Sequencing and Assembling 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, Adriana A. Alberti, et al.

The 3B sequence

❑ Whole 3B Shotgun

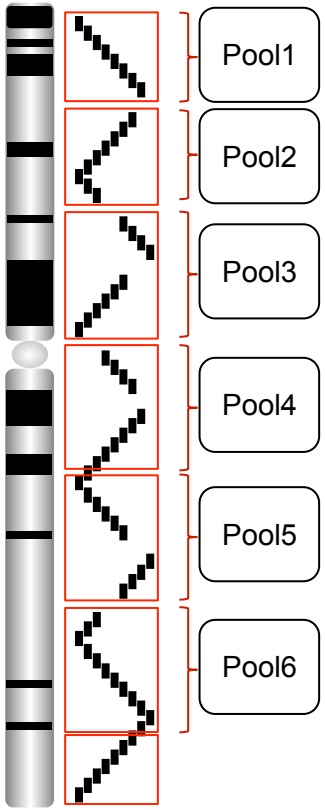
3B DNA

Illumina GAII

PE 2 x 108 nt

82 Gb

❑ MTP BAC Sequencing



8452 BACs

932 pools

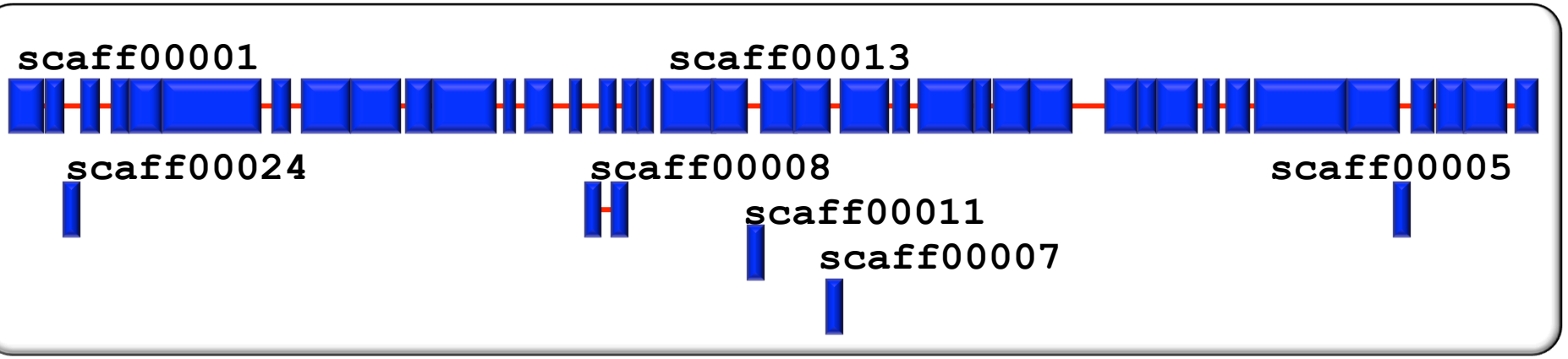
MP 8kb

454 GS-FLX

40 Gb

❑ BAC-ends

❑ BAC pool raw assemblies



v2.1

#scaffolds	16136 scaff
Cumul. size	1 Gb
Gaps	18 %
N50	275 kb

➔ Large scaffolds but small contigs (Newbler fragmentation)

❑ Finishing

v2.1

V. Barbe, S. Mangenot

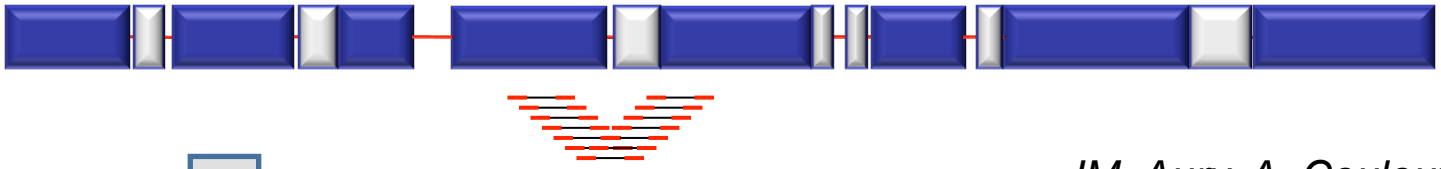
Manual

❑ Manual Curation of the scaffolding

Consider : BAC-ends

Mate pair info (Newbler)

v3.0



JM. Aury, A. Couloux

Automated

❑ Gap closer

❑ Homopolymer correction

v4.0

		#scaff	Mb	N50
v2.1	Raw assemblies	16136	1040	275 kb
v4.0	Finishing + gapCloser	5109	993	463 kb

BAC pool – *Issues*

- Deal with ~8000 *E. coli* clones
- Assignment scaffolds ⇔ BACs

Pooling/tagging

- Incomplete representation of chr.
- Redundancy
- Misassembled BACs
  BAC contamination

