# Sequencing and mapping of bread wheat chromosome 7B

Sekvensering og genkartlegging av brødhvetekromosom 7B

Philosophiae Doctor (PhD) Thesis

Tatiana Belova

Department of Plant Sciences
Faculty of Veterinary Medicine and Biosciences
Norwegian University of Life Sciences

Ås 2014

Thesis number 2014:68
ISSN 1894-6402
ISBN 978-82-575-1230-9

# Table of contents

# Acknowledgements

The work presented here has been carried out in the time period 2010-2014 at the Department of Plant and Environmental Sciences (IPV) at the Norwegian University of Life Sciences. It has been funded by grants from the Norwegian Research Council (project no.199387/I99) and Graminor AS.

My main supervisor Odd-Arne Olsen deserves my deepest gratitude for taking me on the boat of this innovative and interesting 7B wheat genome project. Thank you for being a great captain, always supportive, motivated and determined. Thank you for all the scientific and life discussions which have made me become stronger!

Thanks to my dear colleague Simen Sandve for his support, inspiration, creativity, and willingness to share knowledge. Thank you for your contagious motivation and never-ending help with "R-scripting and manuscripting", and for always being available for discussions. It has been great fun to work with you.

I would like to thank my co-supervisors Åsmund Bjørnstad and Sigbørn Lien, and all my co-authors for their contribution to my PhD dissertation. It has been a pleasure to work along with great experts in the genomics field. Special thanks go to Nathan Springer, Shahryar Kianian, Ajay Kumar, Pierre Sourdille, Ethienne Paux, and Francois Balfourier for your great collaboration and hospitality during my stay in your research groups. Thanks to all members of 7B International Wheat Genome Sequencing Consortium (IWGSC) for providing mapping and sequence data to work with. I am thankful to Vova Zeev and Abraham Korol for your significant contribution to the last manuscript.

I would also like to express my gratitude to my colleagues at IPV and Cigene for sharing their knowledge, good moments, and friendly working atmosphere.

I am most grateful to my parents Yurij and Ludmila for their constant support and love. Thanks for supporting my decision to move to Norway and encouragement to be in science. Special thanks to Ragnar Bratlie for your support, care, friendship and always positive attitude to life. And finally I wish to thank all my friends, Viktor D., Katya A., Katya Y., Katya P., Felipe R., Nina Z., Nastya M., Vika P., Yulia P. and others for always supporting me and bringing happiness and laugh into my life.

Tanya July 2014

# List of papers

I. Belova, T., Zhan, B., Wright, J., Caccamo, M., Asp, T., Simkova, H., Kent, M., Bendixen, C., Panitz, F., Lien, S., Dolezel, J., Olsen, O.-A., Sandve, S.R. **Integration of mate pair sequences to improve shotgun assemblies of flow-sorted chromosome arms of hexaploid wheat.** *BMC genomics* 2013, **14**:222

II. Belova, T., Grønvold, L., Kumar, A., Kianian, S., He, X., Lillemo, M., Springer, N.M., Lien, S., Olsen, O.-A., Sandve, S.R. **Utilization of deletion bins to anchor and order sequences along the wheat 7B chromosome.** *TAG Theoretical and applied genetics Theoretische und angewandte Genetik 2014,* **127** (9):2029-2040

III. Belova, T., Frenkel, Z., Zhan, B., Lillemo, M., Korol, A., Paux, E., Balfourier, F., Sourdille, P., Simkova, H., Kubalakova, M, Dolezel, J., Cattonaro, F., Li, L., Min, J., Chen, J., Yang,Y., Xu, X., Kent, M., Lien, S., Sandve, S.R. and Olsen, O.-A. **Anchoring physical contigs of bread wheat chromosome 7B long arm.** Manuscript

# Abbreviations

| | |
|---|---|
| IWGSC | International Wheat Genome Sequencing Consortium |
| BAC | Bacterial artificial chromosome |
| bp | Base pair |
| CSS | Chromosome survey sequences |
| DaRT | Diversity Array Technology |
| EST | Expressed sequence tag |
| MAS | Marker assisted selection |
| MP | Mate pair |
| MTP | Minimum tiling path |
| NGS | Next-generation sequencing |
| PCR | Polymerase chain reaction |
| PE | Paired end |
| RH | Radiation hybrid |
| SNP | Single nucleotide polymorphism |
| TE | Transposable element |

# Summary

The rapid development in DNA sequencing technologies in the recent years have led to the sequencing of several large and complex plant genomes including maize. Recently, the International Wheat Genome Sequencing Consortium (IWGSC) released a draft sequence of bread wheat genome. Using flow-cytometric sorting, wheat chromosome arms were isolated and sequenced with the paired end Illumina technology platform. This resulted in the generation of thousands of sequence contigs with N50 <4 Kb, the so called chromosome survey sequence or CSS of bread wheat. Wheat CSS assemblies are highly fragmented, which decrease the information content of the assemblies. This is caused by the extreme repeat content (>80%) leading to assembly fragmentation even at the single chromosome level. The work presented in this thesis is part of the Norwegian participation in IWGSC and describes integration of mate pair sequences to improve 7B CSS and anchoring of the 7BL BAC-contig physical map to the genetic and molecular maps.

In Paper I, we assess for the first time the effect of integrating mate pair sequences from flow sorted chromosome arms to reduce the fragmentation of the shotgun assemblies of chromosome arms of bread wheat. Three mate pair (MP) libraries with 2 Kb, 3 Kb, and 5 Kb insert size were sequenced to a total coverage of 89X and 64X for the short and long arm of chromosome 7B, respectively. Scaffolding using the SSPACE software tool showed moderate effect on 7B assembly contiguity and gene space fragmentation. We suggest that this effect is related to the use of DNA produced by multiple displacement amplification reaction of flow-sorted chromosome arms of 7B which is known to contain chimeric DNA molecules that significantly reduced usefulness of MP.

In Paper II, we report on the first high-density deletion bin map of a wheat chromosome 7B generated with a high-density Comparative Genome Hybridization (CGH) Nimblegen array. By using the recently published chromosome survey sequences of bread wheat A, B and D subgenomes (IWGSC data repository at http://wheat-urgi.versailles.inra.fr/ ) to design 7B specific probes we assign ~8% of the 7B chromosome sequence into 9 chromosomal bins. Also our study confirmed and further delineated the former mis-estimation of deletion length and deletion type in Del7BL-3, Del7BL-13 and Del7BL-5 deletion stocks.

In Paper III, we have produced the first anchored physical map of wheat chromosome 7B long arm. To achieve this we used a three step strategy of deletion bin mapping, genetic

mapping and finally synteny-based mapping using the closely related species *Brachypodium*, rice and sorghum. A total of 109 out of 125 7BL physical contigs were assigned to a chromosomal position. Among them 92 physical contigs which span ~95% of 7BL sequence scaffolds were ordered.

# Sammendrag

Takket være den svært raske utviklingen av DNA sekvenseringsteknologi de siste årene har flere store plantegenom blitt sekvensert. Det internasjonale hvetgenomsekvenseringskonsortiet (IWGSC) publiserte nylig den første versjonen av genomsekvensen til brødhvete. Arbeidet ble utført ved å isolere kromosomarmene fra brødhvete ved hjelp av flow-cytometrisk sortering, etterfulgt av såkalt paired-end sekvensering med Illuminateknologi. Resultatet, som refereres til som "survey sekvensen (CSS) til brødhvete består av tusenvis av såkalte sekvenskontiger med N50 mindre enn 4 Kb. Dette betyr at sekvensen er høyst fragmentert, noe som reduserer informasjonsinnholdet til sekvensen. Fragmenteringen skyldes at genomsekvensen inneholder mer enn 80% repeterte sekvenser. Arbeidet er en del av den norske deltakelsen i IWGSC og beskriver effekten av å integrere mate pair sekvensdata for å forbedre kromosom 7B CSS sekvensen og forankringen av 7BL BAC fysiske kontiger til det genetiske kartet for 7BL.

I publikasjon I undersøker vi for første gang effekten av å integrere mate-pair sekvensdata for flow-sorterte kromosomarmer for å redusere fragmenteringen til CSS sekvensen. Vi benytter tre mate-pair bibliotek med 2kb, 3kb og 5kb klonede fragmenter og med en dekningsgrad i sekvenseringen på henholdsvis 89 og 64 ganger for den korte og lange armen av kromosome 7B. Forengelse (eng. scaffolding) av de sammenhengene sekvensene vha computerprogrammet SSPACE viste en moderat forbedring av sekvenskvaliteten. Vi konkluderer med at årsaken er at det ble benyttet DNA som etter isolering vha flow-sortering ble amplifisert vha PCR I den såkalte "multiple displacement amplifiserings" reaksjonen (MDA) some er kjent for å gi kimære DNA molekyler, et fenomen some er kjent for å redusere nytten av mate-pair data.

I publikasjon II beskriver vi det første høytetthets-delesjon-binkartet for brødhvetekromosom 7B. Kartet ble generert ved hjelp av Nimbelgens høytetthets comparative genom hybridiserings array (CGH). Basert på den nylig publiserte CSS sekvensen fra IWGSC for A, B and D subgenomenen til brødhvete (IWGSC data deponi ved http://wheat-urgi.versailles.inra.fr/) identifiserte vi 7B spesifikke prober for omlag 8% av kromosom 7B sekvensen fordelt på 9 delesjonsområder. Vår undersøkelse korrigerte også den tidligere feil estimerte lengden og delesjonstypen for Del7BL-3, Del7BL-13 og Del7BL-5.

I manusckript III presenterer vi det første fysiske kartet for den lange armen av brødhvete kromosom 7B. For å oppnå dette benyttet vi en tredelt strategi; først delesjonskartlegging, genetisk kartlegging og tilsist syntenibasert kartlegging ved hjelp av data fra de nært beslektete artene *Brachypodium*, ris og sorghum. I alt ble 109 av de totalt 125 7BL fysiske kontigene for denne kromosomarmen tildelt en kartposisjon. Blant disse ble 92 forlengede kontiger, som tilsammen representerer 95% av den samlede lengden av den produserte sekvensen for 7BL sorter i riktig rekkefølge.

