Improving genome assemblies and capturing genome variation data for applied crop improvement.

David Edwards
University of Western Australia, Australia
Dave.Edwards@uwa.edu.au
Outline

- Sequencing wheat chromosome arms
- Chickpea chromosomal genomics
- Skim GBS based genome assembly
- Validation and improvement of the canola genome
Sequencing wheat chromosome arms

Sequencing wheat chromosome arms

Sequencing wheat chromosome arm 7BS delimits the 7BS/4AL translocation and reveals homoeologous gene conservation

Paul J. Berkman · Adam Skarszewski · Sahana Manoli · Michal T. Lorenc · Jiri Stiller · Lars Smits · Kaitao Lai · Emma Campbell · Marie Kubaláková · Hana Šimková · Jacqueline Batley · Jaroslav Doležel · Pilar Hernandez · David Edwards

Dispersion and domestication shaped the genome of bread wheat

Paul J. Berkman, Paul Visenoi, Hong C. Lee, Jiri Stiller, Sahana Manoli, Michal T. Lorenc, Kaitao Lai, Jacqueline Batley, Delphine Fleury, Hana Šimková, Marie Kubaláková, Song Weining, Jaroslav Doležel and David Edwards
A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome

**The International Wheat Genome Sequencing Consortium (IWGSC)**

An ordered draft sequence of the 17-gigabase hexaploid bread wheat (*Triticum aestivum*) genome has been produced by sequencing isolated chromosome arms. We have annotated 124,201 gene loci distributed nearly evenly across the homeologous chromosomes and subgenomes. Comparative gene analysis of wheat subgenomes and extant diploid and tetraploid wheat relatives showed that high sequence similarity and structural conservation are retained, with limited gene loss, after polyploidization. However, across the genomes there was evidence of dynamic gene gain, loss, and duplication since the divergence of the wheat lineages. A high degree of transcriptional autonomy and no global dominance was found for the subgenomes. These insights into the genome biology of a polyploid crop provide a springboard for faster gene isolation, rapid genetic marker development, and precise breeding to meet the needs of increasing food demand worldwide.
<table>
<thead>
<tr>
<th>Level of sequence duplication (% of assembled chromosome arm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IWGSC</td>
</tr>
<tr>
<td>0.00%</td>
</tr>
<tr>
<td>5.00%</td>
</tr>
<tr>
<td>10.00%</td>
</tr>
<tr>
<td>15.00%</td>
</tr>
<tr>
<td>20.00%</td>
</tr>
<tr>
<td>25.00%</td>
</tr>
<tr>
<td>30.00%</td>
</tr>
<tr>
<td>35.00%</td>
</tr>
<tr>
<td>40.00%</td>
</tr>
<tr>
<td>45.00%</td>
</tr>
</tbody>
</table>

Contig duplication
Genes not supported by read data

Number of Genes with no read supporting them

- 2DL
- 1DL
- 5AL
- 6AS
- 5AS
- 3AL
- 2DS
- 2AL
- 6DL
- 5BL
- 4DL
- 5BL
- 6BL
- 3AS
- 4DS
- 2BS
- 1AS
New assembly

Total assembly size

Assembly size (bp)

Chromosome Arm assembly

- IWGSC
- Non-normalized Velvet
Chickpea (Cicer arietinum) is the second most widely grown legume crop after soybean, accounting for a substantial proportion of human dietary nitrogen intake and playing a crucial role in food security in developing countries. We report the ~738-Mb draft whole genome shotgun sequence of CDC Frontier, a kabuli chickpea variety, which contains an estimated 28,269 genes. Resequencing and analysis of 90 cultivated and wild genotypes from ten countries identifies targets of both breeding-associated genetic sweeps and breeding-associated balancing selection. Candidate genes for disease resistance and agronomic traits are highlighted, including traits that distinguish the two main market classes of cultivated chickpea—desi and kabuli. These data comprise a resource for chickpea improvement through molecular breeding and provide insights into both genome divers.

The Plant Journal (2013) 74, 715-729
doi: 10.1111/tpj.12173

FEATURED ARTICLE

A draft genome sequence of the pulse crop chickpea (Cicer arietinum L.)