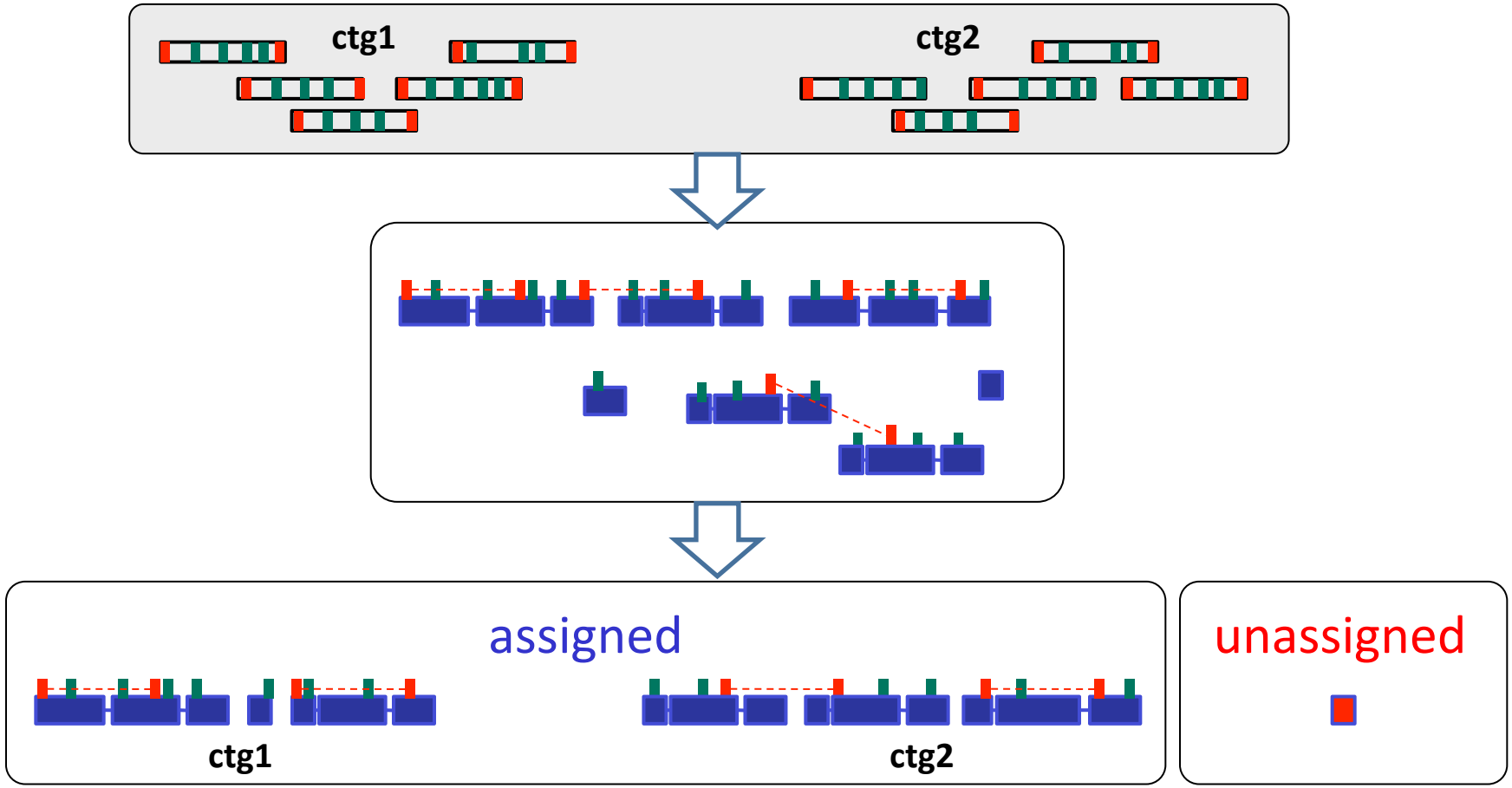
Quality of the physical map

❑ Assigning scaffolds ⇔ BAC-contigs

➢ BAC Ends

➢ Whole Genome Profiling tags

(Philippe *et al*. BMC Genom. 2012)



v4.1

957 Mb

36 Mb

❑ Incomplete representation of the chr.

	Full map	*Sequenced map*
• Fingerprinted BACs	133,000 (19x)	-
• #BAC contigs	1717	1282
• #MTP BACs	9216	8452

❑ Incomplete

- MTP scaff compared to "survey" contigs

✓Match	87%
✓ Absent	13%

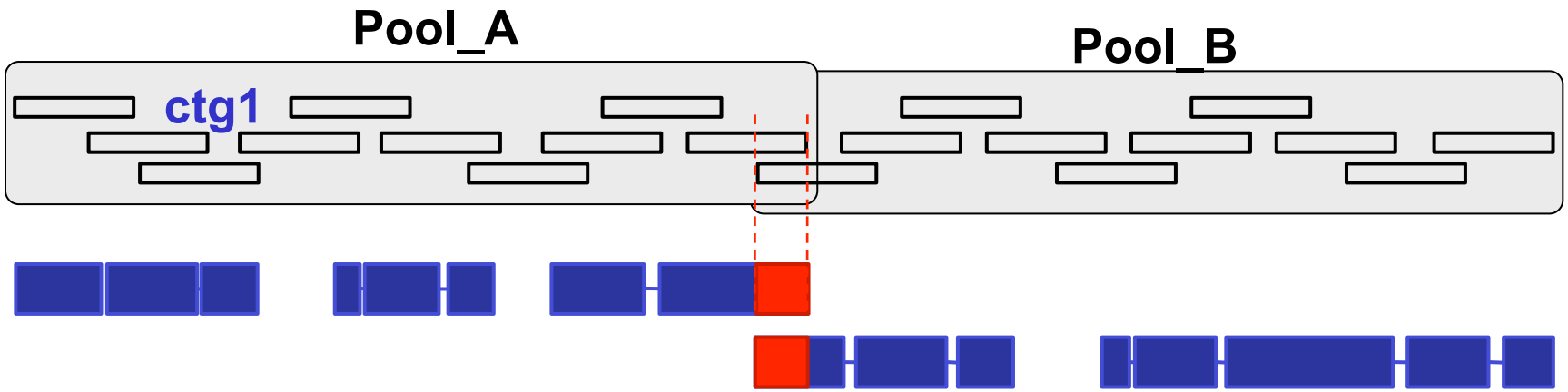
➔ Gaps = 6%
➔ non-3B DNA = 7%

- Inversely (using ~30,000 exons)