# Introduction

Wheat is the most widely cultivated cereal worldwide, being grown in temperate climates from Scandinavia and Russia in the north to Argentina in the south, including upland regions in the tropics [1]. It is one of the oldest domesticated plants and it is proposed that its first cultivation occurred about 10,000 years ago with the transition from hunter-gatherer to agricultural societies. The genome of *Triticum aestivum* (genome AABBDD) arose from at least one homoploid- and two polyploid hybridization events. According to recent findings, a homoploid hybridization between species of the A and the B lineages 5-6 million years ago (Mya) gave rise to the wheat D-genome lineage [2]. The second hybridization is estimated to have occurred approximately 500,000 years ago between the two grass species *Triticum urartu* (the A genome donor), and *Aegilops speltoides* (the B genome donor) giving rise to tetraploid emmer wheat (*T. turgidum*; AABB). The third hybridization is believed to have occurred approximately 10,000 years ago between cultivated tetraploid wheat and the wild grass *Ae. taushii* (D genome) to form modern hexaploid bread wheat (AABBDD) [3-5] (Figure 1). The two last hybridizations were followed by chromosome doubling in the new hybrid, enabling normal bivalent formation at meiosis and thus the production of fertile plants.

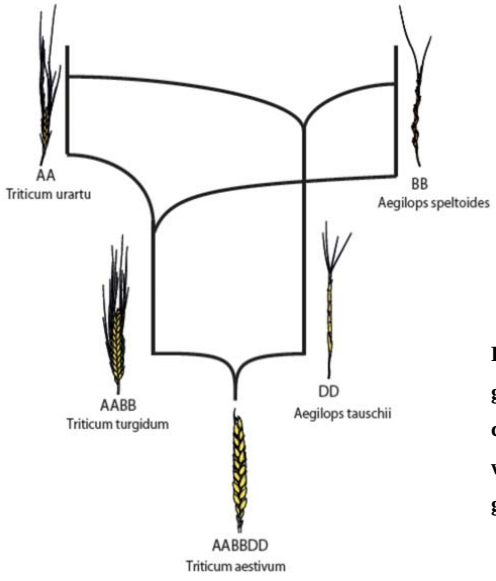


**Figure 1. The evolutionary and genome relationships between cultivated bread, durum wheats and related wild diploid grasses.** The figure is from [3].

Currently, about 95% of the wheat grown worldwide is hexaploid bread wheat mostly used for bread making with the remaining 5% being tetraploid wheat used for pasta making [5]. Wheat grain is rich in protein, minerals and vitamins and accounts for more than 20% of total calories in the human diet. With the global population expected to reach 9.6 million by 2050, wheat breeders, researches and growers need to increase wheat production by 70% to meet future demand [6, 7]. One important tool for breeders to be able to meet production demands is the deployment of molecular breeding methods that allow for faster development of higher yielding and better adapted varieties. Having a physically ordered genome sequence allows the development of molecular markers for marker assisted selection (MAS) and precision breeding. However, despite the fact that wheat has high socio-economic impact, bread wheat is one of the last major crops lacking a high-quality reference genome sequence.

The reason we still lack a genome reference sequence for wheat is that the wheat genome was long considered impossible to sequence due to the large genome size (17Gbp), extreme repeat content (>80% of TE in the genome), and polyploid nature. However, in parallel with revolution in sequencing technology, a number of initiatives such as the International Wheat Genome Sequencing Consortium (IWGSC), The UK wheat consortium (www.wheatisp.org) and the European *TriticeaeGenome FP7* (http://www.triticeaegenome.eu/ ) project were established to develop genomic resources and knowledge to provide foundation for sequencing and physical mapping of the wheat genome.

The IWGSC is an international public-private initiative that was established with the aim to sequence the wheat genome for accelerating wheat improvement (www.wheatgenome.org ). The participating countries of IWGSC include Norway, UK, France, Germany, Italy, Switzerland, Czech Republic, Estonia, Russia, India, China, Japan, Australia, Israel and the United States. Norway is taking responsibility for sequencing and mapping chromosome 7B, the second largest chromosome in wheat after 3B [8]. The work presented in this thesis is part of the Norwegian project to sequence chromosome 7B led by prof. Odd-Arne Olsen.

The complete genome sequence will provide a gene catalogue and be an essential step in understanding the biology of this important crop. Moreover, the availability of a reference genome is expected to allow for discovery of new genes and regulatory sequences and will serve as a foundation for marker development to facilitate trait mapping and make marker-assisted selection in wheat more feasible [9].

## Present status of wheat genomic resources

The allohexaploid wheat genome is one of the largest among crop species, 110 and 40 times of *Arabidopsis* and rice, respectively. Despite its hexaploid nature with three sets of 7 chromosomes, bread wheat behaves as a diploid, undergoing bivalent chromosome pairing during meiosis. In the past years, the availability of wheat genomics data in the public databases has grown rapidly. A significant insight into the transcribed portion of genome was obtained through large-scale sequencing of expressed sequence tags (ESTs). Until recently the main genomic resources for wheat to use were 1,268,372 ESTs (http://www.ncbi.nlm.nih.gov/genbank/dbest/dbest_summary/), ~57,000 unigenes (http://www.ncbi.nlm.nih.gov/unigene/statistics/ ), and 17,000 full-length cDNA sequences (http://trifldb.psc.riken.jp) [10]. A set of 16,000 ESTs were also mapped to chromosome specific bins, providing knowledge on the distribution of genes among sub-genomes and genes along the chromosomes. These genomic resources were essential for studies of individual genes, expression analysis, microarray designs [11, 12], and were utilized intensely for marker development [11].

## Sequencing of bread wheat and its diploid relatives

More recently, the bread wheat genome was shotgun sequenced to a 5-fold coverage using Roche 454 technology [13]. In order to assemble these shotgun reads, local assemblies were carried out on similar reads that formed clusters based on sequence similarity to orthologs in Brachypodium (*Brachypodium distachyon*), sorghum (*Sorghum bicolor*), barley (*Hordeum vulgare L.*) and rice (*Oryza sativa*). This assembly had a N50<1Kbp and represented ~22% of the wheat genome. In order to identify the subgenome origin of assemblies, sequences were classified using machine learning algorithms based on their similarity to the genomes sequence of the D genome donor species *Ae. taushii*, A genome relative *Triticum monococcum* and cDNA sequences of the B genome progenitor *Ae. speltoides*. Comparative analysis with these diploid relative genomes and other sequenced grass genomes allowed Brenchley and colleagues to identify around 96,000 genes with two-third of them assigned to the three subgenomes (A, B and D) of hexaploid wheat [13]. Soon after this publication two papers were published in the same issue of Nature, presenting draft genome sequences and analysis of two wheat diploid relatives, *Triticum urartu* [14] and *Ae. tauschii* [15]. The draft for *T. urartu* predicted 34,879 protein-coding genes, while the *Ae. tauschii* genome

was estimated to contain 43,150 protein-coding genes. The genome data of *Ae. tauschii* predicted the presence of genes encoding 159 previously unknown microRNAs, some of which may contribute to the ability of bread wheat to grow in low-nutrients soil [15]. Other findings related to sequencing of the diploid relatives of the bread wheat A and the D genomes are the identification of unique disease resistance genes. The A genome identified 593 R proteins versus just 197 in *Brachypodium* and 460 in rice; and there are twice as many R gene analogues in the D genome as in rice and six times as many as in maize [14, 15].

For the first time, the complexity of hexaploid bread wheat genome was reduced to individual chromosomes and subsequently sequenced by IWGSC initiative [16]. In this initiative, using flow-cytometric sorting wheat chromosome/chromosome arms were isolated and then sequenced to a depth of between 30X and 241X with the Illumina technology platform. The sequence reads were assembled into so-called chromosome survey sequence (CSS) assemblies. In total, 124,201 gene loci were identified in CSS assemblies with higher number on the B subgenome (44,523; 35%) compared to the A and D subgenomes which contained 40,253 (33%) and 39,425 (32%), respectively. Noteworthy, the distribution at the chromosomal level didn't follow this pattern. Authors explained it by preexisting differences in the subgenomes prior to polyploidization. The study didn't reveal any pronounced bias in gene content, structure, or composition between the different wheat subgenomes. Also no evidence for transcriptional dominance of an individual subgenome was observed. Using a combination of high density wheat SNP mapping and synteny to sequenced grass genome more than 75,000 genes were positioned along wheat chromosomes.

The Brenchley et al. [13], Ling et al. [14], Jia et al. [15] and IWGSC [16] publications represented the first attempts to sequence and produce a draft version of the bread wheat genome and its progenitors, providing a framework for identifying genes, developing molecular markers and further genome analysis. The IWGSC work not only detected and described a large proportion of the gene complement of bread wheat but also provided their chromosomal assignment. This serves a first major milestone in facilitating the isolation of genes controlling agronomically important traits. However, these studies also clearly demonstrated that whole genome shotgun sequencing of bread wheat genome is not sufficient to produce assemblies with significant level of contiguity (N50<100Kb). Even when the sequencing was performed for individual chromosomes, assemblies were very

fragmented with N50 less than 4Kb. A major challenge in applying whole genome/whole chromosome shotgun sequencing to large and complex plant genomes such as wheat is highly repetitive structure of the genome. In wheat, transposable elements can range in size from 100bp to several hundred kilobases [17]. Another complication comes with the difficulties of whole genome/whole chromosome shotgun sequencing to resolve highly similar families of paralogous and/or homeologous genes. Both of these situations can lead to underrepresentation of gene space due to collapse of duplicated regions in the assembly. For example, it was shown that *de-novo* assemblies of the human genome were 16.2% shorter than the reference genome because sequences with identity exceeding 85% resulted in sequence collapse [18]. Moreover, only 57% of the genes had sufficient representation in the assembly, while over 2300 exons were completely absent in the study by Li et al. [19]. In general, the level of fragmentation and miss-assembly can be very high and lead to the difficulties of genome analysis.

