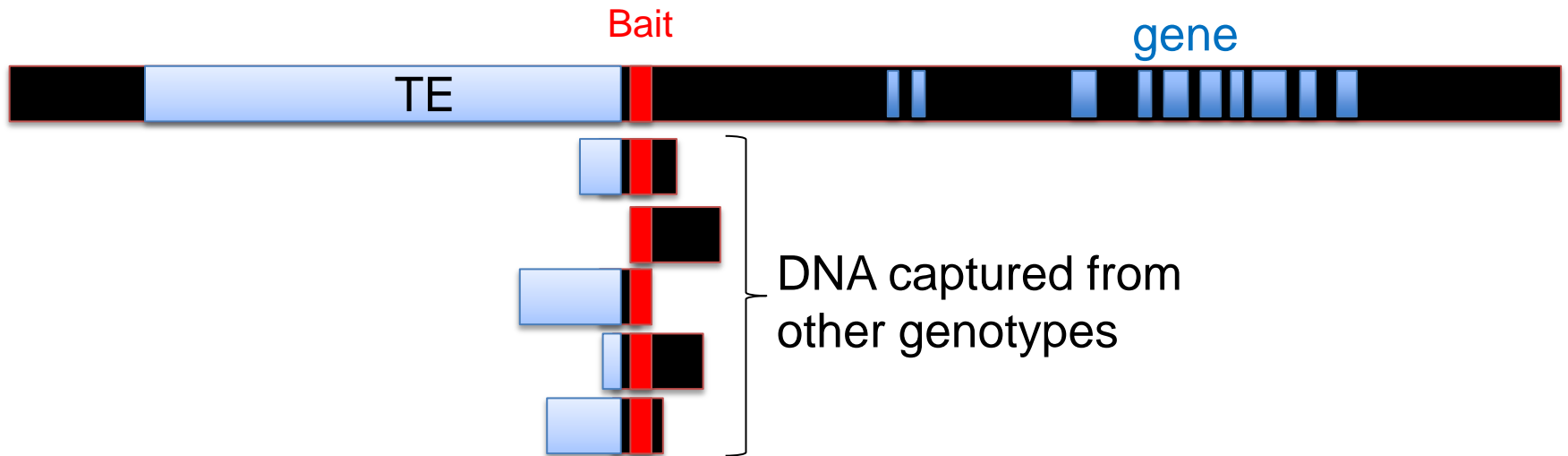
□ Unordered BAC-contigs assembled into **xxxx** scaffolds

□ 1 pseudomolecule

Order/Orientate scaffolds

- SNP discovery with *SureSelect*® Seq. Capture

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



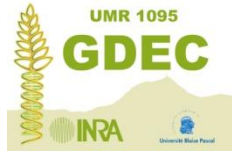
- **xxxxx** SNPs found on scaffolds accounting for **xxx Mb**
- **xxxx** chr-wide SNPs chosen for genotyping

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