✓Full	89%
✓ Partial	7%
✓ Absent	4%

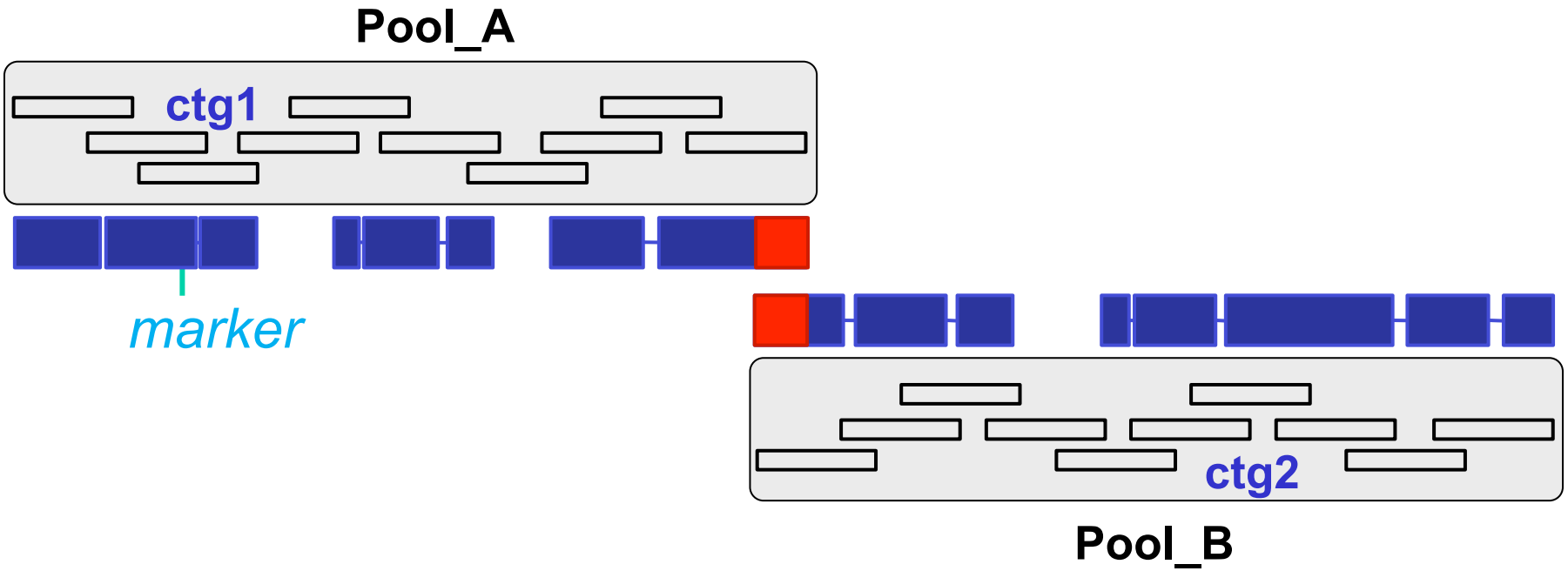
❑ Redundancy

- Expected redundancy



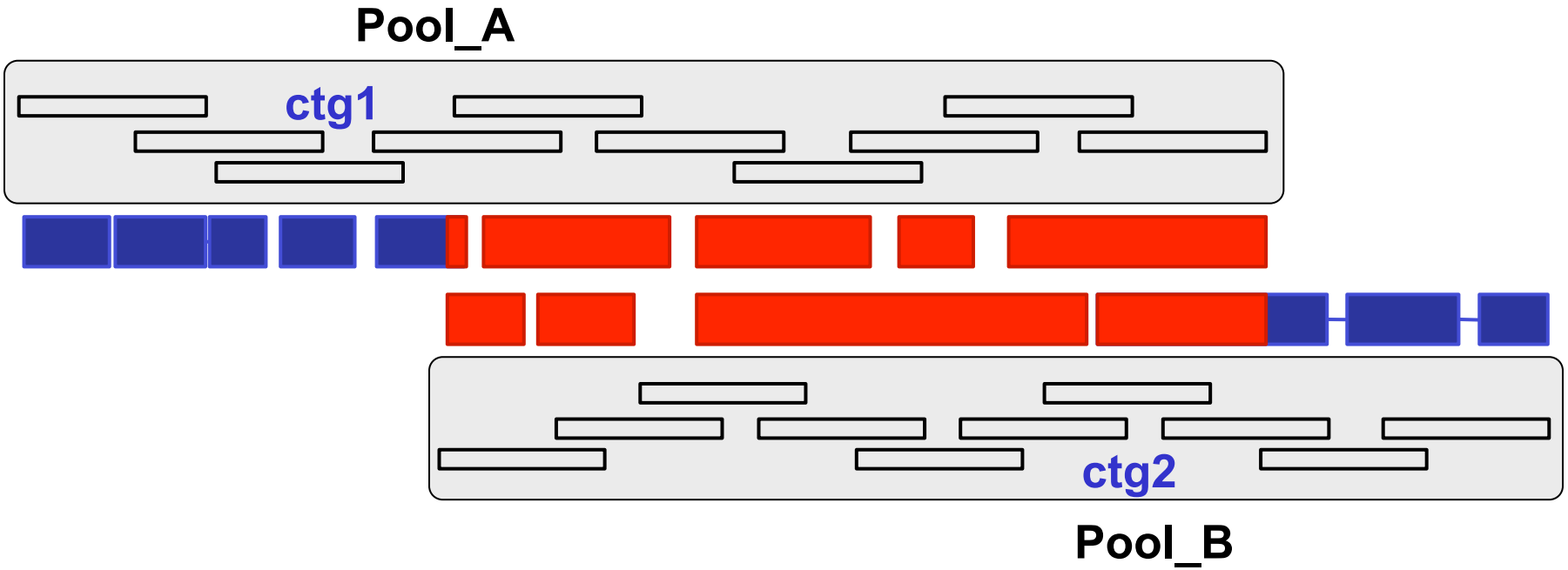
❑ Redundancy

- ○ Unexpected redundancy



❑ Redundancy

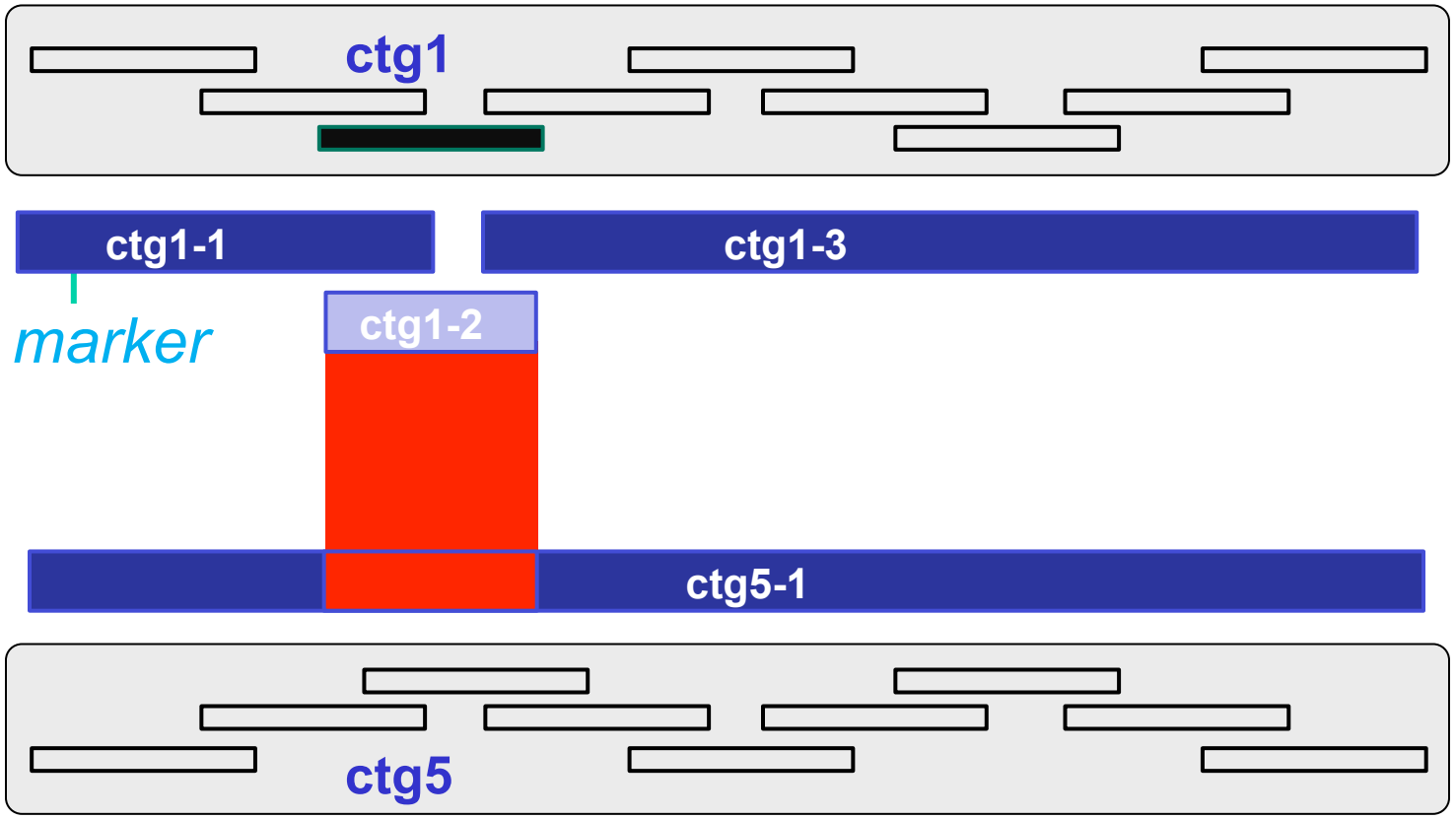
○ Unexpected redundancy



➢ `scaffAssembler.pl`

❑ Redundancy

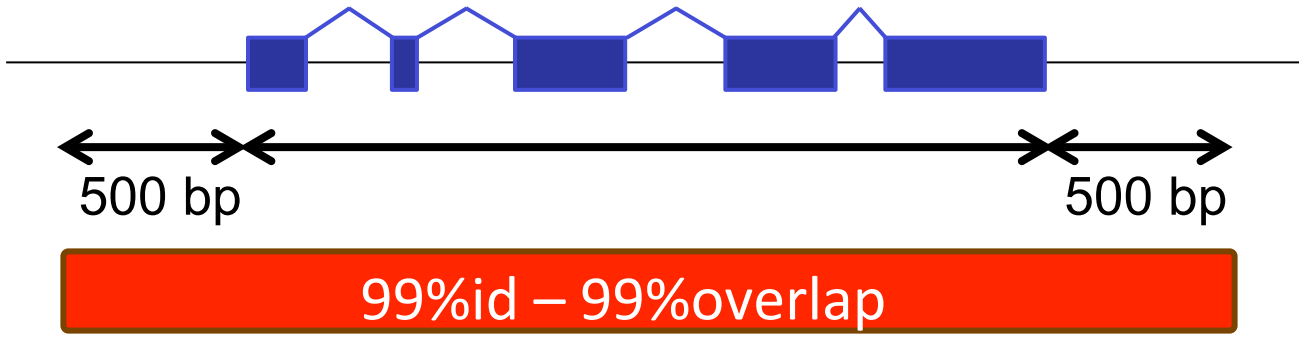
- Misassembled BACs / contaminations



➔ Produce wrong associations betw contigs