## BAC-by-BAC sequencing: map first, sequence later.

While future long-read single molecule sequencing technologies may enable reconstruction of large and complex genomes using only whole genome shotgun sequencing [20], the presently only realistic approach to obtain a complete reference genome sequence of bread wheat is a physical map based sequencing strategy. For complex genomes, physical maps constructed based on restriction fragment fingerprints of BAC clones are fairly robust because even in the presence of interspersed repeat sequences along the BAC inserts (typically 100-220Kb long) a unique restriction pattern is generated. The technologies for physical map construction include SNaPshot [21], whole-genome profiling [22, 23], optical mapping [24, 25] and genome mapping [26]. SNaPshot is a restriction fingerprinting method which uses restriction digestion of the DNA from individual BAC clones by cutting with multiple restriction endonucleases and sizing of the fragments with capillary electrophoresis [27]. Based on the pattern of restriction fragment overlaps minimum tiling path (MTP) which represents a set of BACs that cover entire chromosome with a minimum overlap is identified. Next, the BACs in the MTP are sequenced in pools or BAC-by-BAC to reduce the complexity of the assembly of BAC-sequences. The first plant genome to be fully sequenced using the BAC-by-BAC method was *Arabidopsis thaliana* [28]. The same strategy was later applied to rice, poplar and maize [29-31]. IWGSC has also chosen a BAC-by-BAC strategy for sequencing bread wheat genome. One key challenge with this

strategy when applied to the entire wheat genome is its polypoid structure. Due to low divergence between homoeologous chromosomes, regions from homoeologous chromosomes will have too many restriction fragments in common and will be assembled into single chimeric contigs. Instead, technological advances in flow sorting of chromosomes and the availability of individual chromosome and chromosome arm genetic stocks was used to reduce the complexity of the hexaploid genome allowing production of physical maps of individual wheat chromosomes/chromosome arms [32]. In order to sort individual chromosomes using this method, mitotic chromosomes are stained with DNA fluorochrome and introduced to a flow chamber which results in distribution of fluorescence signal intensity ("the flow karyotype") with each chromosome ideally recognized by individual peak. One complication initially was that the bread wheat flow karyotype only clearly separated chromosome 3B from the remaining 20 chromosomes forming three composite peaks (Figure 2A).

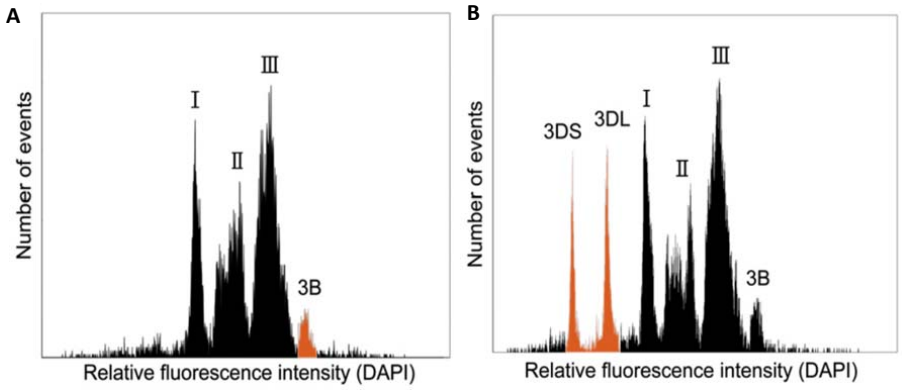


**Figure 2. Flow karyotyping in bread wheat.** A) The wheat cv. CS (2n=6x=42) flow karyotype consists of one single chromosome peak (3B) and three composite peaks consisting of the remaining 20 chromosomes (peaks I-III). B) The double ditelosomic line dDt3D carries the two arms of chromosome 3D in the form of two distinct telosomes, each of which is smaller than any of the 20 entire wheat chromosome, forming discrete, sortable peaks. The figure is reproduced with permission from [33].

## Physical map construction using FPC and LTC software packages.

The assembly of BAC fingerprints into physical contigs for large and repeat-rich genomes is a complex task. One of the standard programs for creating contigs from fingerprinted clones is FPC (FingerPrintedContigs) [34, 35], applied for example to the *Brachypodium* [36], rice [37] and *sorghum* genomes [38]. FPC groups related clones into contigs by using a pair-list algorithm to compare all fingerprints within a database to each other and calculate the coincidence score (Sulston score; i.e. probability that the number of shared bands is a coincidence). Two clones are potentially physically overlapping if their coincidence score is below a given threshold [35]. FPC starts the assembly process using stringent Sulston score cutoff and relaxes stringency to elongate fingerprint contigs. Although many physical maps have been build using FPC algorithm, quite a lot of errors have been identified in such contig assemblies [39]. Furthermore, the application of FPC is even more limited when dealing with complex genomes such as wheat and barley. Due to the high level of repetitive DNA in wheat genome, the criteria FPC uses for BAC contig assembly often result in short and unreliable assemblies. In addition, the presence of repetitive and poorly fingerprinted "questionable" clones (Q-clones) can lead to false overlaps and thus wrongly assembled contigs [40].

Recently, a software package called LTC (Linear Topological Contig) was developed to reduce the rate of false overlaps between BAC clones using new cutoff calculation method [40]. The main improvement of the LTC algorithm over FPC is that LTC initiates clustering using a liberal cutoff (opposite of FPC strategy) and then iteratively increases stringency until fingerprint contigs take on a linear structure (Figure 3). In LTC, "non-linear" clusters are split into sub-clusters with linear topological structure. LTC has been shown to outperform FPC by building contigs that are longer with more reliable ordering, and being more robust to errors caused by false and missing bands, therefore leading to more reliable MTP [41-43].

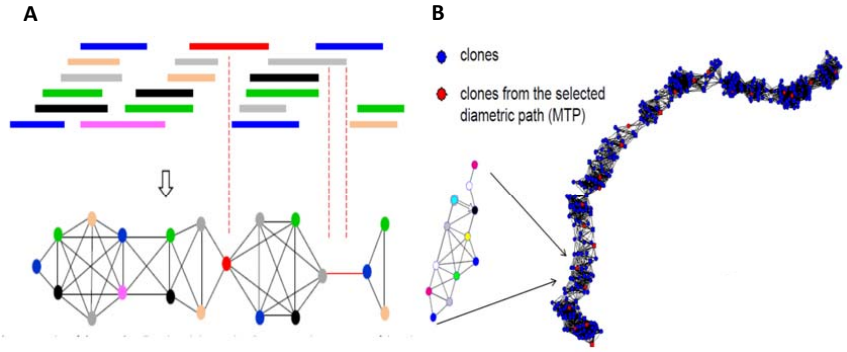


**Figure 3.** A) Top: physical clone overlaps. Bottom: network representation of clones (nodes) and clone overlaps (edges). Colors are used to show correspondence between physical and network representations of clone overlaps. Weak connections caused by low coverage are marked in red. The figure is taken from [43]. B) An example of the network representation of significant clone overlaps of BAC contig. Vertices represent the clones, edges represent the highly significant overlaps. Figure is taken from [44].

## Moving from multiple physical contigs to pseudochromosome.

Although BACs are ordered and oriented within BAC contigs, the order of the BAC contigs themselves remains unresolved. Thus, once physical maps have been assembled it is essential to integrate physical contigs with the genetic maps to determine the order and orientation of the BAC contigs to reconstruct the chromosome sequence. In wheat, genetic mapping is problematic as a result of low recombination rates in the centromeric and pericentromeric regions, which can span up to 50% of the chromosome length [45]. Studies of recombination rates for wheat chromosome 3B showed a recombination-rate gradient with the highest recombination in distal subtelomeric chromosome regions [45].

To facilitate construction of high-density genetic maps in wheat, many efforts have been undertaken. The first markers used for genetic mapping were based on differences in restriction fragment polymorphism (RFLP) [46]. Later, PCR-based markers such as RAPDs based on polymorphism in primer binding sites [47], AFLP markers based on polymorphism in restriction endonuclease recognition sites [48] and SSR markers which represent microsatellites [49] were used for genetic mapping. More recently, SNP markers became the

markers of choice because they are abundant and amenable to high-throughput genotyping. With advances in next-generation sequencing, high-throughput identification and genotyping of SNP markers have progressed to a significant degree in wheat [50-53]. The first high-density 9,000 SNP Infinium assay was developed by an USA/Australia collaborative project and was applied for genotyping of a diverse set of tetraploid and hexaploid wheats. More recently, a genotyping array containing about 90,000 gene-associated SNPs discovered using transcriptome data from 19 accessions of hexaploid and 18 accessions of tetraploid wheat was developed [52]. A total of 46,977 SNPs from the wheat 90K array were genetically mapped using a combination of eight mapping populations [52]. Genotyping of SNPs in polyploid wheat by hybridization methods is complicated by the presence of homoeologous and paralogous copies of genes because probes can hybridize not only to target locus, but also to it homoeologous and/or paralogues gene copies.

## Approaches complementing recombination mapping in wheat.

The limited resolution of genetic approaches to mapping of the wheat genome is increasingly being complemented by other mapping approaches. The polyploid nature of wheat and its tolerance to various forms of aneuploidy have been exploited for developing wheat cytogenetic stocks, including monosomic [54], nullisomic-tetrasomic [55], ditelosomic [56] and deletion lines [57]. Wheat deletion stocks were generated by monosomic addition of a gametocidal chromosome (*Aegilops cylindrica*) to Chinese Spring. The presence of this *Aegilops cylindrica* chromosome induces single chromosomal breaks in gametes that lack the alien chromosome followed by the concomitant loss of the segment distal to the breakpoint. Based on a set of more than 400 deletion lines, the wheat genome was subdivided into 159 chromosome bins of approximately 40Mb [58]. Deletion stocks have been extensively used for molecular mapping in Chinese Spring, providing information on the physical positions of genes and markers to specific chromosome arms and chromosomal bins [58-61]. However, due to the relatively large deletion sizes their application to high-resolution mapping is limited since the loci within each bin cannot be ordered.

An alternative to recombination based approaches, radiation-hybrid mapping (RH) has been used successfully to develop integrated physical maps in animals [62-64]. The radiation hybrid method does not depend on the meiotic recombination rates, but rather on co-

retention of markers in radiation induced deletions of chromosomes to order and determine the physical distance between markers [65]. The advantage of this method is that fewer lines (i.e. individuals) can be used to generate high-resolution maps and that genotyping is based on the presence-absence polymorphism with no need for polymorphic markers. RH mapping was first performed on the human X chromosome [62], and have later been used for mapping animal genomes such as zebra fish [63] and the porcine genome [64]. The potential of RH mapping for high-resolution mapping in plants has been shown in wheat studies including chromosome 1D [66], 3B [67] and D-genome [68]. In plants, radiation hybrid panels are developed by seed and pollen irradiation [69] and through *in vitro* procedures [70]. In seed irradiation, donor seeds are irradiated and plants germinated from these seeds are crossed with recipient plants, while in pollen irradiation, the recipient plant is pollinated with irradiated pollen of a donor plant. After irradiation, the ends of the induced chromosome breaks are assumed to rejoin by homology directed repair or non-homologous end joining which may result in the loss of DNA fragments of different size [71]. The first RH panel in plants was produced for maize chromosome 9 in an oat monosomic addition line and characterized with 21 maize chromosome specific markers [72]. The estimates of mapping resolution for this panel were at the 0.5- to 1.0-Mb level. Later, RH panels were produced in wheat with estimates of mapping resolution of ~199Kb and ~140Kb for 1D and wheat D-genome, respectively [66, 68]. It is noteworthy that although the theoretical resolution in these panels is extremely high, no reports have demonstrated the use of RH mapping to order and orient BAC contigs at a chromosome scale in wheat.

