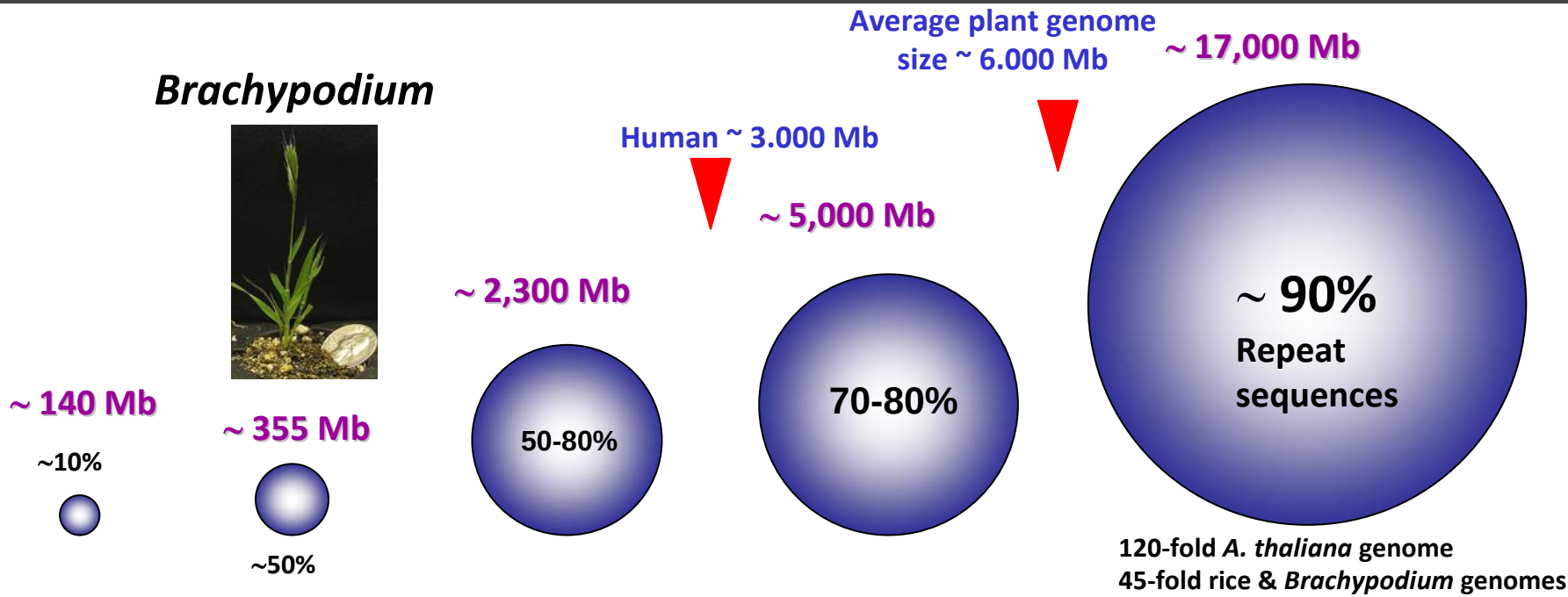
Optimizing the construction of robust physical maps in wheat

By Romain Philippe

January-14-2012

The Challenge……



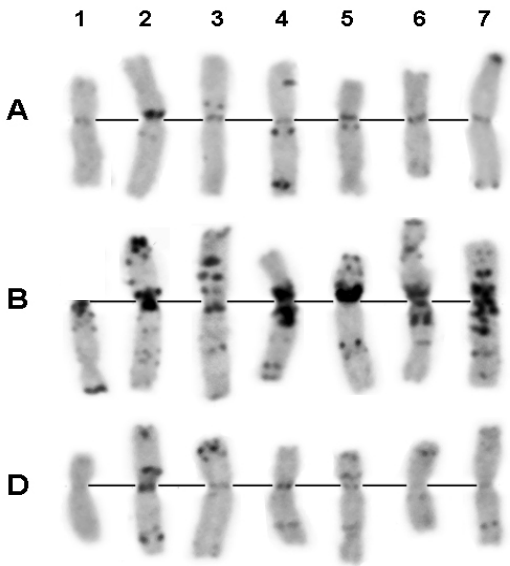
A. thaliana **(2x)** **Rice(2x)** **Maize (2x)** **Barley (2x)** **Bread wheat (6x)**





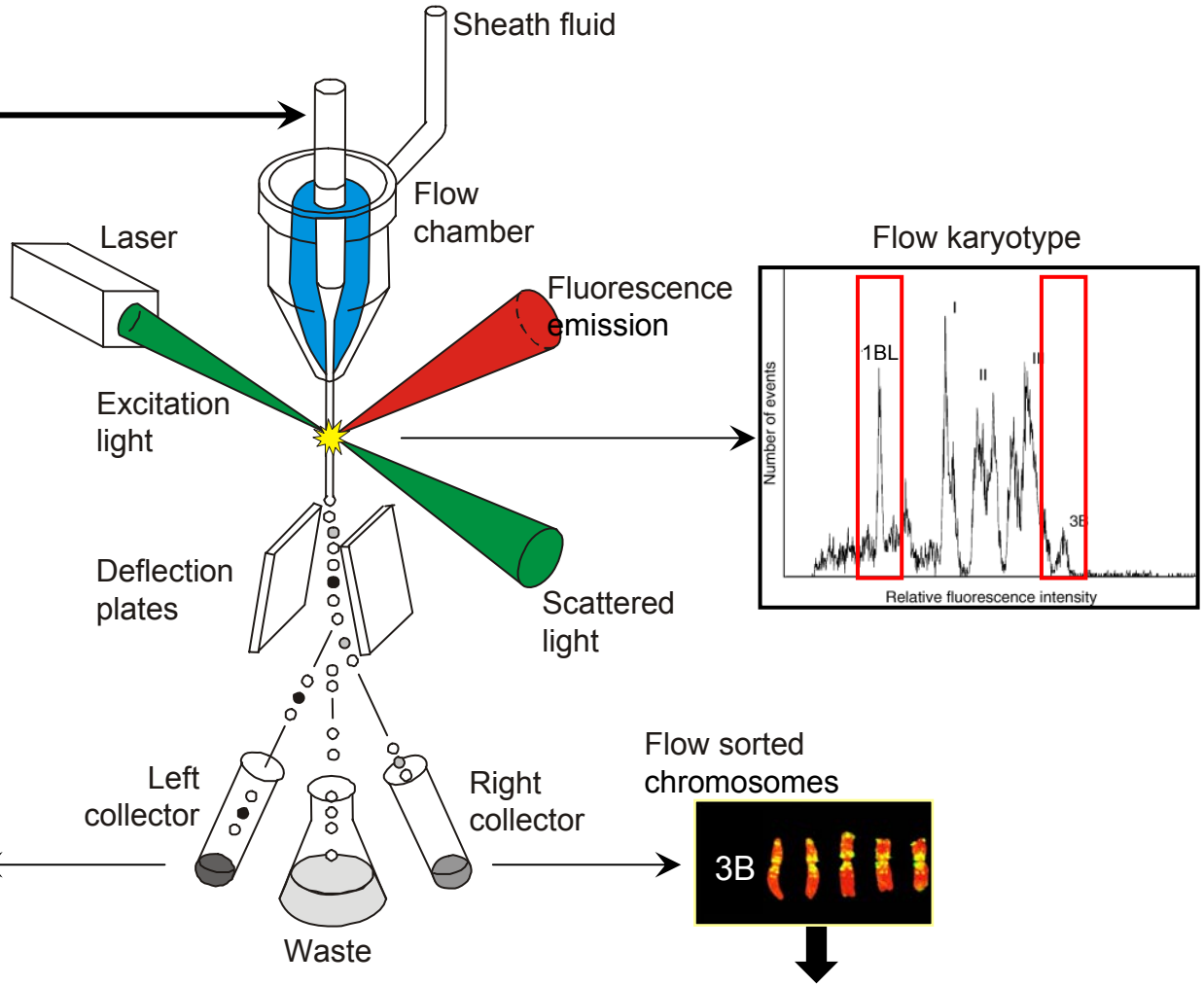
The wheat genome is a challenge for genomic studies and sequencing

Dissecting the hexaploid wheat genome through flow sorting of individual chromosomes



DT1BL line

Laboratory of
Molecular Cytogenetics and Cytometry
Institute of Experimental Botany AS CR



Construction of chromosome or chromosome arm-specific BAC libraries

(kindly provided by J. Dolezel)

A Physical Map of the 1-Gigabase Bread Wheat Chromosome 3B

Etienne Paux,[1] Pierre Sourdille,[1] Jérôme Salse,[1] Cyrille Saintenac,[1] Frédéric Choulet,[1] Philippe Leroy,[1] Abraham Korol,[2] Monika Michalak,[3] Shahryar Kianian,[3] Wolfgang Spielmeyer,[4] Evans Lagudah,[4] Daryl Somers,[5] Andrzej Kilian,[6] Michael Alaux,[7] Sonia Vautrin,[8] Hélène Bergès,[8] Kellye Eversole,[9] Rudi Appels,[10] Jan Safar,[11] Hana Simkova,[11] Jaroslav Dolezel,[11] Michel Bernard,[1] Catherine Feuillet[1]