❑ Redundancy

- ➢ Estimate redundancy with annotated genes



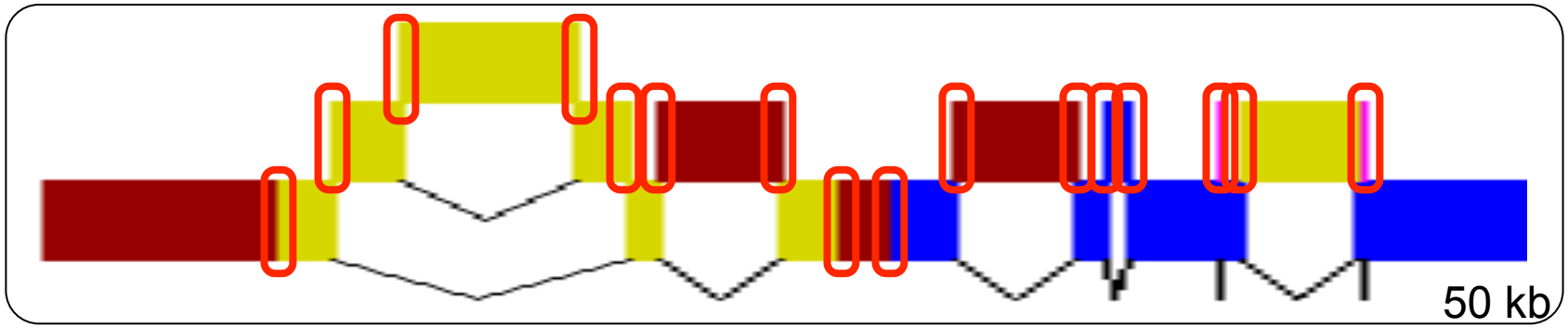
17% redundant gene copies (unexpected)

❑ Redundancy

○ Unexpected redundancy

➢ Problem: "all by all" alignment (1 Gb *vs* 1 Gb)

- ~~Solution1: Mask TEs?~~
- Solution2: Compare TE junctions=ISBPs



➔ Search for shared ISBPs between scaffolds

		#scaff	Mb	N50
v2.1	Raw assemblies	16136	1040	275 kb
v4.0	Finishing + gapCloser	5109	993	463 kb
v4.1.2	Assigning scaff ⇔ ctg	-	-	-
v4.2.2	Merging expected overlap	4747	981	495 kb
v4.4.3	Merging all vs all	2808	833	892 kb

v4.4.3

redundancy **6%**

- Misassemblies
- Duplicated regions

➔ work on a set of **non-redundant genes**

❑ Ordering scaffolds

- How many scaffolds with genes?

  with genes → 675 Mb (81%)

  wo genes → 158 Mb (19%)

- SNP discovery (Agilent-*SureSelect*®)