Finally, by taking advantage of evolutionary conserved gene order (collinearity) between grass genomes, genomics studies in the *Triticeae* have shown that comparative genomics approaches can be of use in the process of sequence contig ordering. To date, five *Triticeae* genomes have been sequenced, namely rice [31], *Brachypodium* [73], *Sorghum* [38], maize [30] and foxtail millet [74]. A synteny driven approach, the so-called GenomeZipper, where virtual gene order in a genome is created based on the syntenic information from the sequenced model grasses has been applied on the barley genome [75, 76], wheat chromosomes 4A [77] and 1BL [78], rye chromosomes [79] and *Lolium* [80]. Although synteny-based mapping approaches can be powerful, inversions and translocations of genes and gene blocks in wheat relative to other grass genomes are common [81].

## Reference genomes: gaps and errors.

The cost of generating a high-quality genome sequence is a major consideration when deciding on a sequencing strategy. Even though many basic questions can be answered using a low-cost whole-genome sequencing (WGS) assembly, a high-quality reference genome sequence is essential for understanding and correctly interpreting the biology of an organism. Fragmented genome sequences with high error rates may not be effectively used as it does not provide complete and reliable information and conclusions based on such assemblies can be incorrect. Genome errors such as erroneous nucleotide substitutions, insertions or deletions or larger-scale translocations may mislead genome annotations and analyses [39, 82, 83].

Although highly desirable, a perfect reference genome for an organism is difficult and costly to obtain. For example, the human reference genome sequence has the highest quality of all the mammalian genome sequences, but still contains many errors and gaps. It was shown that the "finished" assembly contained over 300 gaps in euchromatic portion of the genome, tilling path errors and regions represented by uncommon alleles. Some structurally complex regions were not resolved within the human assemblies until large insert clones were recovered and completely sequenced [84]. In general, whole genome shotgun assemblies are more prone to errors than genome assemblies obtained from BAC-by-BAC sequencing. Many large and long-range mis-assemblies were detected in the WGS sequence of rice, including missing sequences, spurious inversions, multiple assignment of identical sequence contigs (i.e. spurious duplication) and mis-assignment of sequence contigs (i.e. spurious translocation) [85]. In contrast, the map-based BAC-by-BAC strategy reduces the complexity of the assembly process by portioning the genome into smaller pieces. In this situation assembly errors are likely to be localized to individual sequenced BAC clones or merges between BAC clones [86].

## Approaches of anchoring physical map to molecular maps.

Physical maps can be anchored to molecular maps through different methods including experimental anchoring (when sequence of BACs is unknown) or by computational *in silico* anchoring (when the BAC sequence is known). *In silico* anchoring refers to the homology searches of BAC sequences against the marker sequences on the genetic maps to define the chromosomal position of the BAC. Experimental anchoring includes library screening of

BAC libraries (or BAC pools) with molecular markers by PCR-based or/and hybridization-based approaches. For example, integration of physical map of 1B was performed by hybridization of three dimensional MTP BAC DNA pools to the Nimblegen 40K array, containing 39,179 wheat NCBI UniGenes [43], while for 3B anchoring of the physical map was performed by screening three-dimensional BAC pools with PCR-based markers [67]. All the anchoring of MTP BAC contigs for chromosome 7B was performed *in silico* using BLASTN homology searches against selected marker sequences.

## The value of the finished bread wheat genome.

A high quality reference sequence is not relatively complete but provides as complete as possible access to gene models of a genome, the regulatory elements that control their function and a framework for understanding genomic variation. For breeders, access to a genome sequence allows high resolution identification of existing genetic variation as well as the monitoring of this variation in breeding programs. Additional benefits include direct access to all gene content, predicted gene function and mapping information. Knowledge of promoter sequences carries the possibility to monitor epigenetic status of genes and gene expression level monitoring using RNAseq or microarrays. Also, with decreasing cost of sequencing, re-sequencing to access genetic information is also becoming feasible. For breeders, approaches such as genome wide association studies (GWAS) [87], marker-assisted selection [88] and genomic selection [89] are becoming more realistic also for wheat in the not too distant future.

# Objectives of the thesis

## Principal objective

The current study is a part of the Norwegian participation in the International Wheat Genome Sequencing Consortium (IWGSC) which aims to sequence bread wheat genome. The principal objective of this study is to evaluate the sequencing methodology for wheat chromosome 7B and to produce an anchored physical map for this chromosome based on the sequenced BACs.

## The specific objectives are as follows:

1) To investigate whether integration of long range mate pair libraries improves the shotgun sequence assembly of wheat chromosome 7B
2) To produce a cytogenetic deletion bin map of chromosome 7B
3) To investigate the utility of radiation-hybrid population as tool for chromosome 7B mapping
4) To anchor 7BL physical map to genetic maps and molecular maps

# Results and discussion

## Integration of mate pair sequences to improve shotgun assemblies of flow-sorted chromosome arms of hexaploid wheat

Due to revolutionary advances in next-generation sequencing (NGS) technologies, whole-genome shotgun (WGS) strategies have become the methods of choice for sequencing of many organisms, as it allows sequencing of genome irrespective of its size within a short time and for relatively cheap price. Recently, using chromosome sorting and NGS technologies, NGS-based sequencing of the wheat chromosomes/chromosome arms was performed by the IWGSC initiative. The purpose of this work was to provide the first draft sequence of the bread wheat genome for each of the 21 chromosomes [16]. In this initiative isolated wheat chromosome/chromosome arms were sequenced to a depth of between 30X and 241X with the Illumina technology platform. The paired end sequence reads were assembled with the short-read assembly tool ABySS [90]. This resulted in generation of thousands of sequence contigs for each wheat chromosome with N50 less than 4Kb excluding contigs shorter than 200bp [16]. Rather than long contiguous sequences spanning large chromosome regions as obtained in vertebrate genome assemblies [91], the wheat assemblies of short *de novo* reads are highly fragmented. The main reason for the low assembly quality is the high complexity of the wheat genome harboring high amounts of repetitive elements with high sequence identity that during the assembly process collapse into single sequence. One approach permitting us to "jump" across repetitive DNA in order to link (scaffold) contigs for a more contiguous assembly is to use long fragment mate pair (MP) sequencing libraries [92]. In an ideal situation, the result of such scaffolding strategy is one to a few scaffolds per chromosome with gaps of correct length separating the contigs.

In **Paper I**, we investigate to what extent long fragment MP libraries improve wheat paired-end (PE) assemblies by scaffolding. To address this question we have used MP libraries of 2-, 3- and 5-Kb insert size from DNA produced by a multiple displacement amplification (MDA) reaction of flow-sorted chromosome arms of 7B. Several assemblies using different stringency parameters were performed with the SSPACE assembly scaffolding tool [93]. Our result show that addition of 2-, 3- and 5-Kb MP libraries produced from MDA DNA of flow-sorted wheat chromosome arms improved assembly statistics, but that the improvement was greatly dependent on scaffolding stringency. For example, the

assembly N50 was improved substantially at low stringency by 6-7.5-fold, while at the highest stringency, N50 was increased only by 1.3-1.8-fold compared to PE assemblies. Even at low stringency the observed improvement was lower than that reported for recently sequenced plant genomes including cucumber [94], cacao [94], watermelon [95] and bamboo [95]. It is also noteworthy that improvement in the assembly statistics was accompanied by increased assembly errors/reduced assembly correctness. Thus, when scaffolds achieved with different stringency criteria were compared with 50 sequenced random BAC clones from 7BL, a strong correlation between estimated scaffold reliability and scaffold assembly stringency was observed (**paper I** Fig.3B).

We hypothesize that the relatively modest improvement of assembly quality is related to the use of MDA DNA for our MP libraries, which is known to contain chimeric DNA molecules that significantly reduced usefulness of MP. Although the MP libraries had high sequencing coverage, the proportion of properly oriented read pairs was low (~ 40%). No evidence for non-wheat origin or other contamination of incorrectly oriented reads was found. However, although the performance of MP libraries was not ideal, mate pair reads successfully linked up exons from fragmented gene sequences and connected genes from different contigs (**paper I** Table 4).

In general, MP libraries could be a good alternative over traditional bacterial artificial chromosome sequencing because the libraries for sequencing can be produced by relatively simple procedures without the need for laborious cloning, colony picking, DNA clone isolation, etc. However, before investing money and time into sequencing and integration of MP libraries into assemblies of complex repeat-rich genomes, it is important to consider what degree of assembly improvement and quality may be expected for a given project.

From our work we conclude that the effect of 2-, 3-, 5-Kb MP short libraries is rather moderate. However, we do anticipate that for wheat assembly large insert MP sequencing can be more beneficial than short MP libraries. As shown for the rat genome [96] short-insert libraries (PE and 3Kb MP) were much less efficient in spanning long repetitive elements, such as LINEs or LTRs, than large insert MP libraries (>=15Kb). However it is also should be noted that the MDA DNA source is insufficient to provide DNA for long range “linking” libraries, as DNA fragments longer than 5Kb appear substantially underrepresented after MDA.

Subsequent to the publication of paper I, under the framework of the 7B IWGSC project, 10Kb and 20Kb insert fragment libraries were produced from pooled DNA of neighboring MTP BACs (7B IWGSC, unpublished), allowing us to perform a pilot experiment to assess the quality and impact of 10- and 20-Kb mate pair inclusion. The SSPACE k=5 assembly had a substantial N50 increase by 2.7-fold compared to the N50 of the assembly obtained after addition of the 2+3+5 Kb MP libraries. In contrast to 2-,3- and 5K libraries, where only ~1.5% of reads were used for scaffolding, the proportion of used reads was 72% and 59.4% for 10K and 20K libraries, respectively. Also worth mentioning is that the number of erroneously oriented MP reads was as low as 1.6% and 3.82% for the 10Kb and 20Kb insert libraries, respectively. After inclusion of all MP libraries, the largest scaffold was 8.4X larger than longest PE contig.