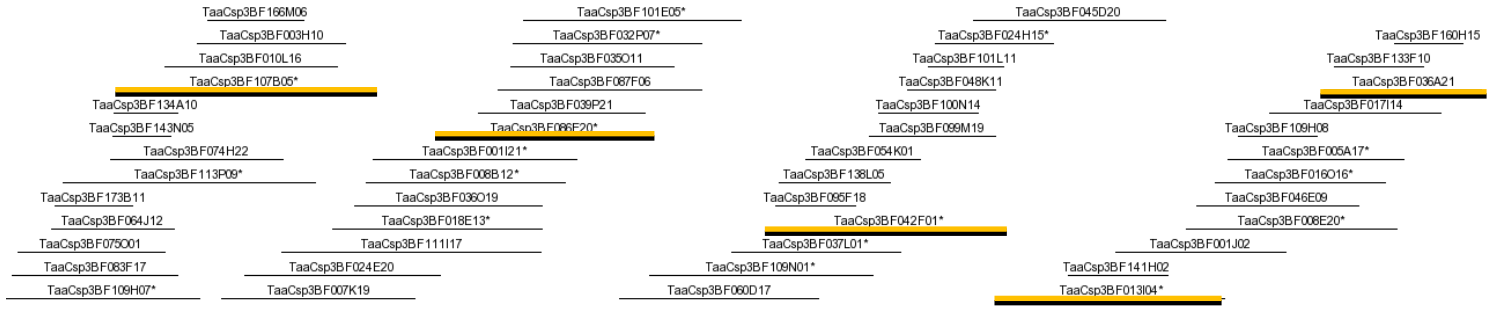
SCIENCE VOL 322 3 OCTOBER 2008

- ✓ 131 792 BACs (19.2 X coverage) fingerprinted with SNaPshot technology
- ✓ 1 283 contigs (average size = 749 kb) built with FPC software
- ✓ 961 Mb coverage (97% chromosome)
- ✓4367 molecular markers (SSRs, ISBPs, ESTs, DArTs,…)
- ✓MTP (8448 clones)

http://urgi.versailles.inra.fr/projects/Triticum/index.php

Improving physical mapping in wheat

✓ Pilot sequencing showed that:



- 10% of the BACs in contigs are mis-assembled

↳ More robust physical map

KeyGene

AMPLICON express

Sequence-based physical mapping of complex genomes by whole genome profiling

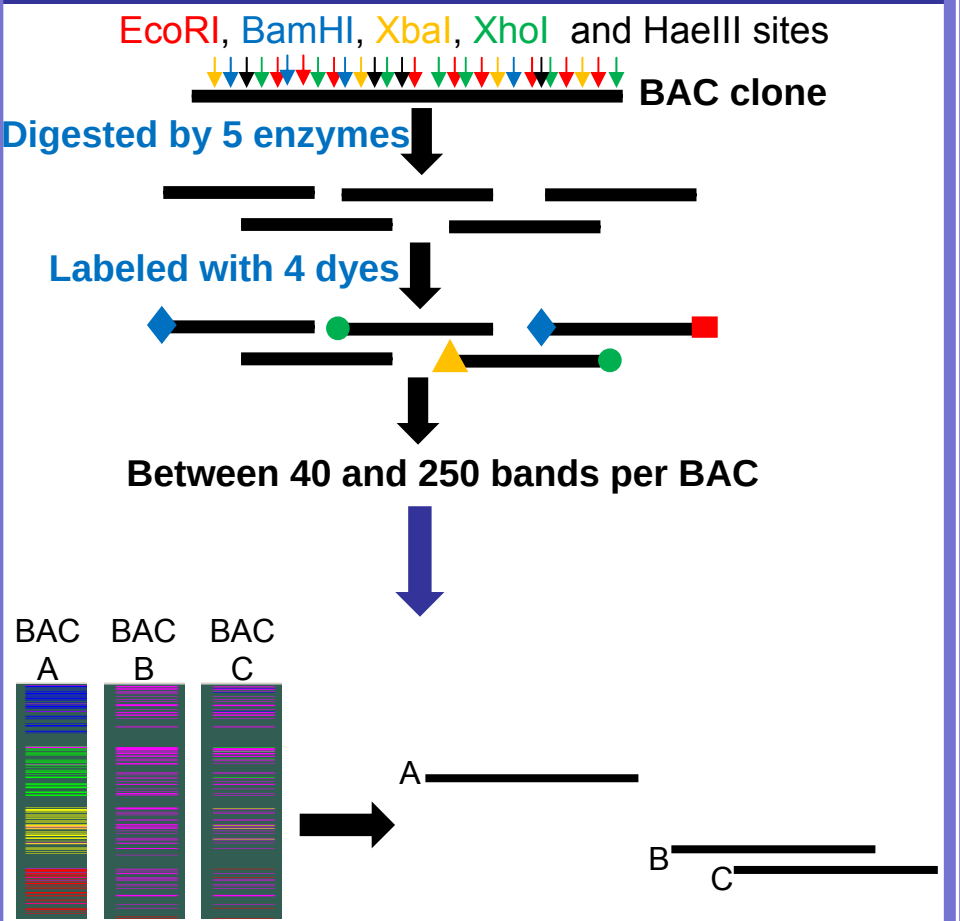
Jan van Oeveren,[1] Marjo de Ruiter,[1] Taco Jesse,[1] Hein van der Poel,[1] Jifeng Tang,[1] Feyruz Yalcin,[1] Antoine Janssen,[1] Hanne Volpin,[1] Keith E. Stormo,[2] Robert Bogden,[2] Michiel J.T. van Eijk,[1] and Marcel Prins[1,3]

[1]*Keygene N.V., Wageningen, The Netherlands;* [2]*Amplicon Express Inc., Pullman, Washington 99163, USA*

CSH PRESS GENOME RESEARCH

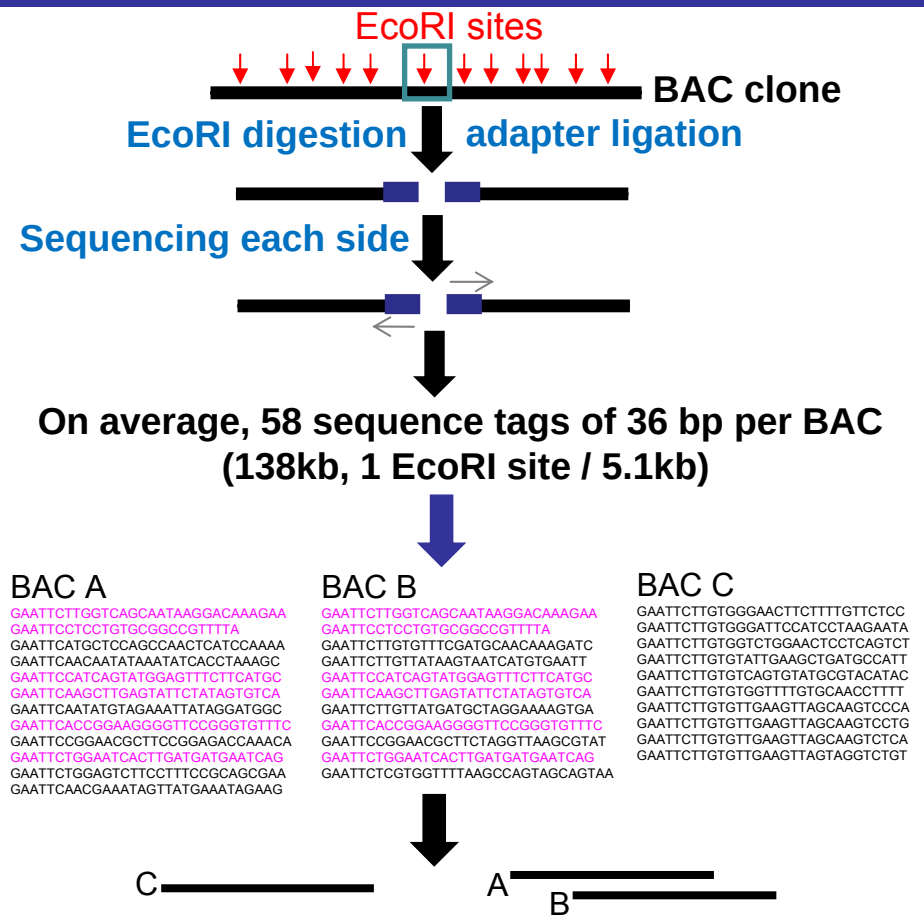
Whole Genome Profiling (WGP) : a new sequence-based physical mapping technology