39,077 SNPs

❑ Ordering scaffolds

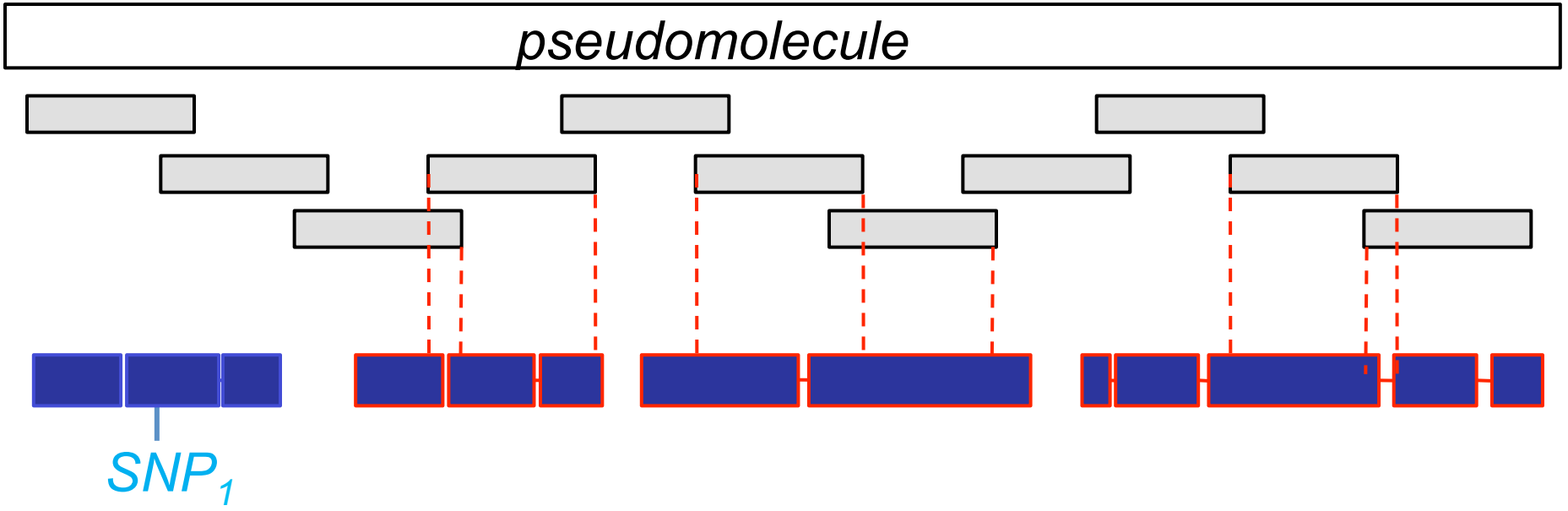
➢ Genetic mapping (*P. Sourdille INRA-GDEC*)

- Cs-Re map (1891 SNPs)
- Neighbor map (3865 markers)
- Use phys. map info to infer scaff position

Pseudomolecule

804 sc 599 Mb 72%

964 sc 679 Mb 82%



pseudomol.pl

❑ Ordering scaffolds

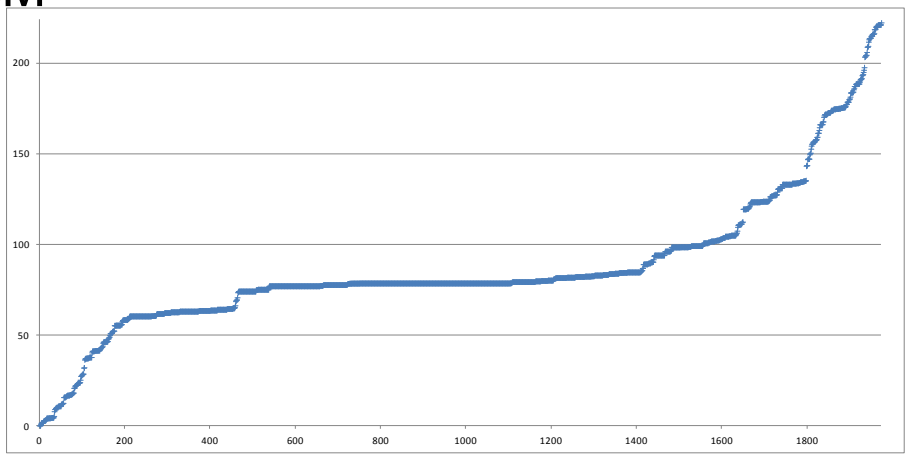
➢ Genetic mapping

	Pseudomolecule		
○ Cs-Re map (1891 SNPs)	804 sc	599 Mb	72%
○ Neighbor map (3865 markers)	964 sc	679 Mb	82%
○ Use phys. map info to infer scaff position	1295 sc	760 Mb	91%