The findings of Paper I have important implication on how to direct and improve future wheat chromosome sequencing and assembly. Moreover, the availability of *de novo* scaffolds and contigs allowed us to generate large set of 7B chromosome specific markers to be further applied in 7B mapping studies. In the future, we can expect improvements in the production of mate pair libraries, both in terms of accuracy of the insert-size and of the suppression of errors. Among third generation technologies, in my opinion, the greatest potential lies with single molecule sequencing performed by e.g. PacBio RS, which can generate read length of 10kb and longer. However until now, the error rate of single-molecule reads is in the range of 10-18%, which has limited their application [97].

## Utilization of deletion bins to anchor and order sequences along the wheat 7B chromosome

Independent of the sequencing strategy used, it is still not possible to generate one continuous sequence per chromosome, especially for large and complex genomes such as wheat. After assembly of the sequencing data from e.g. individual BACs, the assembly of the chromosome is expected to consist of thousands of unordered contigs/scaffolds lacking chromosome positional information. To assemble a reference sequence where contigs/scaffolds are placed in order and the gaps between them are estimated, it is necessary to anchor sequences on high resolution genome maps. This is usually done by approaches based on genetic and physical mapping [98-101]. In wheat, suppression of recombination in the (peri) centromeric region impedes efforts to resolve the order of sequences using recombination based genetic maps. Several studies have shown that the

recombination frequency in wheat chromosomes is lowest in the centromeric region and reaches its maximum towards the distal end [45]. Such suppression of recombination limits genetic anchoring resolution for the (peri) centromeric area, leading to a high number of plysical contigs to be anchored to the same genetic position without an individual order. One approach to overcome the limitations imposed by meiotic mapping that was applied to the integration of BAC-based physical maps of wheat (e.g. 3B, 1BL, 1BS) is to combine meiotic mapping, deletion bin mapping, radiation hybrid mapping and mapping based on synteny with fully sequenced model grass genomes (like *Brachypodium*, rice and sorghum) [43, 67, 78].

In **paper II**, we developed a genotyping array containing 49,500 wheat chromosome 7B specific probes and used it to genotype cv. Chinese Spring (CS) 7B deletion stocks to anchor sequence contigs/scaffolds to deletion bins. In total, we genotyped eleven 7B deletion stocks of cv. Chinese Spring subdividing the chromosome into nine deletion bins. In total, 3,671 sequence contigs and scaffolds that are described in **paper I** were mapped to nine deletion bins of 7B. The bin map produced in **paper II** is the highest density deletion bin map for any wheat chromosome so far, providing 100 times more bin-mapped 7B sequences compared to the previous study by Hossain et al. [102]. Our 7B deletion bin map allowed us to perform *in silico* anchoring of 7B BAC sequences to deletion bins and significantly contributed to the successful anchoring of the 7BL physical map (**paper III**).

Genotyping of polyploid wheat with hybridization based methods is complicated by the highly similar DNA sequences of homoeologous or paralogues due to the low specificity of oligonucleotide probe hybridization. Hence, to maximize probe specificity we used the chromosome survey sequences of 7A and 7D chromosomes (IWGSC data repository at http://wheat-urgi.versailles.inra.fr/) to remove probe sequences with high sequence similarity across sub-genomes. Our strategy to call presence/absence genotype variation in the CS deletion stocks relied on the model-based clustering method (Mclust) which separates classes of “absent” and “present” probes. We show in the paper that by using hybridization signals from multiple probes located on the same sequence scaffold or contig leads to more robust and simple presence/absence calling in polyploid wheat, i.e better separation of “present” and “absent” clusters (**paper II** Fig.2). For example, frequencies of correct assignment of scaffolds/contigs to 7B chromosome arms were higher when we used three (99.97%) compared to two probes (99.1%) per estimated log2 ratio of signal intensities of deletion lines relative to wild type. Validation of our bin mapped results

suggested a high accuracy of the assignment of 7B contigs and scaffolds to 7B deletion bins (error rate of <2.5%). We estimated the gene density along 7B, the highest density was found to be in the distal regions of the chromosome with a lower gene density in the centromeric compartments. These findings are consistent with previous studies that wheat genes occur more frequently in distal parts of the chromosomes [43, 78]. In addition, our study confirmed and further delineated the former mis-estimation of deletion length and deletion type in Del7BL-3, Del7BL-13 and Del7BL-5 deletion stocks.

## The utility of radiation-hybrid population as tool for chromosome 7B mapping

During my PhD work I have attempted to utilize radiation hybrid population for mapping 7B chromosome. One seed (~1100 plants) and one pollen (~60 plants) radiation hybrid panel were produced for wheat chromosome 7B by A.Kumar and S.Kianian (NDSU, USA). The parent plants used for the radiation hybrid panel were the tetraploid wheat cultivar Langdon (LDN; AABB; 2n=4x=28:13"+7B") and Langdon chromosome substitution line (LDN 7D(7B); 2n=4x=28:13"+7D") in which the 7B chromosomes are substituted with 7D chromosomes. To generate the seed panel, the plants that germinated from irradiated LDN seeds were crossed with the LDN 7D(7B) plants in which a pair of chromosome 7B is substituted by a pair of 7D chromosomes of the hexaploid cultivar Chinese Spring (Fig.4).

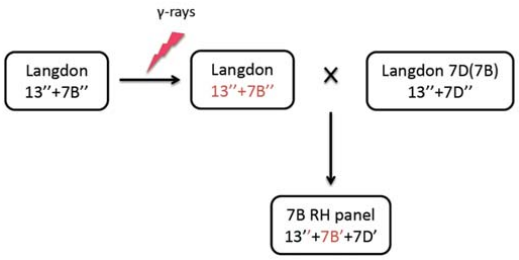


**Figure 4. Schematic presentation of RH seed panel development.** Langdon seeds were irradiated, germinated and viable plants crossed with the Langdon substitution line to yield the RH progeny. Red color represents potentially fragmented chromosomes and the symbol (') indicates the chromosome copy number.

To generate the pollen panel, LDN plants were grown to flowering and dehiscent wheat spikes were excised from the plant with stems kept in water. Entire spikes were irradiated

with ϒ-rays and pollen from the irradiated spikes was immediately used to pollinate emasculated spikes of LDN 7D(7B). F1 hybrid seeds were harvested and planted.

In order to investigate whether RH plants carried deletions, we genotyped a subset of 259 seed RH lines and 53 pollen RH lines with 30 and 31 7B specific SNPs using the Sequenom assay (data not shown). The average deletion frequency was 0.27% (range 0-10%) for seed and 9.07% (range 0-48.4%) pollen panels (Fig. 5). The percentage of plants retaining all tested markers was 95% and 62% for seed and pollen radiation hybrids, respectively. A subset of 96 RH plants (from SNP genotyping plus random lines) including tetraploid LDN and 7D(7B) substitution lines, as well as ditelosomic 7B lines were subjected to DaRTs genotyping. Based on the genotypes of LDN and 7D(7B) (1 vs. 0), among 37,770 DaRTs, 1619 7B specific markers were selected. The average deletion frequency for RH plants from seed panel and pollen panel was 3.4% and 15.5%, respectively (Fig.5). Based on a cutoff value of 15% for false discovery rate ($d_{fdr}$) , a set of 1,619 markers was subdivided into clusters of putatively linked markers (by single linkage method using likelihood-based metric). Two large clusters, cl_1 and cl_2 were obtained with 782 and 778 markers respectively; one cluster cl_3 with 3 markers and 58 clusters with a single marker. Recalculation of $d_{fdr}$ within cl_l resulted in only non-significant linkages. This result can be explained by the observation that marker genotypes within this cluster were very similar (i.e. little diversity in deletion type and size). Markers from cl_1 belong to 7BL BACs. The network of marker linkages for cl_2 consisted of two parts connected via only 3 markers. Markers from the first part (573 markers) originated from 7BS while the second part (200 markers) had markers mapped to both 7BS and 7BL (most likely centromeric region). Unfortunately, ordering of markers within these clusters was difficult because of the poor linkage resolution resulting in a complex cluster-topology with non-linear structure (Fig.6). Obtained RH maps had low resolution with many physical scaffolds mapped to the same locus. Since RH mapping was not informative, we have not included this data in building the integrated 7B physical map.

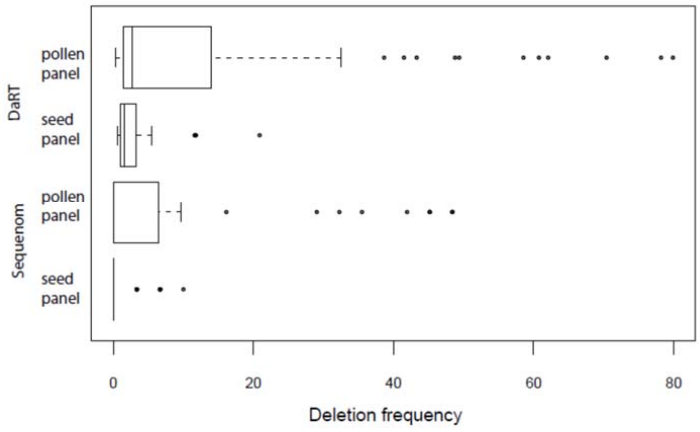


**Figure 5. Deletion frequencies of RH lines from seed and pollen panels based on DaRT and Sequenom genotyping.**

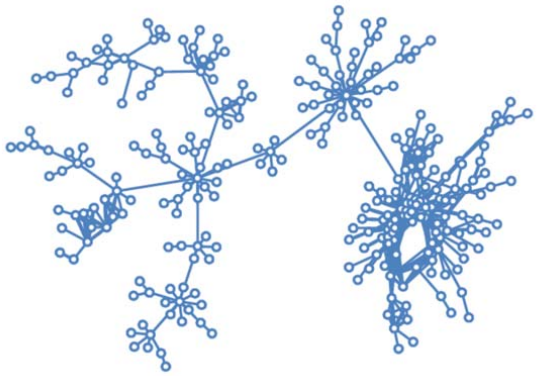

**Figure 6. A network representation of RH map for 7BL cluster.** Vertices correspond to markers, edges reflect RH distances.

