Optimizing the construction of robust physical maps in wheat

By Romain Philippe

January-14-2012
The wheat genome is a challenge for genomic studies and sequencing.
Dissecting the hexaploid wheat genome through flow sorting of individual chromosomes

Sheath fluid
Deflection plates
Excitation light
Waste
Left collector
Right collector
Laser
Scattered light
Flow sorted chromosomes
Flow karyotype
Number of events Relative fluorescence intensity
1BL
3B
Flow sorted chromosome arms
Construction of chromosome or chromosome arm-specific BAC libraries

DT1BL line

(kindly provided by J. Dolezel)
A Physical Map of the 1-Gigabase Bread Wheat Chromosome 3B

Etienne Paux, Pierre Sourdille, Jérôme Salse, Cyrille Saintenac, Frédéric Choulet, Philippe Leroy, Abraham Korol, Monika Michalak, Shahryar Kianian, Wolfgang Spielmeyer, Evans Lagudah, Daryl Somers, Andrzej Kilian, Michael Alaux, Sonia Vautrin, Hélène Bergès, Kellye Eversole, Rudi Appels, Jan Safar, Hana Simkova, Jaroslav Dolezel, Michel Bernard, Catherine Feuillet

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)

Improving physical mapping in wheat

✓ Pilot sequencing showed that:

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

More robust physical map

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 Yalcın,² Antoine Janssen,¹ Hanne Volpin,¹ Keith E. Storomo,² Robert Bogden,² Michiel J.T. van Eijk,¹ and Marcel Prins¹ ³

¹Keygene N.V., Wageningen, The Netherlands; ²Amplicon Express Inc., Pullman, Washington 99163, USA
Whole Genome Profiling (WGP) : a new sequence-based physical mapping technology

**SNaPshot**
- EcoRI, BamHI, XbaI, XhoI and HaeIII sites
- Digested by 5 enzymes
- Labeled with 4 dyes
- Between 40 and 250 bands per BAC

**WGP**
- EcoRI sites
- BAC clone
- EcoRI digestion
- Adapter ligation
- Sequencing each side
- On average, 58 sequence tags of 36 bp per BAC (138kb, 1 EcoRI site / 5.1kb)

**BAC A**
- GAATTCCGGAGCTTCCTCGCTTCTGCTATGACCT
- GAATTCCGGAGCTTCCTCGCTTCTGCTATGACCT
- GAATTCCGGAGCTTCCTCGCTTCTGCTATGACCT
- GAATTCCGGAGCTTCCTCGCTTCTGCTATGACCT

**BAC B**
- GAATTCCGGAGCTTCCTCGCTTCTGCTATGACCT
- GAATTCCGGAGCTTCCTCGCTTCTGCTATGACCT
- GAATTCCGGAGCTTCCTCGCTTCTGCTATGACCT
- GAATTCCGGAGCTTCCTCGCTTCTGCTATGACCT

**BAC C**
- GAATTCCGGAGCTTCCTCGCTTCTGCTATGACCT
- GAATTCCGGAGCTTCCTCGCTTCTGCTATGACCT
- GAATTCCGGAGCTTCCTCGCTTCTGCTATGACCT
- GAATTCCGGAGCTTCCTCGCTTCTGCTATGACCT

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

WGP is theoretically more robust than SNaPshot
WGP pilot project in wheat

≈ 811 Mb

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

12 sequenced contigs (18Mb)
(Choulet et al. Plant Cell 2010)
WGP performs better than SNaPshot for physical mapping in wheat
Can WGP help to generate a high quality reference sequence at reduced sequencing depth?