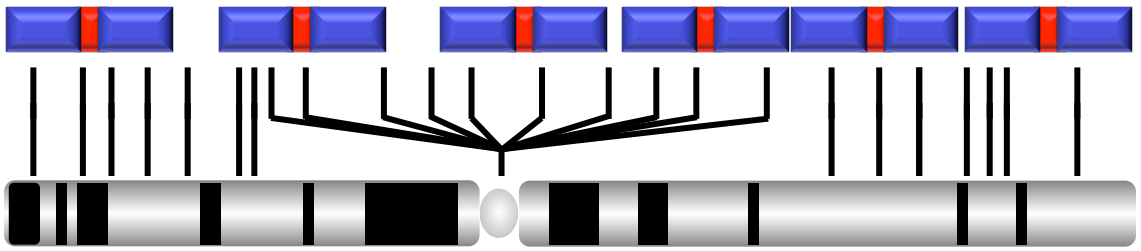
❑ Ordering scaffolds

➢ Genetic mapping

223 cM



0 cM



1295 scaff

366 bins

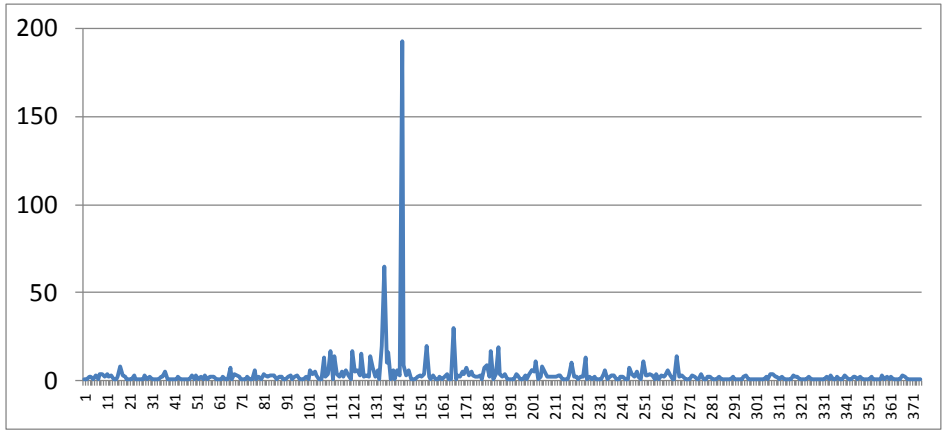
❑ Ordering scaffolds

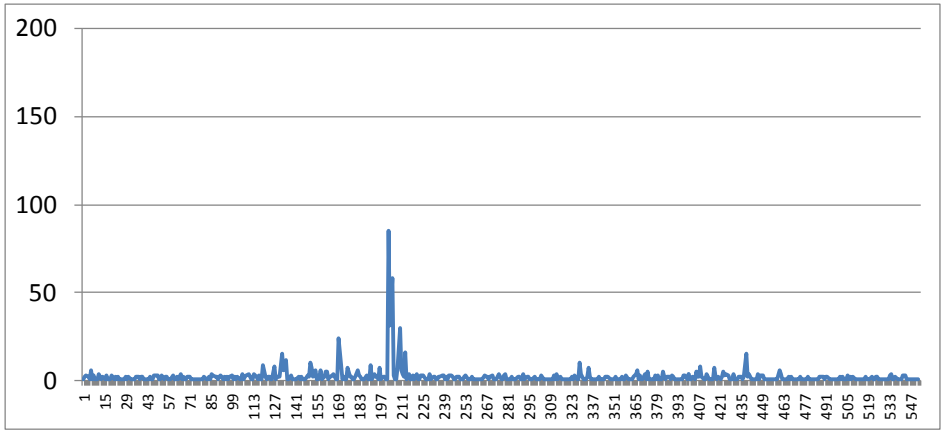
➢ LD mapping (*F. Balfourier - INRA-GDEC*)

