The Big B of Bread Wheat -3B - exploring the structure, function and evolution of the hexaploid wheat genome

Paux E[1], Sourdille P[1], Salse J[1], Saintenac C[1], Choulet F[1], Leroy P[1], Korol A[2], Spielmeyer W[3], Lagudah E[3], Somers D[4], Kilian A[5], Allaux M[6], Bergès H[7], Appels R[8], Dolezel J[9], Bernard M[1] and Feuillet C[1]

[1]INRA, UMR1095 GDEC, Clermont-Ferrand, France, [2]Institute of Evolution, University of Haifa, Israel, State Agricultural and Biotechnology Centre, [3]CSIRO Plant Industry, Canberra, Australia, [4]Agriculture and Agri-Food Canada, Cereal Research Centre, Winnipeg, Canada, [5]Diversity Arrays Technology Pty Ltd, Yarralumla, Australia, [6]INRA, Unité de Recherches en Génomique-Info, Évry, France, [7]INRA-Centre National de Ressources Génomiques Végétales, Toulouse, France, [8]Centre for Comparative Genomics-Murdoch University, Australia, [9]Laboratory of Molecular Cytogenetics and Cytometry, IEB, Olomouc, Czech Republic.

ABSTRACT

Wheat genomics offers powerful tools to understand the molecular basis of phenotypic variation, accelerate gene cloning and marker-assisted selection as well as to improve the exploitation of genetic diversity for efficient crop improvement. While rice and maize improvement is profiting already from information derived from genome sequences, wheat is lagging behind without a genome sequence project underway. Physical maps anchored to genetic maps are the substrate for genome sequencing and they provide efficient tools for marker development, map-based cloning, QTL mapping, as well as for structural, functional, and comparative genomics studies. In the framework of the IWGSC, we have developed a physical map of chromosome 3B and established the proof of concept for physical mapping of the 21 bread wheat chromosomes through a chromosome-based approach. The 3B physical map consists of 1036 contigs with an average size of 783 kb that cover 811 Mb *i.e.* 82% of the chromosome with about ~10x coverage. The physical map was anchored to deletion and genetic maps with 1443 markers thereby providing a framework for efficient map-based cloning and marker development through BAC-end and contig sequencing. The 3B physical map is currently used for map-based cloning, recombination and LD studies as well as to analyse the wheat genome composition, organisation, function and evolution.

INTRODUCTION

In the past 7 years, large international collaborative efforts have resulted in the development of dense wheat genetic maps with hundreds of genes and QTLs identified, and the construction of several BAC libraries from diploid and polyploid wheat. The availability of these resources accelerated the isolation of agronomically important genes by positional cloning. Despite these first breakthroughs, the impact of gene isolation and marker development on wheat improvement remains low, especially for complex agronomic traits such as yield and quality which are based on quantitative trait loci (QTL) that are extremely challenging to identify and isolate. Physical maps anchored to genetic maps are the substrate for accelerated map-based cloning and QTL mapping, comparative genomics studies, genome sequencing and, structural and functional genome analyses. In 2005, the International Wheat Genome Sequencing Consortium (IWGSC; http://www.wheatgenome.org) was launched with the aim of developing a physical map of the hexaploid bread wheat (*Triticum aestivum* L.) genome and establishing the foundation for future genome sequencing.

To establish the physical map of the 17 Gb bread wheat genome with a 10 X genome coverage, more than 1.4 million BAC clones with an average size of 120 kb need to be fingerprinted, assembled into contigs, and anchored to genetic maps. While fingerprinting millions of BAC clones is technically feasible using high information content fingerprinting (HICF) techniques[1], their specific assembly into contigs faithfully representing individual chromosomes and the anchoring of the homoeologous BAC contigs onto the genetic maps represents a daunting task. An alternative to reduce the complexity of the task is to use a "chromosome-based strategy" which relies on the isolation of individual chromosomes or chromosome-arms by flow cytometry for the construction of chromosome-specific BAC libraries[2]. In this way, the wheat genome can be dissected into fragments as small as 220 Mb, *i.e.* 1.3% of the total genome, offering a significant reduction in complexity, and thereby facilitating the development of the wheat physical map in a step-wise manner, one chromosome (arm) at a time, and providing the opportunity to split the task between laboratories in an internationally coordinated effort.