SNaPshot



Tolerance of 0.4 bp between two bands to be considered as identical

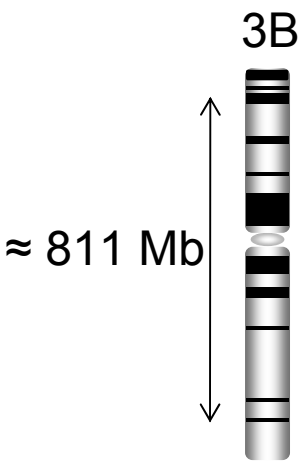
WGP



Tolerance of 0 mismatch (two identical sequences)

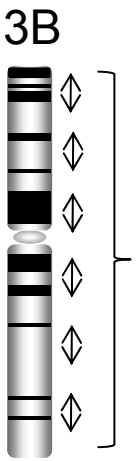
WGP is theoretically more robust than SNaPshot

WGP pilot project in wheat



Total 3B BAC fingerprints
= 56,952 FP (9.6X)

Not random selection of 28.3%

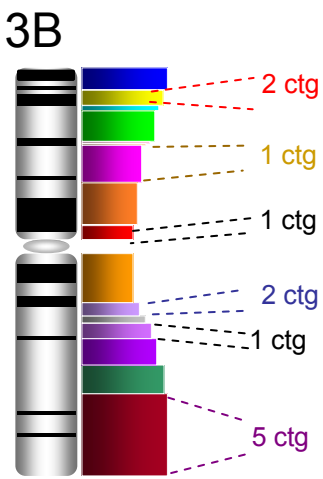


Subset of 3B BAC fingerprints
= 16,128 FP (9.6X)
≈ 230 Mb

AMPLICON express

Tag sequencing

KeyGene

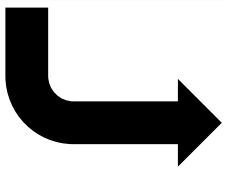


12 sequenced contigs (18Mb)
(Choulet et al. Plant Cell 2010)

Comparison between SNaPshot and WGP efficiency

✓ Physical map assembly done with FPC software

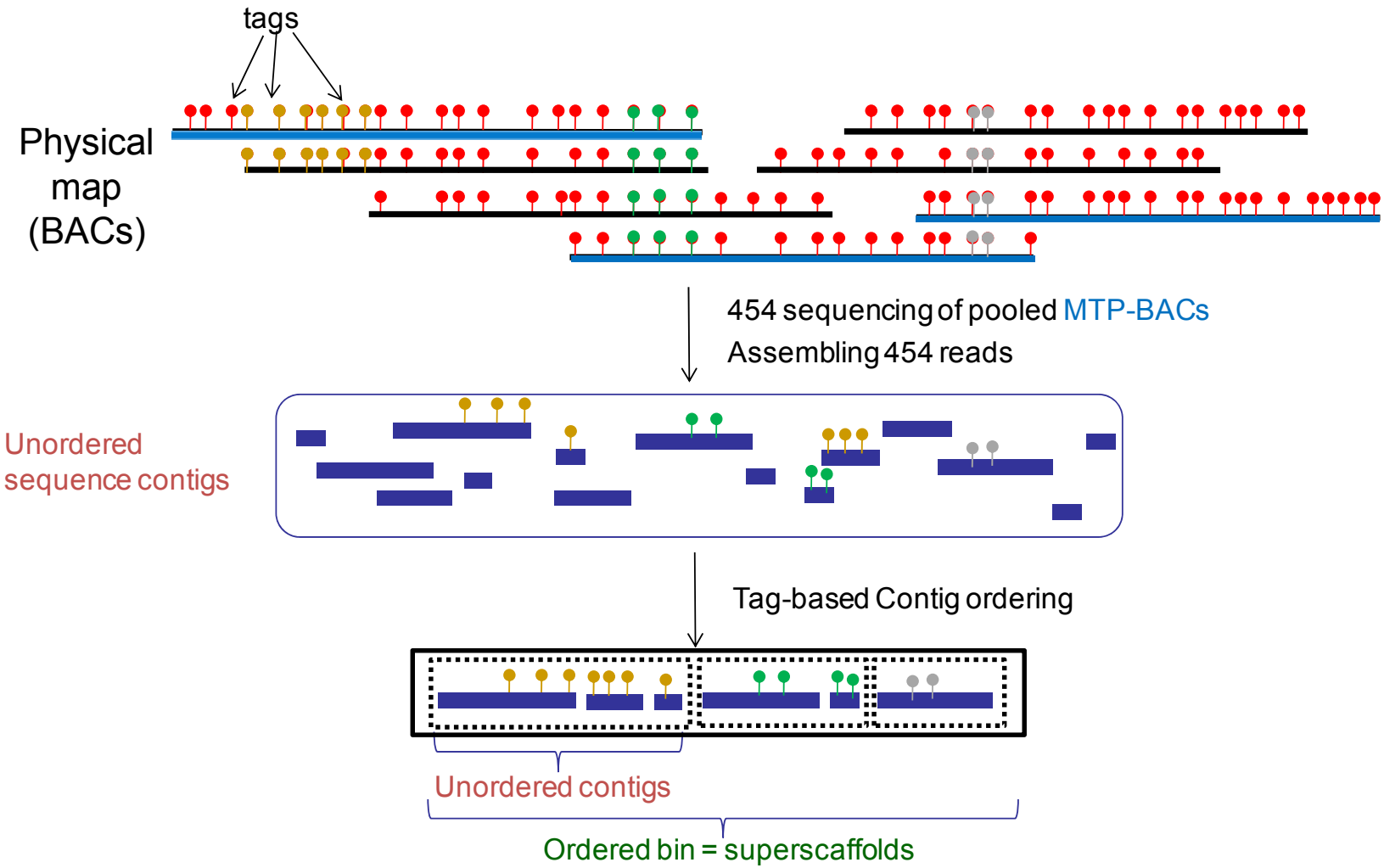
	SNaPshot (1^{e-25})	WGP (1^{e-11})
Estimated coverage in length	236 Mb ± 65	199 Mb ± 42
Number of contigs	631	434
Number of singletons	2112 (18.8%)	4145 (36.9%)
Average contig size (Kb)	374	469
N50 (Kb)	455	567
L50	164	115
Comparison to the 12 sequenced contigs:		
Coverage percentage in length	95.8%	94.9%
Number of chimerical contigs for 10 Mb	0.6	0.6
Percentage of mis-assembled BACs	9.5%	2.7%



✓ Equivalent coverage in length
✓ Equivalent number of chimerical contigs
✓ Less mis-assembled BACs in WGP

WGP performs better than SNaPshot for physical mapping in wheat

Can WGP help to generate a high quality reference sequence at reduced sequencing depth?



WGP to support sequence assembly

454 re-sequencing of 4 reference (Sanger) sequences (600 Kb – 1Mb)

↓

Series of assemblies using Newbler v2.3 (15X-50X, 5X steps)
with and without paired end (PE)

↓

Scaffolding using WGP tags

↓

Comparison with reference sequences

↳ **No PE** + WGP tag integration:

- **enables scaffolding**
- gaps (34%)
- error in bin order (20%)

} **Low quality assembly**

↳ **with PE**, WGP tag integration:

- at low sequencing coverage, **creates Superscaffolds** (≤ 20X) but high percentage of gap remains (>24%) => **Low quality assembly**
- does not improve scaffolding at high sequencing coverage (≥ 25X)

Improving physical mapping in wheat

✓ 3B markers analysis and sequencing showed the **presence of chimerical contigs** in the 3B physical maps (SNaPshot and WGP).