National Institute of Plant Genome Research, Aruna Asaf Ali Marg, New Delhi 110067, India
Chickpea desi vs kabuli
Chickpea kabuli reference
Kabuli reference
Kabuli reference

Desi
Kabuli
A chromosomal genomics approach to assess and validate the desi and kabuli draft chickpea genome assemblies

Pradeep Ruperao, Chon-Kit Kenneth Chan, Sammar Azam, Miroslava Karafiátová, Satomi Hayashi, Jana Čížková, Rachet K. Sørensen, Hana Šimková, Chi Song, Jan Vršánek, Annapurna Chitikineni, Paul Vosendi, Poorn M. Gaure, Teresa Millán, Karam B. Singh, Bunyamin Taran, Jun Wang, Jacqueline Batley, Jaroslav Doležel, Rajeev K. Varshney and David Edwards
Skim GBS

- Determine SNPs by sequencing parents and running SGSautoSNP
- Low coverage skim sequence segregating population
- Map reads to the reference genome
- Call genotype where reads cover previously defined SNP
- Impute and clean to define haplotype blocks
Genotype calling

Call genotype of previously predicted SNPs
## Haplotype blocks

<table>
<thead>
<tr>
<th></th>
<th>T</th>
<th>A</th>
<th>G</th>
<th>G</th>
<th>T</th>
<th>C</th>
<th>C</th>
<th>A</th>
<th>G</th>
<th>G</th>
<th>A</th>
<th>T</th>
<th>A</th>
<th>A</th>
<th>A</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>G</td>
<td>C</td>
<td>G</td>
<td>G</td>
<td>C</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TN1</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>A</td>
<td>T</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>TN2</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>A</td>
<td>T</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>TN3</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>A</td>
<td>T</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>T</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TN4</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>A</td>
<td>T</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>TN5</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>G</td>
<td>C</td>
<td>G</td>
<td>G</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>TN6</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>A</td>
<td>T</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>TN7</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>A</td>
<td>G</td>
<td>G</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>G</td>
<td>C</td>
<td>G</td>
<td>G</td>
<td>C</td>
<td></td>
</tr>
</tbody>
</table>
Pre-imputation
After imputation and cleaning
Genome assessment with isolated chromosome sequence data
A) Heat maps on released genome (chromosome’s) size & quality
B-H) Heat maps on improved genome assessment with isolated chromosomes sequence data
I) Gene density plots
J) SNP density plots
Genome assessment with isolated chromosome sequence data
A) Heat maps on released genome (chromosome’s) size & quality
B-H) Heat maps on improved genome assessment with isolated chromosomes sequence data
I) Heat map produced with WGS (ICC 4958)
J) Gene density plots
Kabuli v2.6.2 overall chromosome length has increased from 303.1 Mbp to 423.2 Mbp by placing 1,987 contigs.

Desi v1.1 overall chromosome length has increased from 124.3 Mbp to 416.9 Mbp by placing 133,840 contigs.
Desi vs Kabuli comparison
Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome

Boulos Chalhoub, 1,2,3 France Denoël, 2,3,4,5 Shengyi Liu, 5,6 Isobel A. P. Parkin, 6,7 Haihao Tang, 5,8 Xiyin Wang, 5,8 Julien Chiquet, 13 Harry Beltram, 1 Chaobo Tong, 5 Birgit Sanse, 1,2 Margot Correà, 2 Corinne Da Silva, 1 Jérémy Just, 3 Cyril Falentin, 3 Chun Shin Koh, 14 Isabelle Le Claric, 1 Maria Bernard, 2 Pascal Bento, 7 Benjamin Noel, 2 Karine Labadie, 2 Adriana Alberti, 8 Mathieu Charles, 1 Dominique Arnault, 1 Hui Guo, 9 Christian Davioud, 26 Salam Alamery, 7 Kamel Jabbari, 1,3,9 Meiblei Zhao, 19 Patrick P. Ediger, 26 Houda Chelafla, 2 David Taieb, 2 Gilles Lassasse, 23 Imène Mestiri, 23 Nicolas Schuel, 15 Marie-Christine Le Paslier, 15 Guangyi Fan, 23 Victor Renault, 23 Philippe E. Bayer, 1 Azniezka A. Golisz, 17 Sahana Manoli, 7 Tae-Ho Lee, 9 Vinh Ha Dinh Thii, 18 Sahaneh Chalabi, 1 Qi Dong, 24 Chuchuan Fan, 24 Reece Tollaere, 17 Yunhai Lu, 1 Christophe Rattail, 1 Jinxiong Shen, 24 Christine H. D. Sidebottom, 34 Xinfa Wang, 2 Aurélie Canaguier, 4 Aurélie Chauveau, 15 Aurélie Berard, 15 Gwennédile Deniot, 1 Mei Guan, 25 Zongzhong Lin, 26 Fengming Sun, 26 Yong Pan, 26 Erk Lyons, 27 Christopher D. Town, 1 Ian Bancroft, 28 Xiaowu Wang, 29 Jinhong Meng, 24 Jianing Ma, 36 Chris Fries, 36 Graham J. King, 35 Dominique Bruneau, 35 Réjine Delouche, 26 Michal Renard, 27 Jean-Marc Aury, 2 Keith L. Janssen, 27 Jacqueline Batley, 19,22 Rod J. Snowdon, 35 Jorg Tost, 35 David Edwards, 35,36 Yongming Zhou, 36 Wei Hua, 35 Andrew G. Sharpe, 1 Andrew H. Paterson, 1 Christopher Guan, 35 Patrick Whisson 1,2,3,4.

Oilseed rape (*Brassica napus* L.) was formed ~7500 years ago by hybridization between *B. rapa* and *B. oleracea*, followed by chromosome doubling, a process known as allopolyploidy. Together with more ancient polyploidizations, this conferred an aggregate 72+ genome multiplication since the origin of angiosperms and high gene content. We examined the *B. napus* genome and the consequences of its recent duplication. The constituent *A* and *C* subgenomes are engaged in subtle structural, functional, and epigenetic cross-talk, with abundant homoeologous exchanges. Incipient gene loss and expression divergence have begun. Selection in *B. napus* oilseed types has accelerated the loss of glucosinolate genes, while preserving expansion of oil biosynthesis genes. These processes provide insights into allopolyploid evolution and its relationship with crop domestication and improvement.
• Both parental individuals high coverage (~50x)
• 92 double haploid Tapidor x Ningyou individuals, average coverage 1.6x
• Novel algorithm: contigPlacer
• Uses tagging SNPs per contig and compares genotype patterns with placed contigs, places unplaced contigs
Darmor genome

- 10 A-chromosomes, 9 C-chromosomes,
- 22 sets of unplaced contigs
- Assembled size: 850.29 Mbp
- 645.95 Mbp placed (75.8%), 204.33 Mbp
  unplaced contigs
• Identified 1,006,985 SNPs

• ~60% alleles called, after imputation ~80%

• Data in contigPlacer
• Before contigPlacer: 645.95 Mbp placed (75.8%), 204.33 Mbp unplaced contigs

• After contigPlacer: 798.95 Mbp placed (93.9%), 51.33 Mbp unplaced

• 98.5% of unplaced contigs with initial chromosome assignment mapped to the same chromosome
Fixing minor structural errors is important for accurate trait fine mapping.
Conclusions

• Many high quality published genomes can be improved

• Chromosomal genomics, skimGBS and bioinformatics can validate and improve genome assemblies

• Quality genome assemblies improve trait association, they are essential for pan genome assembly and assessment of structural variation
Acknowledgements

Philipp Bayer
Pradeep Ruperao
Juan Montenegro
Kenneth Chan
Michal Lorenc
Agnieszka Golicz
Kaitao Lai
Paul Visendi
Paula Martinez
Jenny Lee

Paul Berkman
Jiri Stiller
Sahana Manoli

Jacqueline Batley
Alice Hayward
Emma Campbell
Jessica Dalton-Morgan
Satomi Hayashi
Reece Tollenaere

Hana Šimková
Marie Kubaláková
Jaroslav Doležel

Tim Sutton
Deepa Jaganathan
Rajeev Varshney
(and colleagues)

Contact:
Dave.Edwards@uwa.edu.au