Here, we report the construction of a physical map of the largest wheat chromosome, 3B (~1Gb) and demonstrate that the chromosome-based approach is feasible and suitable for the construction of the 21 hexaploid wheat chromosome physical maps. We provide examples of the use of the chromosome 3B map for structural and evolutionary analyses of the wheat genome and for genome-based tool development for wheat breeders.

CONSTRUCTION OF A 10 X PHYSICAL MAP OF CHROMOSOME 3B

The 67,968 3B BAC clones of a chromosome 3B-specific BAC library from cv. Chinese Spring[3] were fingerprinted using a slightly modified High Information Content Fingerprinting (HICF) SNaPshot protocol[1]. A total of 56,952 high quality fingerprints (84%) were obtained and assembled using the Fingerprinted Contigs Program (FPC)[4]. The automated build of chromosome 3B had 1991 contigs containing 44,008 clones that represented a total contig size of 960 Mb.

The automated assembly was used to define a Minimal Tiling Path (MTP) consisting of 7440 BAC clones. The MTP was further used for systematic BAC-end sequencing, three dimensional pooling for efficient PCR-based screening and anchoring and for the establishment of a tiling array. Contigs were then manually merged at lower stringency using information provided by contig anchoring. This resulted in a final assembly of 1036 contigs with an average size of 783 kb. This assembly covers 811 Mb, and, with an estimated size of 995 Mb[5] for chromosome 3B, represents 82% of the chromosome.

Full benefit of a physical map is only achieved when contigs are ordered along the chromosomes. This is performed by anchoring the contigs with molecular markers followed by integration of the anchored contigs onto genetic maps. The 3B physical map was anchored by PCR screening of the BAC library with molecular markers. To reduce significantly the number of PCR reactions and establish the most cost-efficient screening strategy, three BAC clone pooling schemes were considered: six-dimensional (6-D) pooling and three-dimensional (3-D) pooling of the entire BAC library as well as 3-D pooling of the minimal tiling path (MTP). Six-D pooling[6] reduces the number of PCR needed to get the complete unambiguous address of a BAC clone in a single round, by a factor of 384. It is the most efficient but also the most expensive strategy and requires specific automated platforms that are not available in every laboratory. Three-D pools that exist in pools of 384-well plates, rows, and columns are easier and cheaper to develop. For the 3B BAC library, the three dimensional pools represent 177 plate pools, 2832 (16x177) lane pools and 4248 (24 x 177) column pools. With an average number of 8 positive plates per marker, the total number of PCR reactions required for screening these pools with the 1000 markers is more than 500,000. Thus, the strategy is straightforward but remains laborious and costly. To reduce further the number of PCR reactions, we evaluated the possibility of using the MTP determined after the first automated assembly which at this stage consisted of 7440 BACs. When pooled in three dimensions, the MTP represented only 20 pool plates, 16 lane pools and 24 column pools. Thus, only 60 PCR reactions are needed per marker for a total of less than 60,000 PCR reactions. One possible drawback of the approach, however, is that in the MTP,

a given region is only covered by a single BAC clone which might not grow or might produce low quality DNA. Consequently, the risk of negative amplification with one marker is higher than in strategies employing the complete BAC library where a region is covered 4 to 12 times. To avoid this problem, we developed a complementary strategy that relies on the combination of results from plate pools screening of the entire BAC library (177) and FPC data using a new software called "elephant" (electronic physical map anchoring tool)[7]. The program works as follows:

● Using the FPC file, "elephant" partitions the contigs into short sections, hereafter referred to as "segments", by splitting the contig at each branching point (e.g. a clone finishing or joining the assembly) (Figure 1A) and establishes a list of clones for each segment.
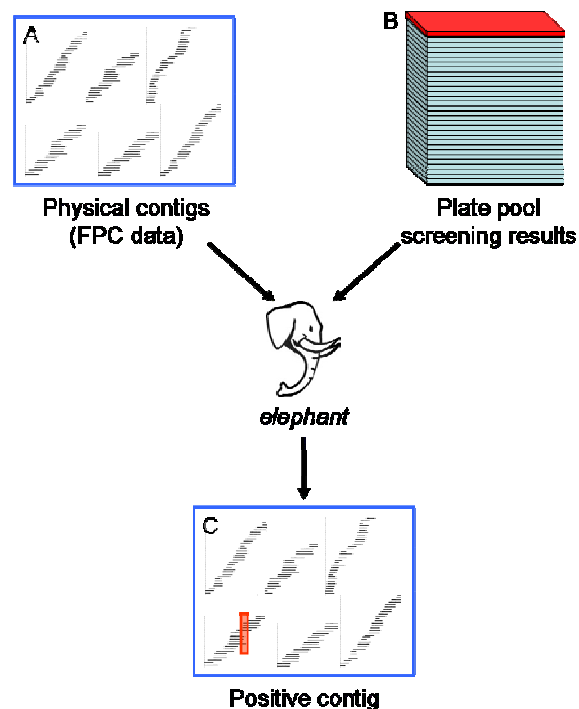

Figure 1: Schematic representation of the Elephant Strategy.

● For each marker, elephant combines the results from pool screening with the pool composition to establish a list of candidate clones harbouring the markers (Figure 1B).

● Then, "elephant" compares the two lists and scores each segment. All segments with a score above the threshold of 13 are selected as candidates.

● The output text file reports the candidate contigs for each marker with the start and end position of the corresponding segments, as well as a list of positive, negative, and missing clones in these segments (Figure 1C).

● Validation of a candidate contig can then be performed by testing one single clone instead of 384 clones from a positive plate.

With this approach, it is possible to identify missing BAC addresses through a single PCR experiment on plate pools (177 individual PCRs) with a reduction of the number of PCR by 384 fold. The "elephant" package is under a GPL license and is freely available at http://urgi.versailles.inra.fr/tools/elephant/download/.

More than 2000 molecular markers (SSRs, ESTs, ISBPs[8]) were assayed and 1443 were finally linked to 680 BAC contigs representing 611 Mb and 75% of the 3B physical map. Very few contigs were anchored with the three types of markers and most of the contigs were anchored by a single type of marker indicating that the different classes of markers cover different regions of the genome and should be used in combination to ensure optimal representation of the chromosome. About half of the contigs were anchored with ISBP markers[8] only, demonstrating the importance of developing this type of markers in a physical mapping project.

The physical map was then integrated onto genetic maps following two strategies: meiotic mapping and deletion mapping. The 680 contigs harbouring a marker were assigned to one of 16 deletion bins using Chinese Spring cytogenetic stocks[9]. In total, 599 were unambiguously assigned to one of the 16 physical intervals delimited by the deletions bins. If deletion mapping is interesting because it is independent from recombination and polymorphism, the reduced number of lines available to date and the large size range of the deletions obtained with this method strongly limit its resolution. One of the most promising high resolution mapping approaches for wheat is radiation hybrid (RH) mapping. A RH panel is currently under construction for 3B and preliminary assays using markers originating from contigs of the 3BL7-0.63-1.00 deletion bin and from sequenced regions indicates a resolution of about 263 kb per break (S. Kianian, pers. comm.). If the panel can be increased and the results are confirmed, RH mapping will become the main resource for anchoring wheat physical maps in the future.

Further integration of the 3B physical map was performed by meiotic mapping using two types of linkage maps. The first one consisted of a single reference genetic map developed from a F2 population (CsRe) derived from the cross between Chinese Spring (Cs), the reference cultivar chosen by the IWGSC for genome sequencing and the French cultivar Renan (Re). To date, 102 SSR and ISBP markers have been mapped using 376 F2 individuals of the CsRe population. Eighty nine of them were anchored to a contig allowing the integration of 75 individual contigs to the genetic map. Using the genetic mapping information, it was possible to order 80% of them. We are currently developing RILs from 1800 additional F2 individuals of the CsRe population. This reference stock should provide enough resolution to integrate all contigs except those located in the 6 most centromeric bins where recombination is totally suppressed. It will be available to the international community to support the IWGSC physical mapping effort.