↳ More robust physical map

✓ Average contig size is smaller in wheat than in small genomes (1500 Kb in Brachypodium vs 749 Kb in wheat chromosome 3B)

↳ Increase the average contig size

Frenkel *et al. BMC Bioinformatics* 2010, **11**:584
http://www.biomedcentral.com/1471-2105/11/584

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

LTC: a novel algorithm to improve the efficiency of contig assembly for physical mapping in complex genomes

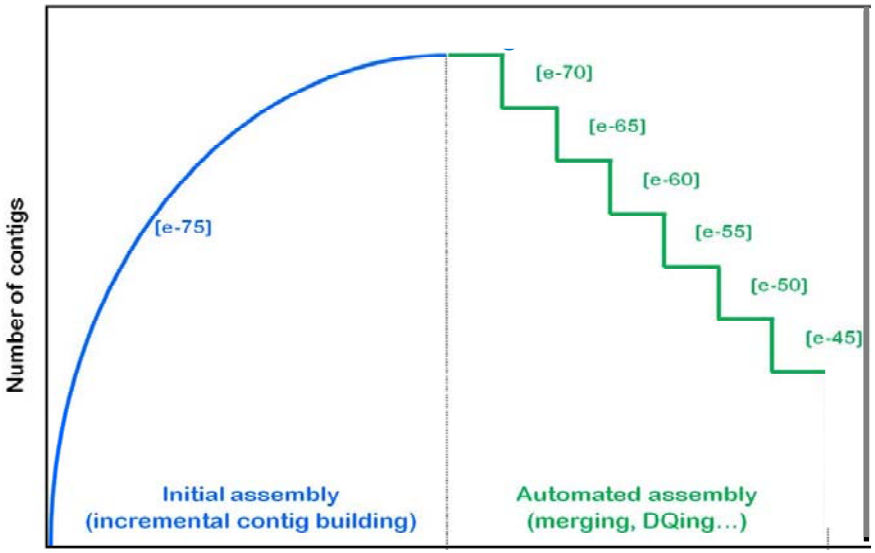
Zeev Frenkel[1*], Etienne Paux[2], David Mester[1], Catherine Feuillet[2], Abraham Korol[1]

University of Haifa

Triticeae Genome

FPC vs LTC

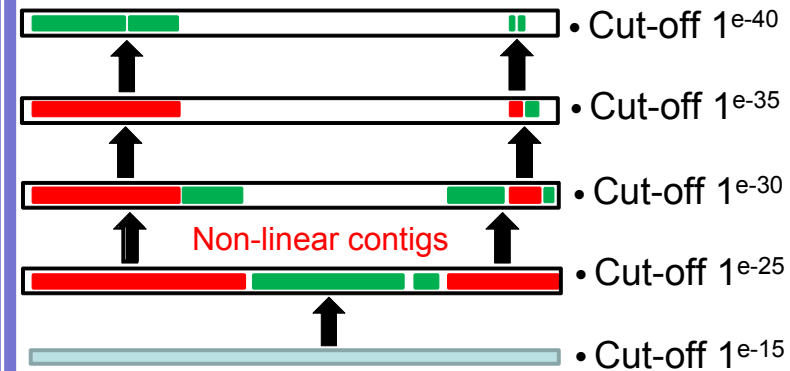
FPC



✓ **Based only on BACs similarity**

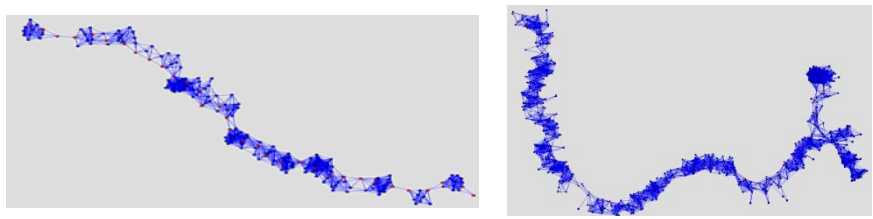
• High initial stringency to limit chimerical contigs
• Adding singletons to contigs extremity at each step
• Merging contigs at each step

LTC



✓ **Based on BACs similarity and contigs linearity**

• Low initial stringency to limit the number of contigs
• Checking contigs linearity