The RH mapping method depends on the size of the chromosomal deletions that are present in the mapping population. In order to develop a high resolution RH map high level of chromosome fragmentation, homogeneous breakage along the chromosome and different size of deletions are desirable. The results from our study show that the deletion frequency for 7B RH seed panel was lower than reported for D-genome chromosomes, 0.27%

compared to 2.1%. However, the deletion frequency achieved in our 7B pollen material was in the same range as reported for D-genome pollen panels (range 7.4-32%). In the previous studies of Kumar et. al [103] a few wheat 3B BAC contigs could be placed and positioned relative to each other on a RH map. It was anticipated that a subset of ~100 informative RH plants is sufficient to produce a RH map of the single chromosome. However no published studies on RH mapping of many physical contigs do exist in wheat. Our study shows that it is very challenging, if not impossible to obtain a homogeneous mapping of BAC contigs from representation of ~100 RH lines. In ideal situation of high density genotyping linkages between markers should be supported by linkages of other adjacent markers. One would expect a linear network structure of the marker connections. That was not the case in our study. When deletions are small and non-interconnected mapping becomes very challenging. Therefore, large overlapping deletions, which interconnect smaller deletions, are necessary. In our study, the deletion frequency of pollen RH plants was high, however the diversity of deletion types for panel was very poor.

Based on our work I could conclude that for high quality RH mapping in wheat firstly large population size of both pollen and seed panel are needed. Secondly, extensive preliminary screening of RH lines to identify plants with large and diverse deletions has to be performed. This requires markers with known positions spanning the entire chromosome and screening of possibly thousands of plants. Thirdly, an efficient and cost-effective genotyping method with low error rate to call presence/absence is necessary.

## Ordering and orienting physical contigs along bread wheat chromosome 7B long arm

In **paper III**, we generated an anchored physical map of bread wheat chromosome 7B long arm using a combination of different mapping data. The physical map of 7BL contained 45,087 BACs assembled into 125 MTP long physical contigs which were BAC-by-BAC sequenced. The sequence assembly of MTP BACs resulted in 40,677 scaffolds covering ~97% of the 7BL estimated length. In our anchoring strategy we used three step strategy, including applying deletion bin mapping then genetic mapping and finally synteny-based mapping. In total, 105 7BL physical scaffolds were anchored to seven 7BL deletion bins spanning ~97% of the 7B sequence scaffolds. The integration of the 7B physical map with genetic map was accomplished on the basis of three crosses: an F8 population derived from the cross between Chinese Spring (Cs) and the French Cultivar Renan(Re) genotyped with

ultradense 420K SNP array, and two additional F6 genetic crosses of Sy*Naxos and SHA3/CBRD*Naxos genotyped with 90K SNP Illumina array. Ninety six physical contigs were genetically anchored with 96% of them ordered for a total span of ~95% of 7BL sequence scaffolds. The synteny based mapping using species of *Brachypodium*, rice and sorghum provided especially valuable information for regions with limited genetic resolution, i.e. centromeric regions. The comparison between the 7B genetic map and 7B *Triticeae* prototype map indicates high degree of collinearity, however rearrangements were also present (**paper III** Fig.4). In total, among 125 7BL physical contigs, 16 physical contigs covering ~9Mb or 1.7% of the 7B sequence scaffolds were not anchored due to lack of any sequence overlap or/and genetic position, synteny information or bin map information. This is quite a small fraction compared to other wheat chromosomes e.g. 1BL, 1BS and 3B where ~26%, ~22.6% and ~44% of the chromosome length remained without anchoring, respectively. This increase in anchoring efficiency is explained by the completely sequenced 7BL MTP. Even though large proportion of physical contigs was anchored, several improvements should be accomplished to fully anchor and orient the 7BL physical map. Firstly, the order and orientation of the physical contigs which were placed based on little evidence should be improved. Secondly, unanchored physical contigs should be integrated with genetic and molecular maps. This can be achieved by screening BAC pools, deletions stocks and genetic populations with markers designed from selected BAC contigs. Additionally methods of fluorescent in situ hybridization mapping [104], optical mapping [105] and genome mapping on nanochannel arrays [106] can provide an additional layer of mapping information in future studies.

# Conclusions

This study provides important insights for future sequencing and mapping projects on bread wheat and other complex genomes. In our study we have shown that although short insert size mate pair libraries assist in the assembly of sequences, the improvements in the quality of the assembly are small. In contrast, the use of large insert mate pair libraries (10 and 20 kb) has a major effect on the quality of the sequence assembly of wheat genomic DNA due to their ability to span long repetitive sequence elements.

Bread wheat chromosomes 7B is the first wheat chromosome for which BAC clones of the minimum tiling path have been fully individually sequenced. Using a combination of high-

density deletion bin mapping, genetic mapping and synteny-based mapping we have generated the first draft of an anchored physical map of the long arm of wheat chromosome 7B. An anchored physical map provides opportunities for gene isolation and facilitates direct linkage to traits used in the field and breeding. Future work will be focused on the improvement of contig ordering along 7BL by integration of other mapping data or applying additional anchoring strategies.

# References


1. Feldman M, Lupton FGH, Miller TE: **Wheats**. London: Longman Scientific; 1995.
2. Marcussen T, Sandve SR, Heier L, Spannagl M, Pfeifer M, IWGSC, Jakobsen KS, Wulff B, Steuernagel B, Mayer K *et al*: **Ancient Hybridizations Among the Ancestral Genomes of Bread Wheat**. *accepted, Science* 2014.
3. Dvorak J, Terlizzi P, Zhang HB, Resta P: **The evolution of polyploid wheats: identification of the A genome donor species**. *Genome / National Research Council Canada = Genome / Conseil national de recherches Canada* 1993, **36**(1):21-31.
4. Petersen G, Seberg O, Yde M, Berthelsen K: **Phylogenetic relationships of Triticum and Aegilops and evidence for the origin of the A, B, and D genomes of common wheat (Triticum aestivum)**. *Molecular phylogenetics and evolution* 2006, **39**(1):70-82.
5. Shewry PR: **Wheat**. *Journal of experimental botany* 2009, **60**(6):1537-1553.
6. Foley JA, Ramankutty N, Brauman KA, Cassidy ES, Gerber JS, Johnston M, Mueller ND, O'Connell C, Ray DK, West PC *et al*: **Solutions for a cultivated planet**. *Nature* 2011, **478**(7369):337-342.
7. Tilman D, Cassman KG, Matson PA, Naylor R, Polasky S: **Agricultural sustainability and intensive production practices**. *Nature* 2002, **418**(6898):671-677.
8. Safar J, Simkova H, Kubalakova M, Cihalikova J, Suchankova P, Bartos J, Dolezel J: **Development of chromosome-specific BAC resources for genomics of bread wheat**. *Cytogenetic and genome research* 2010, **129**(1-3):211-223.
9. Perez-de-Castro AM, Vilanova S, Canizares J, Pascual L, Blanca JM, Diez MJ, Prohens J, Pico B: **Application of genomic tools in plant breeding**. *Current genomics* 2012, **13**(3):179-195.
10. Mochida K, Yoshida T, Sakurai T, Ogihara Y, Shinozaki K: **TriFLDB: a database of clustered full-length coding sequences from Triticeae with applications to comparative grass genomics**. *Plant physiology* 2009, **150**(3):1135-1146.
11. Bernardo AN, Bradbury PJ, Ma H, Hu S, Bowden RL, Buckler ES, Bai G: **Discovery and mapping of single feature polymorphisms in wheat using Affymetrix arrays**. *BMC genomics* 2009, **10**:251.
12. Rustenholz C, Choulet F, Laugier C, Safar J, Simkova H, Dolezel J, Magni F, Scalabrin S, Cattonaro F, Vautrin S *et al*: **A 3,000-loci transcription map of chromosome 3B unravels the structural and functional features of gene islands in hexaploid wheat**. *Plant Physiol* 2011, **157**(4):1596-1608.
13. Brenchley R, Spannagl M, Pfeifer M, Barker GLA, D'Amore R, Allen AM, McKenzie N, Kramer M, Kerhornou A, Bolser D *et al*: **Analysis of the breadwheat genome using whole-genome shotgun sequencing**. *Nature* 2012, **491**(7426):705-710.
14. Ling HQ, Zhao SC, Liu DC, Wang JY, Sun H, Zhang C, Fan HJ, Li D, Dong LL, Tao Y *et al*: **Draft genome of the wheat A-genome progenitor Triticum urartu**. *Nature* 2013, **496**(7443):87-90.
15. Jia JZ, Zhao SC, Kong XY, Li YR, Zhao GY, He WM, Appels R, Pfeifer M, Tao Y, Zhang XY *et al*: **Aegilops tauschii draft genome sequence reveals a gene repertoire for wheat adaptation**. *Nature* 2013, **496**(7443):91-95.
16. Consortium IWGS: **A chromosome-based draft sequence of the hexaploid bread wheat genome**. *accepted, Science*.
17. Choulet F, Wicker T, Rustenholz C, Paux E, Salse J, Leroy P, Schlub S, Le Paslier MC, Magdelenat G, Gonthier C *et al*: **Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces**. *The Plant cell* 2010, **22**(6):1686-1701.
18. Alkan C, Sajjadian S, Eichler EE: **Limitations of next-generation genome sequence assembly**. *Nature methods* 2011, **8**(1):61-65.

19. Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K *et al*: **De novo assembly of human genomes with massively parallel short read sequencing**. *Genome research* 2010, **20**(2):265-272.
20. Huddleston J, Ranade S, Malig M, Antonacci F, Chaisson M, Hon L, Sudmant PH, Graves TA, Alkan C, Dennis MY *et al*: **Reconstructing complex regions of genomes using long-read sequencing technology**. *Genome research* 2014, **24**(4):688-696.
21. Luo MC, Thomas C, You FM, Hsiao J, Shu OY, Buell CR, Malandro M, McGuire PE, Anderson OD, Dvorak J: **High-throughput fingerprinting of bacterial artificial chromosomes using the SNaPshot labeling kit and sizing of restriction fragments by capillary electrophoresis**. *Genomics* 2003, **82**(3):378-389.
22. Philippe R, Choulet F, Paux E, van Oeveren J, Tang J, Wittenberg AH, Janssen A, van Eijk MJ, Stormo K, Alberti A *et al*: **Whole Genome Profiling provides a robust framework for physical mapping and sequencing in the highly complex and repetitive wheat genome**. *BMC genomics* 2012, **13**:47.
23. van Oeveren J, de Ruiter M, Jesse T, van der Poel H, Tang J, Yalcin F, Janssen A, Volpin H, Stormo KE, Bogden R *et al*: **Sequence-based physical mapping of complex genomes by whole genome profiling**. *Genome research* 2011, **21**(4):618-625.
24. Schwartz DC, Li X, Hernandez LI, Ramnarain SP, Huff EJ, Wang YK: **Ordered restriction maps of Saccharomyces cerevisiae chromosomes constructed by optical mapping**. *Science* 1993, **262**(5130):110-114.
25. Aston C, Mishra B, Schwartz DC: **Optical mapping and its potential for large-scale sequencing projects**. *Trends in biotechnology* 1999, **17**(7):297-302.
26. Lam ET, Hastie A, Lin C, Ehrlich D, Das SK, Austin MD, Deshpande P, Cao H, Nagarajan N, Xiao M *et al*: **Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly**. *Nature biotechnology* 2012, **30**(8):771-776.
27. Luo MC, Ma Y, You FM, Anderson OD, Kopecky D, Simkova H, Safar J, Dolezel J, Gill B, McGuire PE *et al*: **Feasibility of physical map construction from fingerprinted bacterial artificial chromosome libraries of polyploid plant species**. *BMC genomics* 2010, **11**:122.
28. Arabidopsis Genome I: **Analysis of the genome sequence of the flowering plant Arabidopsis thaliana**. *Nature* 2000, **408**(6814):796-815.
29. Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A *et al*: **The genome of black cottonwood, Populus trichocarpa (Torr. & Gray)**. *Science* 2006, **313**(5793):1596-1604.
30. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA *et al*: **The B73 maize genome: complexity, diversity, and dynamics**. *Science* 2009, **326**(5956):1112-1115.
31. International Rice Genome Sequencing P: **The map-based sequence of the rice genome**. *Nature* 2005, **436**(7052):793-800.
32. Vrana J, Kubalakova M, Simkova H, Cihalikova J, Lysak MA, Dolezel J: **Flow sorting of mitotic chromosomes in common wheat (Triticum aestivum L.)**. *Genetics* 2000, **156**(4):2033-2041.
33. Dolezel J, Vrana J, Capal P, Kubalakova M, Buresova V, Simkova H: **Advances in plant chromosome genomics**. *Biotechnology advances* 2014, **32**(1):122-136.
34. Soderlund C, Longden I, Mott R: **FPC: a system for building contigs from restriction fingerprinted clones**. *Computer applications in the biosciences : CABIOS* 1997, **13**(5):523-535.
35. Soderlund C, Humphray S, Dunham A, French L: **Contigs built with fingerprints, markers, and FPC V4.7**. *Genome research* 2000, **10**(11):1772-1787.
36. Gu YQ, Ma Y, Huo N, Vogel JP, You FM, Lazo GR, Nelson WM, Soderlund C, Dvorak J, Anderson OD *et al*: **A BAC-based physical map of Brachypodium distachyon and its comparative analysis with rice and wheat**. *BMC genomics* 2009, **10**:496.

37. Chen M, Presting G, Barbazuk WB, Goicoechea JL, Blackmon B, Fang G, Kim H, Frisch D, Yu Y, Sun S *et al*: **An integrated physical and genetic map of the rice genome**. *The Plant cell* 2002, **14**(3):537-545.
38. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A *et al*: **The Sorghum bicolor genome and the diversification of grasses**. *Nature* 2009, **457**(7229):551-556.
39. Salzberg SL, Yorke JA: **Beware of mis-assembled genomes**. *Bioinformatics* 2005, **21**(24):4320-4321.
40. Frenkel Z, Paux E, Mester D, Feuillet C, Korol A: **LTC: a novel algorithm to improve the efficiency of contig assembly for physical mapping in complex genomes**. *BMC bioinformatics* 2010, **11**:584.
41. Breen J, Wicker T, Shatalina M, Frenkel Z, Bertin I, Philippe R, Spielmeyer W, Simkova H, Safar J, Cattonaro F *et al*: **A physical map of the short arm of wheat chromosome 1A**. *PloS one* 2013, **8**(11):e80272.
42. Lucas SJ, Akpinar BA, Kantar M, Weinstein Z, Aydinoglu F, Safar J, Simkova H, Frenkel Z, Korol A, Magni F *et al*: **Physical mapping integrated with syntenic analysis to characterize the gene space of the long arm of wheat chromosome 1A**. *PloS one* 2013, **8**(4):e59542.
43. Raats D, Frenkel Z, Krugman T, Dodek I, Sela H, Simkova H, Magni F, Cattonaro F, Vautrin S, Berges H *et al*: **The physical map of wheat chromosome 1BS provides insights into its gene space organization and evolution**. *Genome biology* 2013, **14**(12):R138.
44. Zeev Frenkel EP, David Mester, Catherine Feuillet, Abraham Korol **Using LTC software to assemble physical maps in complex genomes such as wheat**. In: *PAG XX: 2012; San Diego*; 2012.
45. Saintenac C, Falque M, Martin OC, Paux E, Feuillet C, Sourdille P: **Detailed Recombination Studies Along Chromosome 3B Provide New Insights on Crossover Distribution in Wheat (Triticum aestivum L.)**. *Genetics* 2009, **181**(2):393-403.
46. Chao S, Sharp PJ, Worland AJ, Warham EJ, Koebner RM, Gale MD: **RFLP-based genetic maps of wheat homoeologous group 7 chromosomes**. *TAG Theoretical and applied genetics Theoretische und angewandte Genetik* 1989, **78**(4):495-504.
47. Williams JG, Kubelik AR, Livak KJ, Rafalski JA, Tingey SV: **DNA polymorphisms amplified by arbitrary primers are useful as genetic markers**. *Nucleic acids research* 1990, **18**(22):6531-6535.
48. Vos P, Hogers R, Bleeker M, Reijans M, van de Lee T, Hornes M, Frijters A, Pot J, Peleman J, Kuiper M *et al*: **AFLP: a new technique for DNA fingerprinting**. *Nucleic acids research* 1995, **23**(21):4407-4414.
49. Song QJ, Shi JR, Singh S, Fickus EW, Costa JM, Lewis J, Gill BS, Ward R, Cregan PB: **Development and mapping of microsatellite (SSR) markers in wheat**. *TAG Theoretical and applied genetics Theoretische und angewandte Genetik* 2005, **110**(3):550-560.
50. Akhunov E, Nicolet C, Dvorak J: **Single nucleotide polymorphism genotyping in polyploid wheat with the Illumina GoldenGate assay**. *TAG Theoretical and applied genetics Theoretische und angewandte Genetik* 2009, **119**(3):507-517.
51. van Poecke RM, Maccaferri M, Tang J, Truong HT, Janssen A, van Orsouw NJ, Salvi S, Sanguineti MC, Tuberosa R, van der Vossen EA: **Sequence-based SNP genotyping in durum wheat**. *Plant biotechnology journal* 2013, **11**(7):809-817.
52. Wang S, Wong D, Forrest K, Allen A, Chao S, Huang BE, Maccaferri M, Salvi S, Milner SG, Cattivelli L *et al*: **Characterization of polyploid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array**. *Plant biotechnology journal* 2014.
53. Trebbi D, Maccaferri M, de Heer P, Sorensen A, Giuliani S, Salvi S, Sanguineti MC, Massi A, van der Vossen EA, Tuberosa R: **High-throughput SNP discovery and genotyping in durum wheat (Triticum durum Desf.)**. *TAG Theoretical and applied genetics Theoretische und angewandte Genetik* 2011, **123**(4):555-569.

54. Sears ER: **The Aneuploids of common wheat**: College of Agriculture, Agricultural Experiment Station; 1954.
55. Sears ER: **Nullisomic-Tetrasomic Combinations in Hexaploid Wheat**. In: *Chromosome Manipulations and Plant Genetics.* Edited by Riley R, Lewis KR: Springer US; 1966: 29-45.
56. Sears ER SL: **The telocentric chromosomes of common wheat**. In: *Proceedings of the 5th International Wheat Genetics Symposium* New Dehli: Ramanujam, S. Indian Society of Genetics and Plant Breeding 1979: 389-407.
57. Endo TR, Gill BS: **The deletion stocks of common wheat**. *J Hered* 1996, **87**(4):295-307.
58. Qi LL, Echalier B, Chao S, Lazo GR, Butler GE, Anderson OD, Akhunov ED, Dvorak J, Linkiewicz AM, Ratnasiri A *et al*: **A chromosome bin map of 16,000 expressed sequence tag loci and distribution of genes among the three genomes of polyploid wheat**. *Genetics* 2004, **168**(2):701-712.
59. Czyczylo-Mysza I, Tyrka M, Marcinska I, Skrzypek E, Karbarz M, Dziurka M, Hura T, Dziurka K, Quarrie SA: **Quantitative trait loci for leaf chlorophyll fluorescence parameters, chlorophyll and carotenoid contents in relation to biomass and yield in bread wheat and their chromosome deletion bin assignments**. *Mol Breeding* 2013, **32**(1):189-210.
60. Sourdille P, Singh S, Cadalen T, Brown-Guedira GL, Gay G, Qi L, Gill BS, Dufour P, Murigneux A, Bernard M: **Microsatellite-based deletion bin system for the establishment of genetic-physical map relationships in wheat (Triticum aestivum L.)**. *Functional & integrative genomics* 2004, **4**(1):12-25.
61. Cui F, Fan X, Zhao C, Zhang W, Chen M, Ji J, Li J: **A novel genetic map of wheat: utility for mapping QTL for yield under different nitrogen treatments**. *BMC genetics* 2014, **15**(1):57.
62. Goss SJ, Harris H: **New method for mapping genes in human chromosomes**. *Nature* 1975, **255**(5511):680-684.
63. Geisler R, Rauch GJ, Baier H, van Bebber F, Bross L, Dekens MP, Finger K, Fricke C, Gates MA, Geiger H *et al*: **A radiation hybrid map of the zebrafish genome**. *Nature genetics* 1999, **23**(1):86-89.
64. Hawken RJ, Murtaugh J, Flickinger GH, Yerle M, Robic A, Milan D, Gellin J, Beattie CW, Schook LB, Alexander LJ: **A first-generation porcine whole-genome radiation hybrid map**. *Mammalian genome : official journal of the International Mammalian Genome Society* 1999, **10**(8):824-830.
65. Cox DR, Burmeister M, Price ER, Kim S, Myers RM: **Radiation hybrid mapping: a somatic cell genetic method for constructing high-resolution maps of mammalian chromosomes**. *Science* 1990, **250**(4978):245-250.
66. Kalavacharla V, Hossain K, Gu Y, Riera-Lizarazu O, Vales MI, Bhamidimarri S, Gonzalez-Hernandez JL, Maan SS, Kianian SF: **High-resolution radiation hybrid map of wheat chromosome 1D**. *Genetics* 2006, **173**(2):1089-1099.
67. Paux E, Sourdille P, Salse J, Saintenac C, Choulet F, Leroy P, Korol A, Michalak M, Kianian S, Spielmeyer W *et al*: **A physical map of the 1-gigabase bread wheat chromosome 3B**. *Science* 2008, **322**(5898):101-104.
68. Kumar A, Simons K, Iqbal MJ, de Jimenez MM, Bassi FM, Ghavami F, Al-Azzam O, Drader T, Wang Y, Luo MC *et al*: **Physical mapping resources for large plant genomes: radiation hybrids for wheat D-genome progenitor Aegilops tauschii**. *BMC genomics* 2012, **13**:597.
69. Gao W, Chen ZJ, Yu JZ, Raska D, Kohel RJ, Womack JE, Stelly DM: **Wide-cross whole-genome radiation hybrid mapping of cotton (Gossypium hirsutum L.)**. *Genetics* 2004, **167**(3):1317-1329.
70. Wardrop J, Snape J, Powell W, Machray GC: **Constructing plant radiation hybrid panels**. *The Plant journal : for cell and molecular biology* 2002, **31**(2):223-228.
71. Bleuyard JY, Gallego ME, White CI: **Recent advances in understanding of the DNA double-strand break repair machinery of plants**. *DNA Repair* 2006, **5**(1):1-12.