The optimal use of the 3B physical map for map-based cloning of genes and QTLs present on this chromosome can only be achieved if a majority of contigs are integrated with a maximum of mapped markers. To accomplish this, a "3B neighbour map" was established that combines the position of loci from 13 different genetic maps including the CsRe map. This latter was used as a framework to extrapolate the position of loci from other maps following the same criteria as for the IBM neighbour map of maize[10]. The 3B neighbour map contains 636 SSR, RFLP, STS, DArT and ISBP markers with 30 'anchor' markers which position is robust and that define 31 genetic bins in which the relative loci positions can slightly differ from one map to another. In total, 213 contigs were anchored in these genetic bins.

Based on the results obtained in this project, we propose a strategy for the establishment of an integrative physical map of the bread wheat genome following the chromosome-based approach in five steps: (1) chromosome sorting and BAC library construction, (2) Fingerprinting, contig assembly and MTP design, (3) MTP screening with mapped markers (forward anchoring), (4) BAC end sequencing, marker development from the contigs and genetic mapping (reverse anchoring) and (5) integration of the forward and reverse anchoring data into an integrative physical map. This can be applied now to the rest of the wheat chromosomes as well as to other eukaryotic genomes facing similar size and complexity challenges.

ALIGNMENT OF THE CHROMOSOME 3B PHYSICAL MAP TO THE RICE GENOME

The physical map of chromosome 3B provides a new platform for high resolution comparative analyses between wheat and other grass genomes. Alignment of the 3B map with the rice genome was performed using the genic sequences (ESTs/STS) present on the contigs that were integrated to the CsRe genetic map. A detailed analysis was conducted in the terminal part of the short arm where 14 of 27 contigs contained 49 EST/STS that identified 38 rice genes on chromosome 1. Four inversions of the order along the chromosomes as well as contigs for which the orthologous rice genes are at non collinear positions were observed. These results indicate a higher level of rearrangement than previously identified by comparative mapping using wheat ESTs[11] and suggest that unlike with mammal genomes[12], rearrangements between plant genomes are likely too important to allow ordering the physical map of a genome using another genome sequence.

Further comparisons will be possible soon with the genome sequence of *Brachypodium distachyon* L., an emerging model species for temperate grasses (www.brachypodium.org). Comparative physical mapping will also be interesting with barley that is more closely related to wheat and for which a whole genome

physical map is underway (www.barleygenome.org). These studies will provide new information on the degree of colinearity between the grass genomes and give indications at which level physical mapping in wheat can benefit from other genome sequences and maps.

A FOUNDATION FOR STRUCTURAL AND FUNCTIONAL ANALYSES OF THE WHEAT GENOME

One of the immediate applications of a physical map is positional cloning. With an average contig size of 783 kb, the chromosome 3B physical map, allows, in a single step, to land on any target locus providing that recombination is compatible with fine mapping of the target gene. This was recently demonstrated in the map-based cloning project of the stem rust resistance gene *Sr2*. The gene was located on the short arm of chromosome 3B in a genetic interval of approx 1 cM between makers CA640157 and gwm533[13]. Screening of the 3B MTP with these markers allowed us to identify a contig for each of these markers. Additional ISBP and SSR markers were designed from BAC-end sequences, and used on a high resolution mapping population to define a physical interval of approximately 600 kb that spans the *Sr2* locus (W. Spielmeyer, unpublished data).

The 3B physical map also provides the foundation for sequencing and in-depth studies of the wheat genome composition and organization. Until now, sequencing on the wheat genome has concerned short BAC contigs at target loci in map-based cloning projects or large-scale random sequencing of genes (> 800,000 ESTs), BACs[14] or plasmid ends[8, 15]. If these analyses provided useful information on the overall composition of the wheat genome (gene numbers, gene density, TE content), the distribution of genes and repeats over large stretches of several megabases in wheat remains unknown. Here, we have used the physical map information to select 13 contigs of 800 kb to 3.2 Mb originating from different regions of chromosome 3B for sequencing and annotation. Preliminary analysis of the 3.2 Mb region (Figure 2) revealed the presence of 50 genes and 609 transposable elements (TEs). TEs represent 60% of the sequence, the majority of them being LTR-retrotransposons (72%) and CACTA family DNA transposons (20%).
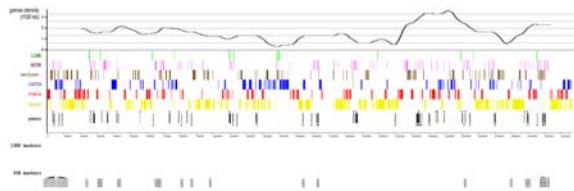

Figure 2: Schematic representation of the annotation of a 3.2 Mb contig of sequence on the short arm of chromosome 3B. TEs are represented as coloured bars (LINE:green;CACTA: blue, Copia: red; GYPSY:yellow)

while genes are indicated with black bars with the gene density distribution shown on the top of the figure.