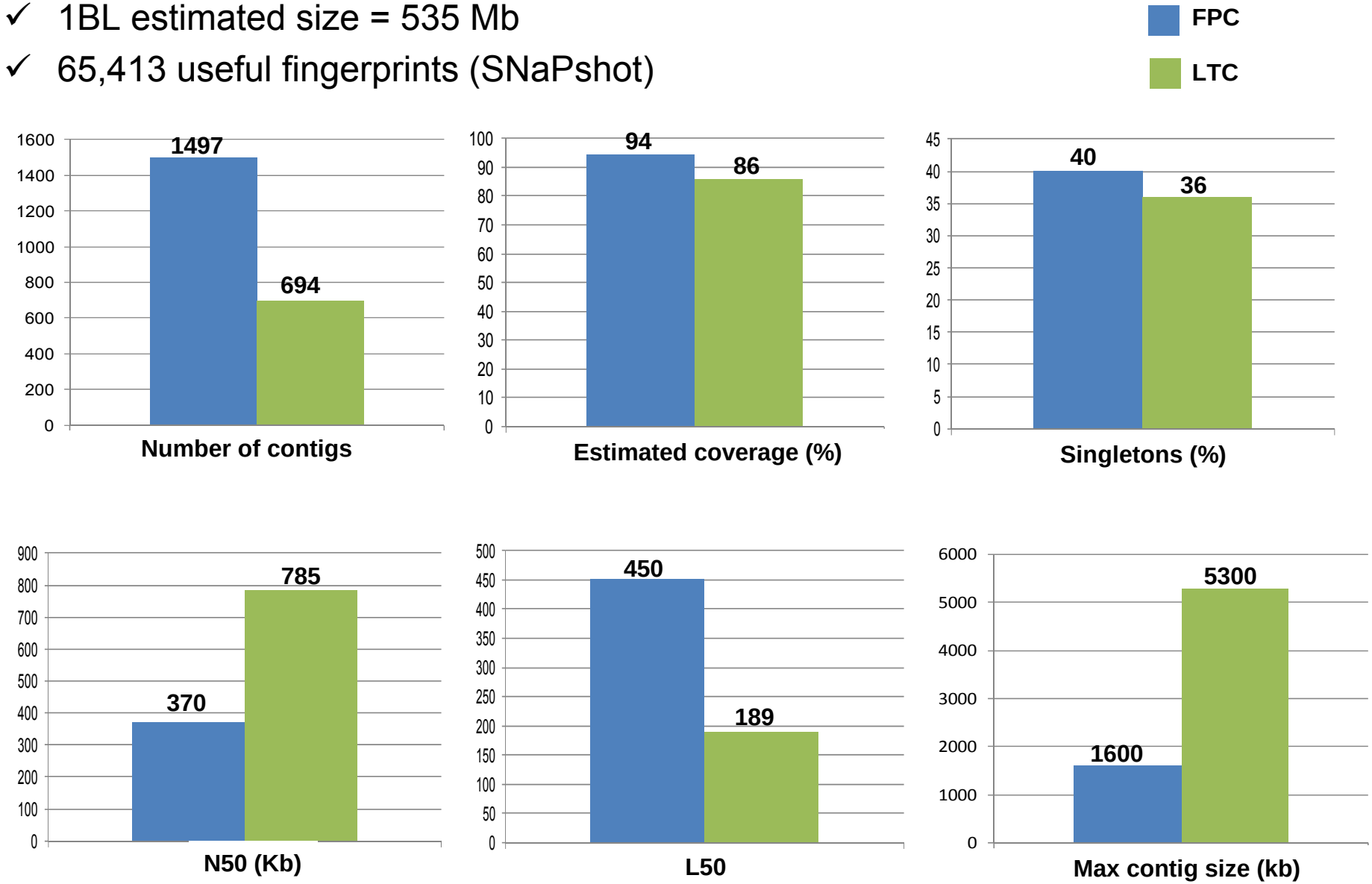


• Splitting non linear contigs by increasing the stringency

↳ **Elimination of chimerical contigs**

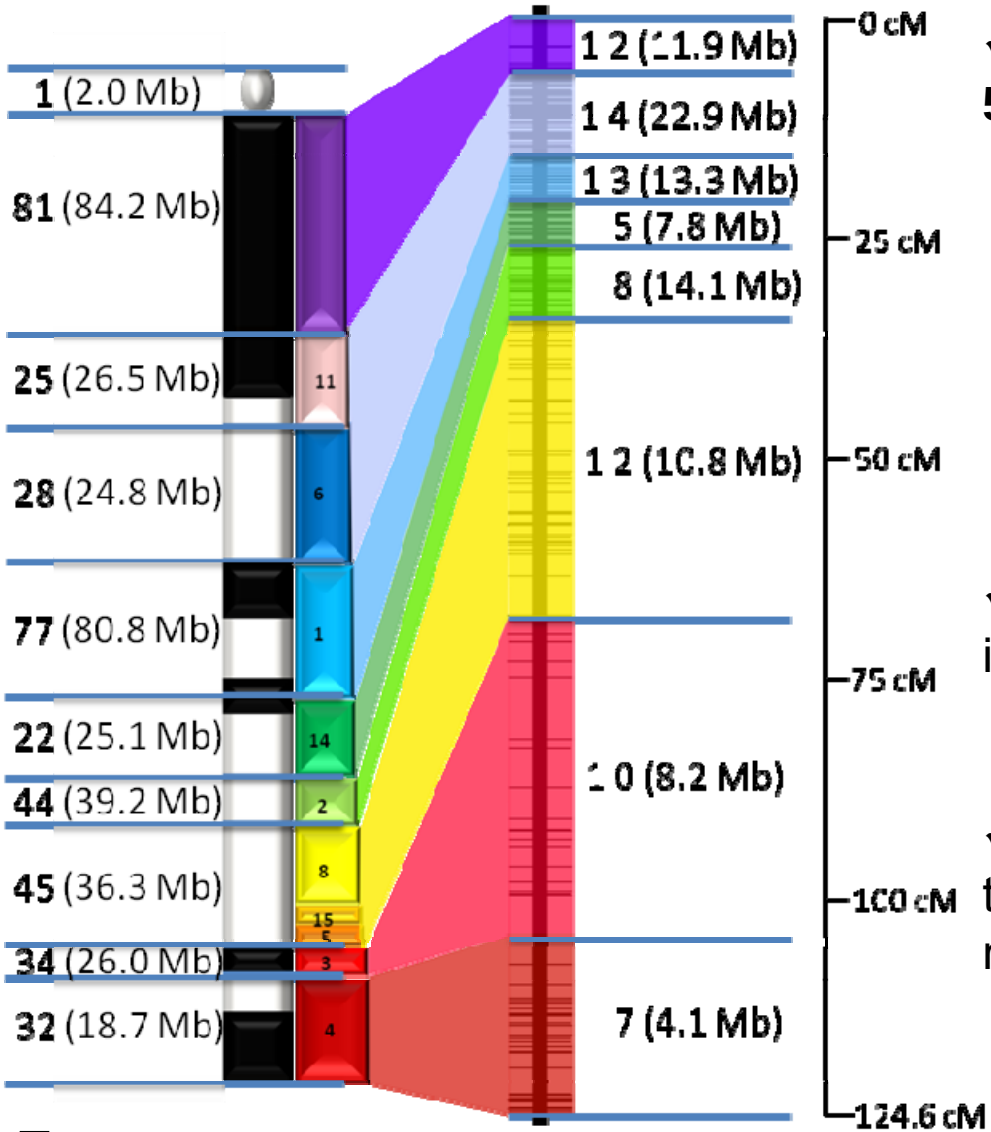
Comparison LTC and FPC 1BL physical map

- ✓ 1BL estimated size = 535 Mb
- ✓ 65,413 useful fingerprints (SNaPshot)



✓ **LTC significantly improves physical mapping in wheat**

Status of 1BL anchoring



✓ **616 contigs** (455 Mb, 85%) containing **5538 markers**:

- **403 PCR markers** (ISBP, SSR, COS, RFLP)
- **1223 unigenes** (Nimblegen chip with 39,179 wheat unigenes)
- **3912 ISBP** (Nimblegen chip with 17,788 1BL ISBP)

✓ **389 contigs** (364 Mb, 80%) anchored in **deletion bins**

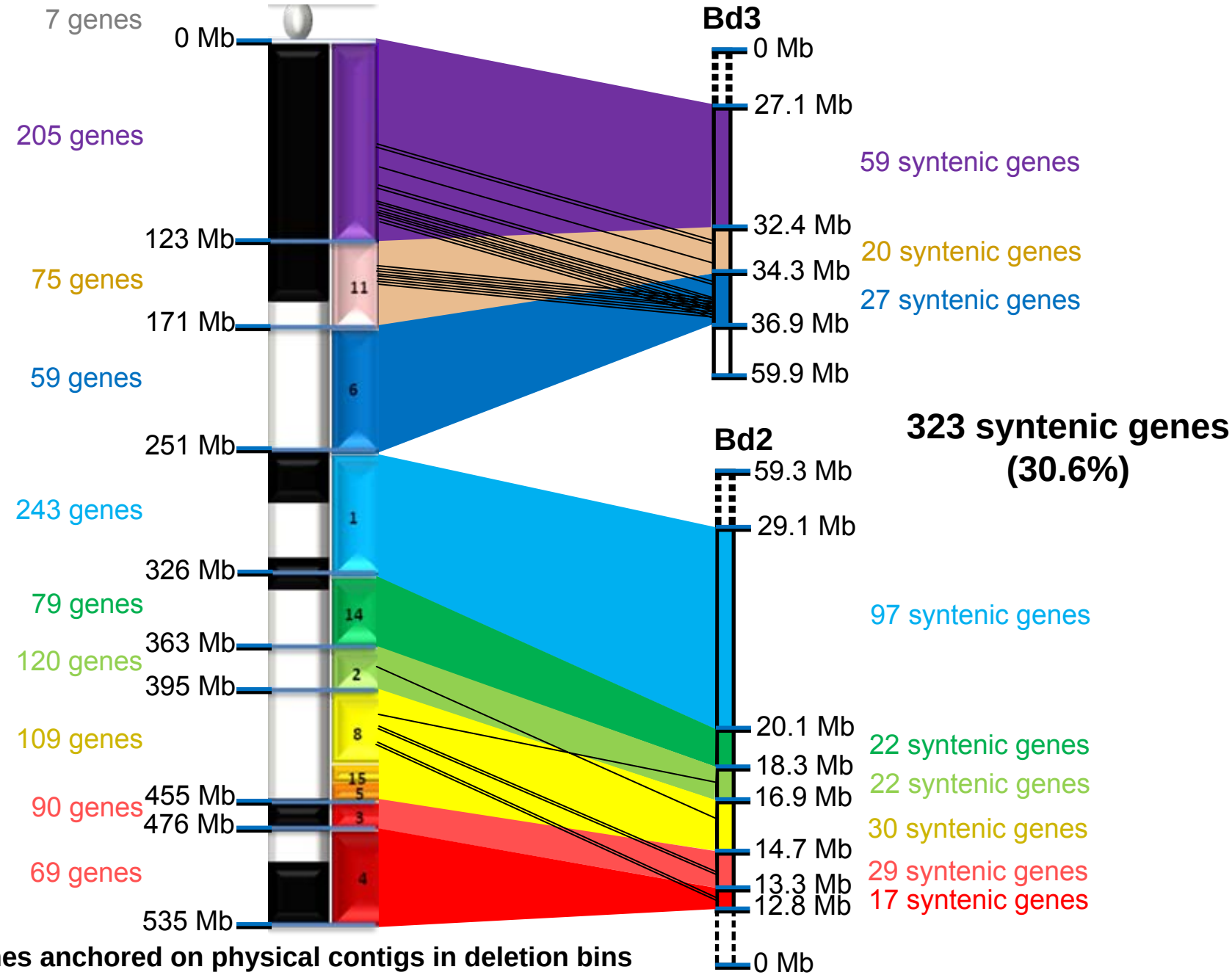
✓ **82 contigs** (94 Mb, 21%) anchored on the **1BL neighbour genetic map** (478 markers)

Sequencing of chromosome 1B (S/L)

Breed Wheat

INVESTISSEMENTS D'AVENIR

Preliminary results on the synteny with Brachypodium

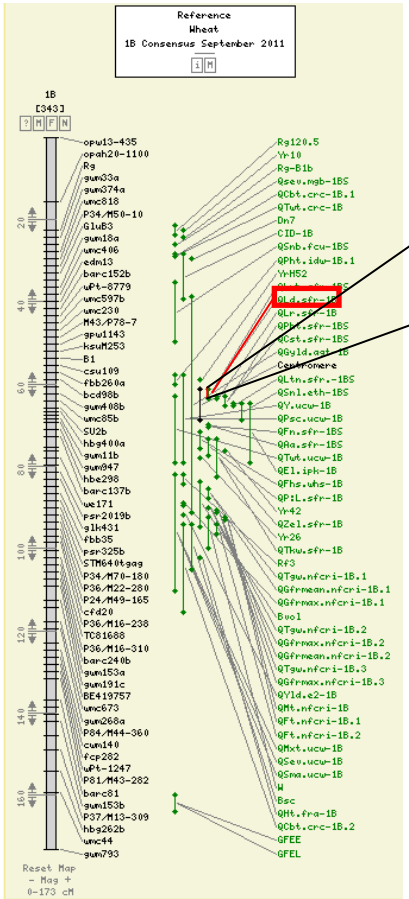


1056 genes anchored on physical contigs in deletion bins

Map-based cloning on 1BL

- 40 QTLs on 1BL available on wheat GeneCatalog (http://ccg.murdoch.edu.au/index.php/CMap)
- 17 QTLs with marker(s) on the 1BL physical map

Example: QLd.sfr-1B (Lodging)



1B wheat consensus map (sept. 2011)

1 ctg (2 Mb, 18 markers)

81 ctg (84.2 Mb, 801 markers)

1BL physical map
26 contigs ordered based on synteny → 47 Mb, 134 genes out of 474 markers

Ctg438, 1167 Kb, 8 genes out of 23 markers

✚ 56 unordered contigs, 40 Mb, 78 genes out of 345 markers

Take-home messages

- ✓ **LTC improves physical mapping** in complex genomes compared to FPC:
  - Increased contig size
  - No chimerical contigs

- ✓ In wheat, **whole genome profiling** is **more efficient** than SNaPshot for **physical mapping:**
  - Increased contig size
  - Decreased percentage of mis-assembled BACs

- ✓ In wheat, **WGP tags improve low** quality sequence assemblies

→ **Combination of WGP fingerprinting with LTC assembly should lead to very high quality physical maps in wheat**