1295 scaff
366 genetic bins

1295 scaff
554 genetic+LD bins

#scaff per bin

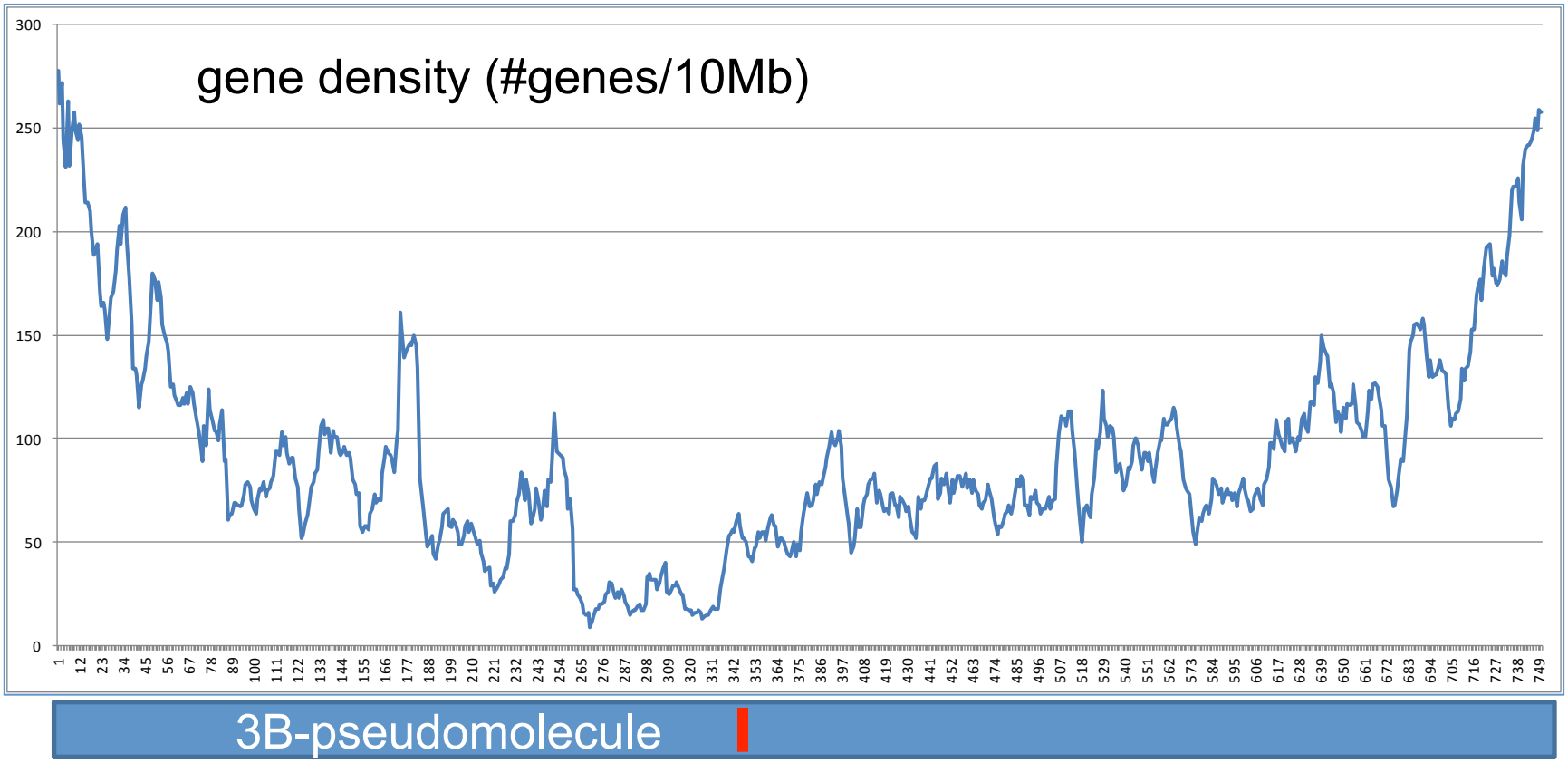




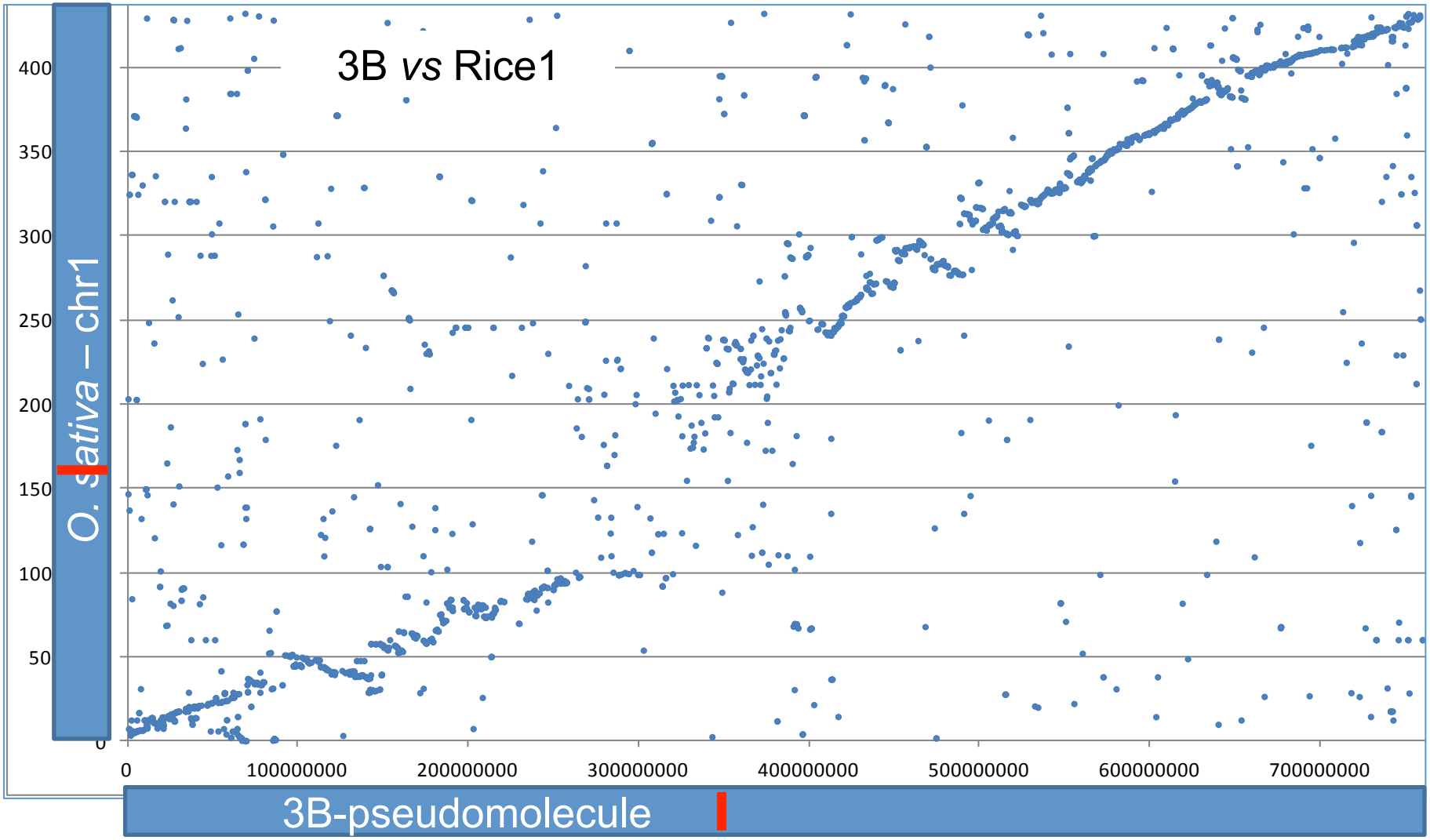
❑ 3B pseudomolecule

➢ **TriAnnot pipeline**

7703 prot-coding genes



❑ 3B pseudomolecule



❑ Ongoing studies

- Gene space
- Comparative genomics (duplicated genes!)
- TE
- Gene expression
- Structural variations
- Recombination studies

+ 15 map-based cloning projects on 3B (collab.)

INRA Clermont-Ferrand

C. Feuillet
E. Paux
P. Sourdille
P. Leroy
N. Guilhot
L. Pingault
J. Daron
N. Glover
S. Theil

Genoscope, Evry

P. Wincker
A. Alberti
V. Barbe
S. Mangenot
JM. Aury
A. Couloux

INRA Versailles

H. Quenesville
M. Alaux
et al.

INRA Toulouse

H. Berges et al.

Inst. Experimental Botany

J. Dolezel et al.