About 30 annotated features are potentially unknown TEs showing that large wheat genome sequences are an important resource for new TE discovery. The analysis also shows that genes and TEs are not randomly distributed along the sequence (Figure 2).

A physical map offers also an unlimited source of information for marker development and breeding. Prior chromosome 3B is fully sequenced, sample sequencing from the minimal tiling path already offers a great source of new markers for wheat breeding. We have reported previously the potential of using BAC end sequences from the MTP to develop genome specific markers that are evenly distributed along the chromosome[8]. In total, we already developed several hundreds of ISBP and SSR markers from chromosome 3B using this information. The potential for SSR and ISBP development is even larger from the sequenced regions. For example, on the 3.2 Mb of contiguous sequence currently under analysis, we identified about 1500 potential ISBP and SSR markers and have already developed more than 100 new markers for breeding in the region (unpublished data). This demonstrates the great potential of wheat genome sequencing for unlimited marker development and marker assisted selection (MAS) in wheat.

CONCLUSIONS AND FUTURE DIRECTIONS

The construction of the first physical map of the largest wheat chromosome has demonstrated that physical mapping in wheat can be successful following a chromosome-based strategy. This opens the route for the construction of the physical map of the 20 remaining bread wheat chromosomes as well as for other eukaryotic genomes that face similar challenges of genome size and complexity. In wheat, the entire strategy of sorting and BAC library construction is in place now for the rest of the chromosomes and projects are already underway (see http://www.wheatgenome.org/projects.php and http://www.ueb.cas.cz/Olomouc1/LMCC/Resources/resourres.html#chrs).

The current 3B chromosome-landing ready physical map is suitable for scientists and breeders to perform efficient map-based cloning of the >40 genes and QTLs of interest identified on 3B, and for marker development at target loci for MAS. It represents a solid platform that can be extended rapidly with array-based technologies for anchoring and high-throughput sequencing techniques. A 15X coverage map is underway for future chromosome 3B sequencing which given the current revolution in sequencing technologies can now be envisaged for a reasonable cost. Functional analysis of chromosome 3B is also underway using tiling arrays of the MTP to identify and characterize the content and expression of the genes located along the chromosome. Finally, the 3B physical map represents an invaluable

tool for comparative analysis in a recent polyploid, domesticated and cultivated species.

REFERENCES

1. M. C. Luo *et al.*, *Genomics* **82**, 378 (2003).
2. J. Dolezel *et al.*, *Chromosome Res.* **15**, 51 (2007).
3. J. Safar *et al.*, *Plant J.* **39**, 960 (2004).
4. C. Soderlund, S. Humphray, A. Dunham, L. French, *Genome Res.* **10**, 1772 (2000).
5. B. S. Gill, B. Friebe, T. Endo, *Genome* **34**, 830 (1991).
6. P. E. Klein *et al.*, *Genome Res.* **10**, 789 (2000).
7. E. Paux *et al.*, *Funct. Integr. Genomics* **8**, 29 (2007).
8. E. Paux *et al.*, *The Plant Journal* **48**, 463 (2006).
9. M. Dilbirligi *et al.*, *Genetics* **166**, 461 (2004).
10. K. C. Cone *et al.*, *Plant Physiol* **130**, 1598 (2002).
11. M. La Rota, M. E. Sorrells, *Funct. Integr. Genomics* **4**, 34 (2004).
12. S. G. Gregory *et al.*, *Nature* **418**, 743 (2002).
13. R. Kota, W. Spielmeyer, R. A. McIntosh, E. S. Lagudah, *Theor. Appl. Genet.* **112**, 492 (2006).
14. K. M. Devos *et al.*, *Proc. Natl. Acad. Sci. U. S. A.* **102**, 19243 (2005).
15. W. Li *et al.*, *Plant J.* **40**, 500 (2004).