✓ **1BL physical map available** (616 contigs covering 85% of the 1BL chromosome arm and containing 1223 unigenes out of 5538 markers)

http://urgi.versailles.inra.fr/cgi-bin/gbrowse/wheat_FPC_1BL_pub/

Funded by

GAP Division

Acknowledgments

Isabelle Bertin
Etienne Paux
Pierre Sourdille
Frédéric Choulet
Nicolas Guilhot
Catherine Feuillet

Jaroslav Dolezel
Hana Simkova
Jan Bartos

Jan van Oeveren
Jifeng Tang
Alexander Wittenberg
Antoine Janssen
Michiel van Eijk
Edwin van der Vossen

Eduard Akhunov

Abraham Korol
Vova Frenkel

Hélène Bergès
Arnaud Bellec
Sonia Vautrin

Robert Bogden
Keith Stormo

Adriana Alberti
Patrick Wincker

Federica Cattonaro
Simone Scalabrin

WGP data

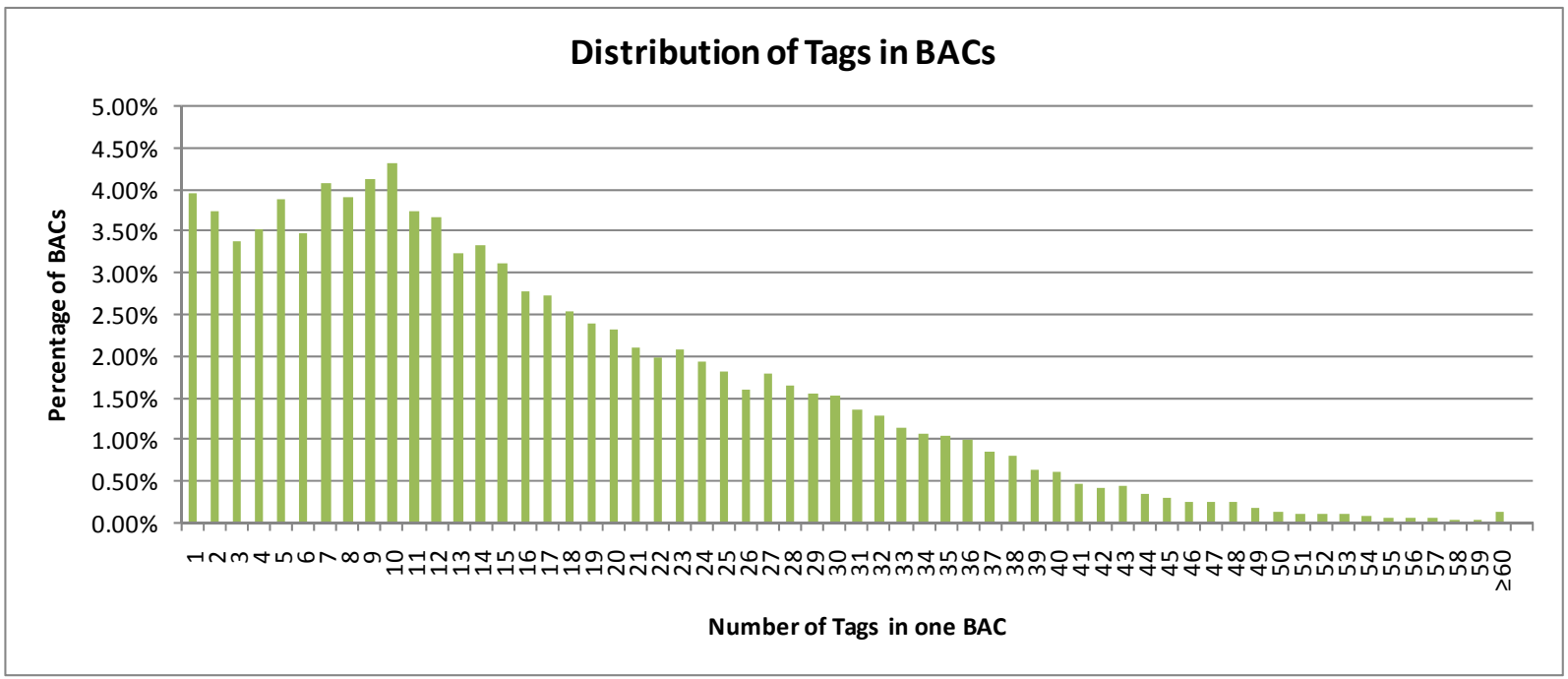
- 111,678 tags assigned to 14,199 BACs.

- Keygen have done a tag filtering : elimination of tags present in only one BAC (uninformative tags) and tags present in more than 12 BACs (no locus specific tags) :

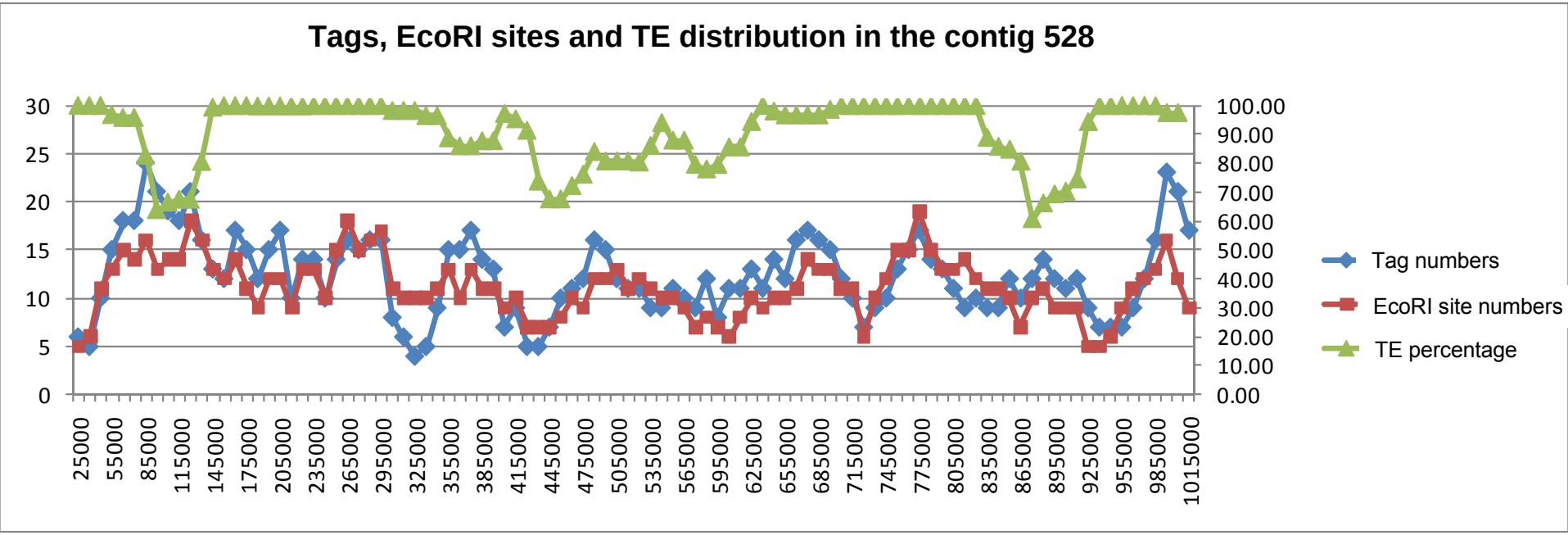
  ✓47,900 tags on 13,888 BACs (8.3X coverage)

  ✓Average number of BACs sharing the same tag: 4.8 (vs 8.3 expected)

  ✓Average number of tags per BAC: 16.4 (vs 58.4 expected : 1 EcoRI site / 5.1 kb).



Repeats do not impact tag distribution



• Average correlation coefficient between tag number and TE percentage: -0.078 (r^2 = 0.01, p-value = 0.00003)

WGP physical map

• Assembly with FPC v9.3 (Soderlund et al., 1997 and 2000) modified by Keygene to handle WGP data.

• FPC parameters established by van Oeveren et al. (2011) to built physical map in Arabidopsis, melon, tomato, allotetraploid rape seed and lettuce :

- ✓ tolerance of 0
- ✓ single cutoff at 1e-06
- ✓ single DQing step at 1e-06

	WGPv1
Number of BACs used	13,888
Percentage of assembled BACs	79.2%
Number of contigs	786
Comparison with the 12 sequenced contigs	
Number of chimerical contigs for 10 Mb	5.6 (35.7%)
Percentage of mis-assembled BACs	26.8%

↳ FPC parameters and/or data filtering were not enough stringent to built a robust physical map in wheat.

WGP physical map

• Elimination of BACs with ≤ 4 tags (bad quality BAC fingerprints) and with ≥ 40 tags (possible mixed BACs) : final subset of 11,238 BACs.