Physical map (BACs)

454 sequencing of pooled MTP-BACs

Assembling 454 reads

Tag-based Contig ordering

Ordered bin = superscaffolds
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 ($\leq 20X$) but high percentage of gap remains (>24%) => Low quality assembly
- does not improve scaffolding at high sequencing coverage ($\geq 25X$)
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

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

Zeev Frenkel, Etienne Paux, David Mester, Catherine Feuillet, Abraham Korol

University of Haifa

BMC Bioinformatics

Open Access
FPC vs LTC

**FPC**
- High initial stringency to limit chimerical contigs
- Adding singletons to contigs extremity at each step
- Merging contigs at each step

**LTC**
- Low initial stringency to limit the number of contigs
- Checking contigs linearity
- Splitting non-linear contigs by increasing the stringency

**Based on BACs similarity and contigs linearity**

- Cut-off $10^{-40}$
- Cut-off $10^{-35}$
- Cut-off $10^{-30}$
- Cut-off $10^{-25}$
- Cut-off $10^{-15}$

**Elimination of chimerical contigs**

**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
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)
Preliminary results on the synteny with Brachypodium

1056 genes anchored on physical contigs in deletion bins

- 7 genes
- 205 genes
- 75 genes
- 59 genes
- 243 genes
- 79 genes
- 120 genes
- 109 genes
- 90 genes
- 69 genes

Syntenic genes:
- 59 syntenic genes
- 20 syntenic genes
- 27 syntenic genes
- 97 syntenic genes
- 22 syntenic genes
- 30 syntenic genes
- 29 syntenic genes
- 17 syntenic genes

(30.6%)
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)

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

81 ctg (84.2 Mb, 801 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/](http://urgi.versailles.inra.fr/cgi-bin/gbrowse/wheat_FPC_1BL_pub/)
Acknowledgments

Funded by

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

GAP Division
Eduard Akhunov

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

University of Haifa
Abraham Korol
Vova Frenkel

Institute of Experimental Botany
Jaroslav Dolezel
Hana Simkova
Jan Bartos

Oregon State
Eduard Akhunov

Changchun Normal University
Hélène Bergès
Arnaud Bellec
Sonia Vautrin

AmpliCone
Robert Bogden
Keith Stormo

Center for Applied Genomics (CGA)
Adriana Alberti
Patrick Wincker

IGA
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

<table>
<thead>
<tr>
<th>WGPv1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of BACs used</td>
</tr>
<tr>
<td>Percentage of assembled BACs</td>
</tr>
<tr>
<td>Number of contigs</td>
</tr>
<tr>
<td>Comparison with the 12 sequenced contigs</td>
</tr>
<tr>
<td>Number of chimerical contigs for 10 Mb</td>
</tr>
<tr>
<td>Percentage of mis-assembled BACs</td>
</tr>
</tbody>
</table>

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 \( \leq 4 \) tags (bad quality BAC fingerprints) and with \( \geq 40 \) tags (possible mixed BACs) : final subset of 11,238 BACs.

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

<table>
<thead>
<tr>
<th></th>
<th>WGPv1</th>
<th>WGPv2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of BACs used</td>
<td>13,888</td>
<td>11,238</td>
</tr>
<tr>
<td>Percentage of assembled BACs</td>
<td>79.2%</td>
<td>89.2%</td>
</tr>
<tr>
<td>Number of contigs</td>
<td>786</td>
<td>853</td>
</tr>
</tbody>
</table>

Comparison with the 12 sequenced contigs

<table>
<thead>
<tr>
<th></th>
<th>WGPv1</th>
<th>WGPv2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of chimerical contigs for 10 Mb</td>
<td>5.6 (35.7%)</td>
<td>2.8 (15.6%)</td>
</tr>
<tr>
<td>Percentage of mis-assembled BACs</td>
<td>26.8%</td>
<td>14.9%</td>
</tr>
</tbody>
</table>

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 = \((\text{real SSC coverage} - \text{SSC size}) / \text{real SSC coverage}\)

Percentage of error in bin order = \((\text{number of error in bin merger} / \text{total number of bin merger})\)
  
  \[= \frac{3}{12} = 25\%\]
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

1st step: reads assembly without using paired-end information

2nd step: contigs assembly using paired-end information

Integration of WGP data

Superscaffolds + unordered sequence contigs or scaffolds

- Checked bin order in superscaffold
- Measured gap between contigs or scaffolds in superscaffolds

reference sequence
WGP to support sequence assembly

Without Paired-end

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

With Paired-end

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

Unordered scaffolds

Calculation of L90 and N90

Integration of WGP data

Superscaffolds + unordered scaffolds

reference sequence

Sorted scaffolds by length

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)