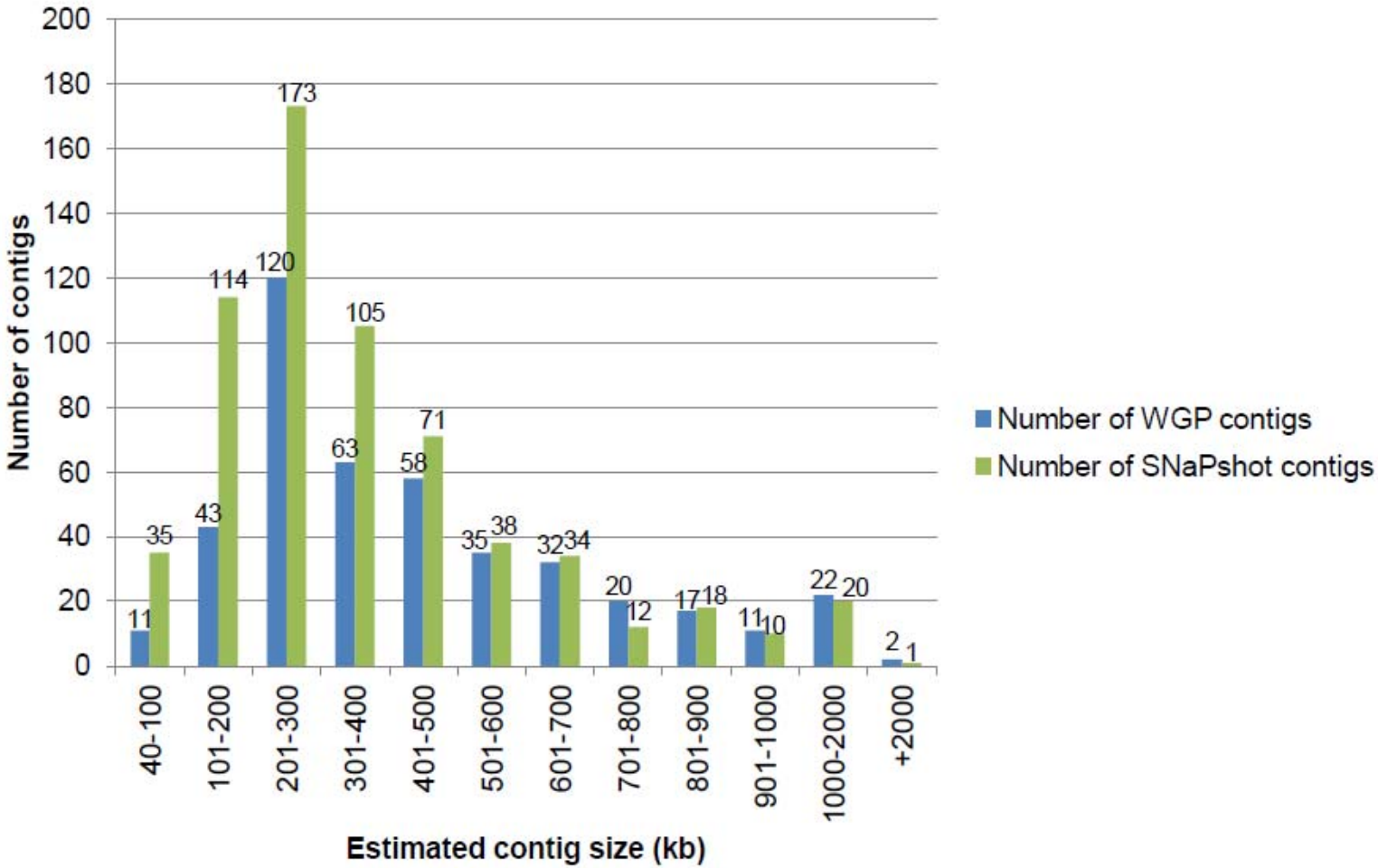
• FPC parameters established by van Oeveren et al. (2011).

	WGPv1	WGPv2
Number of BACs used	13,888	11,238
Percentage of assembled BACs	79.2%	89.2%
Number of contigs	786	853
Comparison with the 12 sequenced contigs		
Number of chimerical contigs for 10 Mb	5.6 (35.7%)	2.8 (15.6%)
Percentage of mis-assembled BACs	26.8%	14.9%

↳ BACs filtering improve the physical map but not enough to built a robust physical map in wheat.

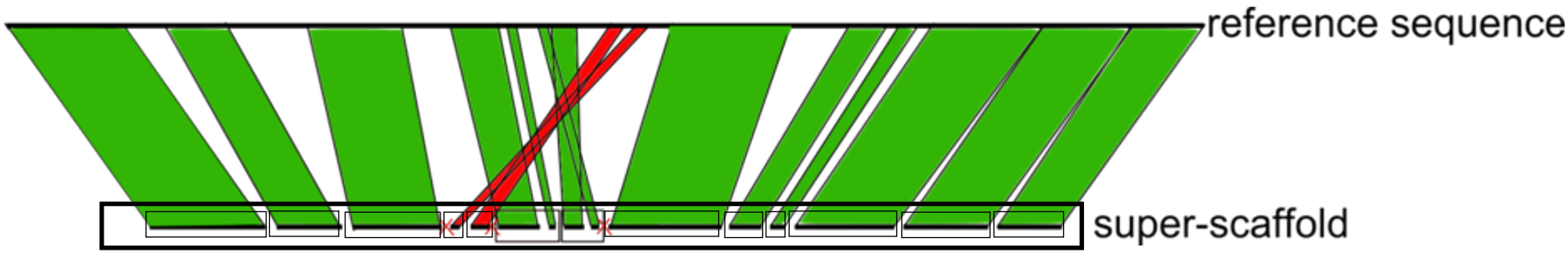
↳ FPC parameters were not enough stringent.

Contigs size distribution



Additional file 1. Distribution of the contig size in the optimal WGP and SNaPshot physical maps.

Without Paired-end

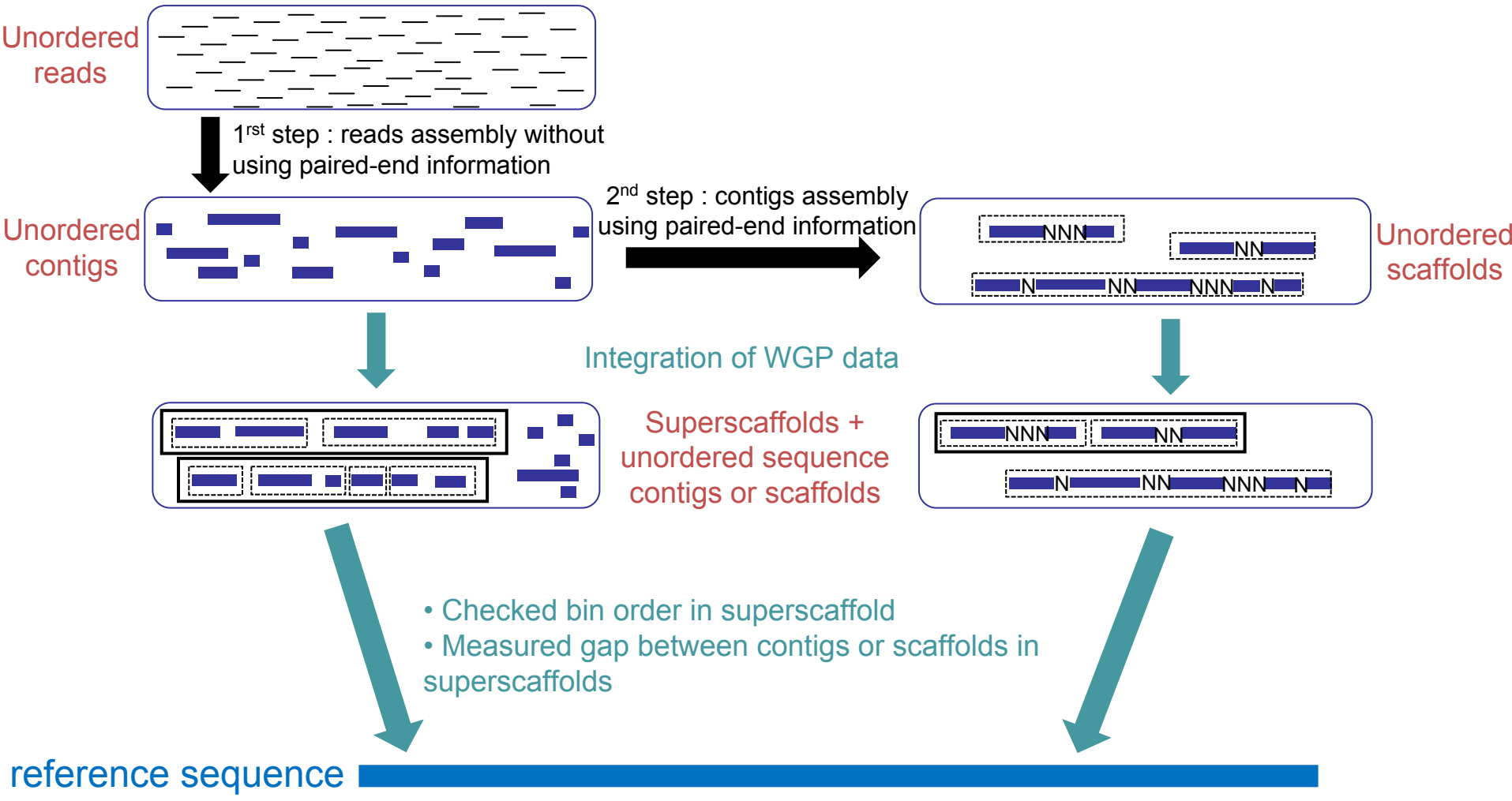


Percentage of gap in superscaffolds = ((real SSC coverage – SSC size) / real SSC coverage)

Percentage of error in bin order = (number of error in bin merger / total number of bin merger)
= 3 / 12 = 25%

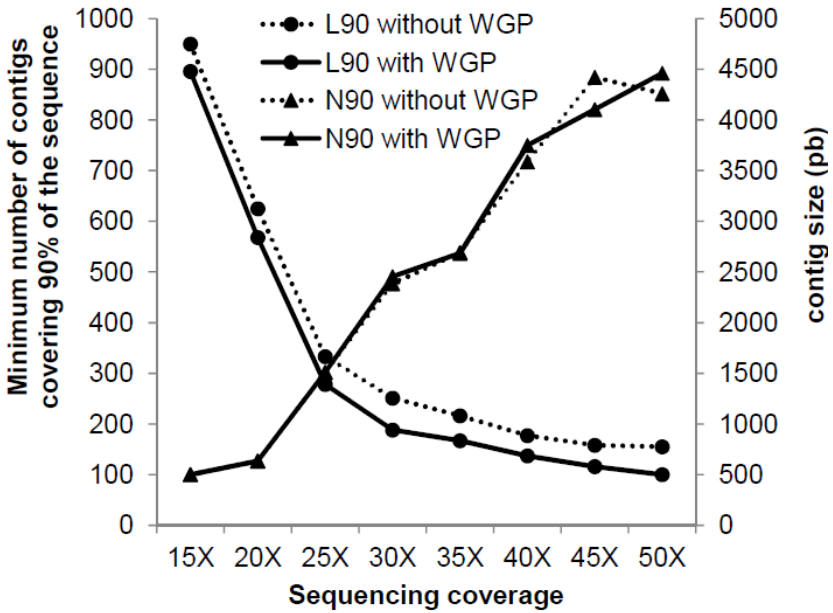
WGP to support sequence assembly

- Re-Sequencing by 454-GS-FLX of 4 reference (Sanger) sequences (600 Kb – 1Mb)
- Series of assemblies using Newbler v2.3:
  - sequence coverage between 15X and 50X (5X steps)
  - with and without paired end (PE) information

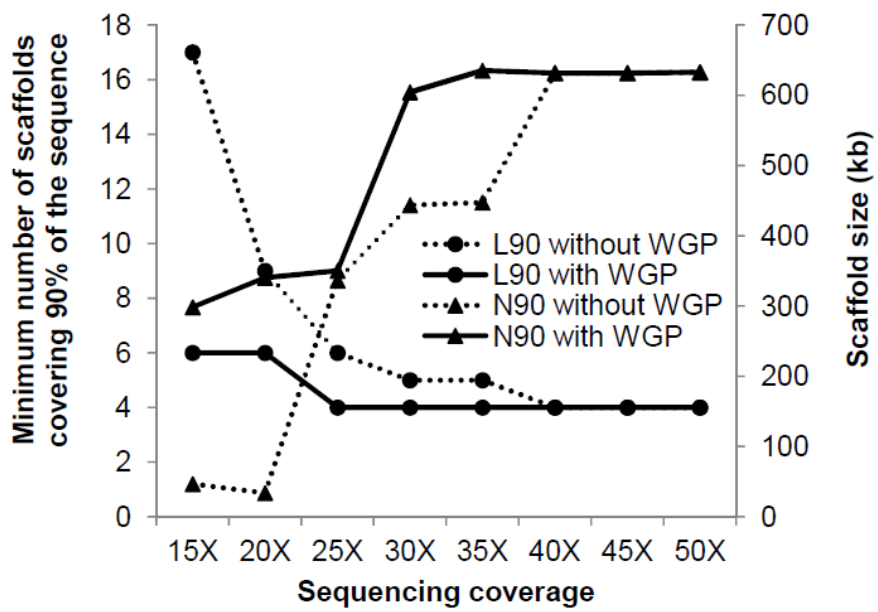


WGP to support sequence assembly

Without Paired-end



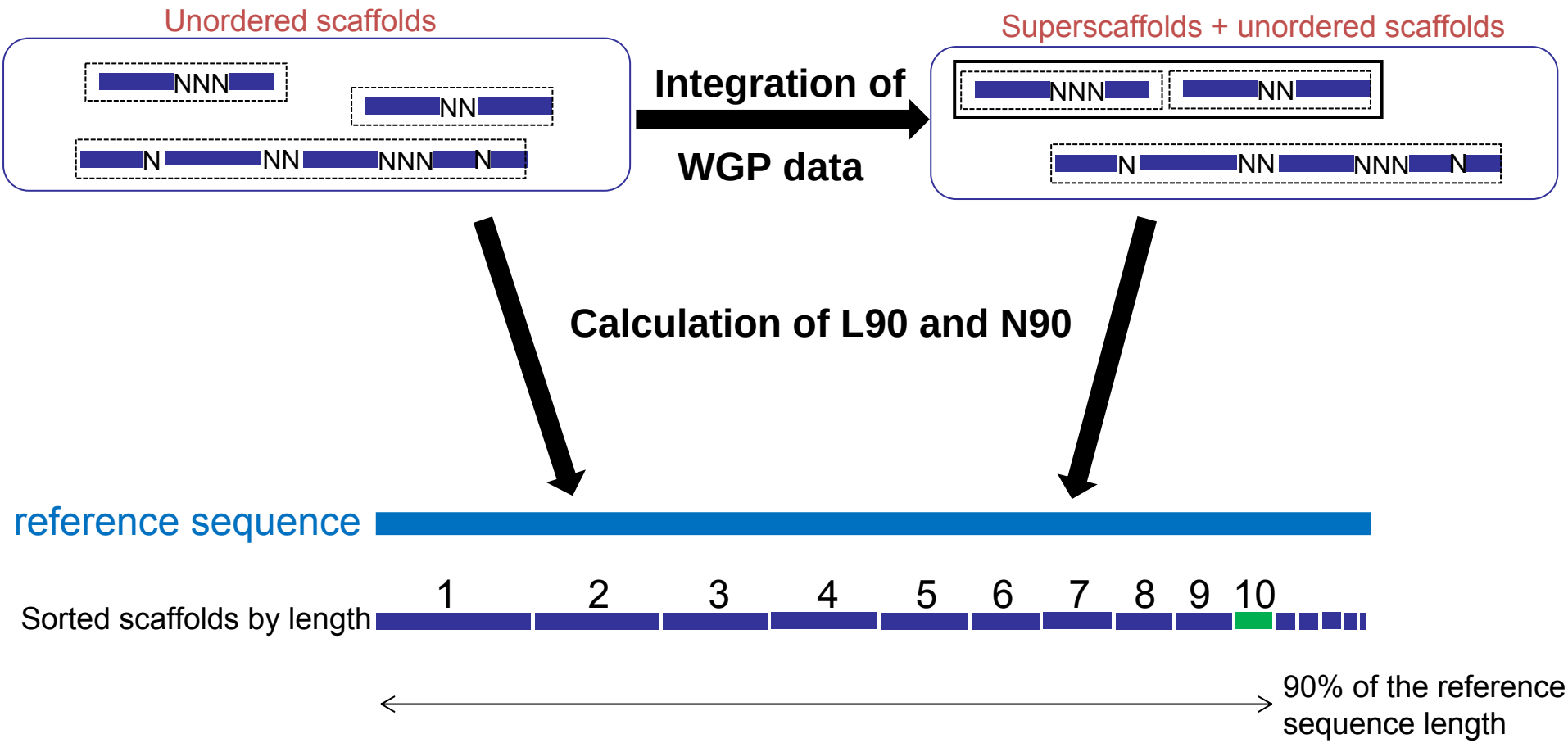
With Paired-end



- Without Paired End sequencing and with PE sequencing at low coverage
  ➡ WGP data integration improve sequence assemblies

- With Paired End deep sequencing (>25X)
  ➡ WGP does not bring any significant improvement in sequence assembly

With Paired-end



N90 = minimum number of contigs to cover 90% of the reference sequence (10)

L90 = length of the shortest contig such that the sum of contigs of equal length or longer is at least 90% of the reference sequence (size of the green contig)