72. Riera-Lizarazu O, Vales MI, Ananiev EV, Rines HW, Phillips RL: **Production and characterization of maize chromosome 9 radiation hybrids derived from an oat-maize addition line**. *Genetics* 2000, **156**(1):327-339.
73. International Brachypodium I: **Genome sequencing and analysis of the model grass Brachypodium distachyon**. *Nature* 2010, **463**(7282):763-768.
74. Zhang GY, Liu X, Quan ZW, Cheng SF, Xu X, Pan SK, Xie M, Zeng P, Yue Z, Wang WL *et al*: **Genome sequence of foxtail millet (Setaria italica) provides insights into grass evolution and biofuel potential**. *Nature biotechnology* 2012, **30**(6):549-+.
75. Mayer KF, Taudien S, Martis M, Simkova H, Suchankova P, Gundlach H, Wicker T, Petzold A, Felder M, Steuernagel B *et al*: **Gene content and virtual gene order of barley chromosome 1H**. *Plant Physiol* 2009, **151**(2):496-505.
76. Mayer KF, Martis M, Hedley PE, Simkova H, Liu H, Morris JA, Steuernagel B, Taudien S, Roessner S, Gundlach H *et al*: **Unlocking the barley genome by chromosomal and comparative genomics**. *The Plant cell* 2011, **23**(4):1249-1263.
77. Hernandez P, Martis M, Dorado G, Pfeifer M, Galvez S, Schaaf S, Jouve N, Simkova H, Valarik M, Dolezel J *et al*: **Next-generation sequencing and syntenic integration of flow-sorted arms of wheat chromosome 4A exposes the chromosome structure and gene content**. *The Plant journal : for cell and molecular biology* 2012, **69**(3):377-386.
78. Philippe R, Paux E, Bertin I, Sourdille P, Choulet F, Laugier C, Simkova H, Safar J, Bellec A, Vautrin S *et al*: **A high density physical map of chromosome 1BL supports evolutionary studies, map-based cloning and sequencing in wheat**. *Genome biology* 2013, **14**(6):R64.
79. Martis MM, Zhou R, Haseneyer G, Schmutzer T, Vrana J, Kubalakova M, Konig S, Kugler KG, Scholz U, Hackauf B *et al*: **Reticulate evolution of the rye genome**. *The Plant cell* 2013, **25**(10):3685-3698.
80. Pfeifer M, Martis M, Asp T, Mayer KFX, Lubberstedt T, Byrne S, Frei U, Studer B: **The Perennial Ryegrass GenomeZipper: Targeted Use of Genome Resources for Comparative Grass Genomics**. *Plant physiology* 2013, **161**(2):571-582.
81. Sachin Kumar HSB, Pushpendra Kumar Gupta: **Comparative DNA Sequence Analysis Involving Wheat, Brachypodium and Rice Genomes Using Mapped Wheat ESTs**. *Triticeae Genomics and Genetics* 2012, **3**(3):25-37.
82. Choi JH, Kim S, Tang H, Andrews J, Gilbert DG, Colbourne JK: **A machine-learning approach to combined evidence validation of genome assemblies**. *Bioinformatics* 2008, **24**(6):744-750.
83. Phillippy AM, Schatz MC, Pop M: **Genome assembly forensics: finding the elusive mis-assembly**. *Genome biology* 2008, **9**(3):R55.
84. Bovee D, Zhou Y, Haugen E, Wu Z, Hayden HS, Gillett W, Tuzun E, Cooper GM, Sampas N, Phelps K *et al*: **Closing gaps in the human genome with fosmid resources generated from multiple individuals**. *Nature genetics* 2008, **40**(1):96-101.
85. Pan Y, Deng Y, Lin H, Kudrna DA, Wing RA, Li L, Zhang Q, Luo M: **Comparative BAC-based physical mapping of Oryza sativa ssp. indica var. 93-11 and evaluation of the two rice reference sequence assemblies**. *The Plant journal : for cell and molecular biology* 2014, **77**(5):795-805.
86. Haiminen N, Feltus FA, Parida L: **Assessing pooled BAC and whole genome shotgun strategies for assembly of complex genomes**. *BMC genomics* 2011, **12**:194.
87. Neumann K, Kobiljski B, Dencic S, Varshney RK, Borner A: **Genome-wide association mapping: a case study in bread wheat (Triticum aestivum L.)**. *Mol Breeding* 2011, **27**(1):37-58.
88. Gupta PK, Langridge P, Mir RR: **Marker-assisted wheat breeding: present status and future possibilities**. *Mol Breeding* 2010, **26**(2):145-161.
89. Poland J, Endelman J, Dawson J, Rutkoski J, Wu SY, Manes Y, Dreisigacker S, Crossa J, Sanchez-Villeda H, Sorrells M *et al*: **Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing**. *Plant Genome-Us* 2012, **5**(3):103-113.

90. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I: **ABySS: a parallel assembler for short read sequence data**. *Genome research* 2009, **19**(6):1117-1123.
91. Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S *et al*: **High-quality draft assemblies of mammalian genomes from massively parallel sequence data**. *Proc Natl Acad Sci U S A* 2011, **108**(4):1513-1518.
92. Wetzel J, Kingsford C, Pop M: **Assessing the benefits of using mate-pairs to resolve repeats in de novo short-read prokaryotic assemblies**. *BMC bioinformatics* 2011, **12**.
93. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W: **Scaffolding pre-assembled contigs using SSPACE**. *Bioinformatics* 2011, **27**(4):578-579.
94. Argout X, Salse J, Aury JM, Guiltinan MJ, Droc G, Gouzy J, Allegre M, Chaparro C, Legavre T, Maximova SN *et al*: **The genome of Theobroma cacao**. *Nature genetics* 2011, **43**(2):101-108.
95. Peng Z, Lu Y, Li L, Zhao Q, Feng Q, Gao Z, Lu H, Hu T, Yao N, Liu K *et al*: **The draft genome of the fast-growing non-timber forest species moso bamboo (Phyllostachys heterocycla)**. *Nature genetics* 2013, **45**(4):456-461, 461e451-452.
96. van Heesch S, Kloosterman WP, Lansu N, Ruzius FP, Levandowsky E, Lee CC, Zhou SG, Goldstein S, Schwartz DC, Harkins TT *et al*: **Improving mammalian genome scaffolding using large insert mate-pair next-generation sequencing**. *BMC genomics* 2013, **14**.
97. English AC, Richards S, Han Y, Wang M, Vee V, Qu J, Qin X, Muzny DM, Reid JG, Worley KC *et al*: **Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology**. *PloS one* 2012, **7**(11):e47768.
98. International Barley Genome Sequencing C, Mayer KF, Waugh R, Brown JW, Schulman A, Langridge P, Platzer M, Fincher GB, Muehlbauer GJ, Sato K *et al*: **A physical, genetic and functional sequence assembly of the barley genome**. *Nature* 2012, **491**(7426):711-716.
99. Sharma SK, Bolser D, de Boer J, Sonderkaer M, Amoros W, Carboni MF, D'Ambrosio JM, de la Cruz G, Di Genova A, Douches DS *et al*: **Construction of reference chromosome-scale pseudomolecules for potato: integrating the potato genome with genetic and physical maps**. *G3* 2013, **3**(11):2031-2047.
100. Wei F, Zhang J, Zhou S, He R, Schaeffer M, Collura K, Kudrna D, Faga BP, Wissotski M, Golser W *et al*: **The physical and genetic framework of the maize B73 genome**. *PLoS genetics* 2009, **5**(11):e1000715.
101. Febrer M, Goicoechea JL, Wright J, McKenzie N, Song X, Lin J, Collura K, Wissotski M, Yu Y, Ammiraju JS *et al*: **An integrated physical, genetic and cytogenetic map of Brachypodium distachyon, a model system for grass research**. *PloS one* 2010, **5**(10):e13461.
102. Hossain KG, Kalavacharla V, Lazo GR, Hegstad J, Wentz MJ, Kianian PM, Simons K, Gehlhar S, Rust JL, Syamala RR *et al*: **A chromosome bin map of 2148 expressed sequence tag loci of wheat homoeologous group 7**. *Genetics* 2004, **168**(2):687-699.
103. Kumar A, Bassi FM, Paux E, Al-Azzam O, de Jimenez MM, Denton AM, Gu YQ, Huttner E, Kilian A, Kumar S *et al*: **DNA repair and crossing over favor similar chromosome regions as discovered in radiation hybrid of Triticum**. *BMC genomics* 2012, **13**:339.
104. Cheng Z, Presting GG, Buell CR, Wing RA, Jiang J: **High-resolution pachytene chromosome mapping of bacterial artificial chromosomes anchored by genetic markers reveals the centromere location and the distribution of genetic recombination along chromosome 10 of rice**. *Genetics* 2001, **157**(4):1749-1757.
105. Zhou S, Wei F, Nguyen J, Bechner M, Potamousis K, Goldstein S, Pape L, Mehan MR, Churas C, Pasternak S *et al*: **A single molecule scaffold for the maize genome**. *PLoS genetics* 2009, **5**(11):e1000711.
106. Hastie AR, Dong LL, Smith A, Finklestein J, Lam ET, Huo NX, Cao H, Kwok PY, Deal KR, Dvorak J *et al*: **Rapid Genome Mapping in Nanochannel Arrays for Highly Complete and Accurate De Novo Sequence Assembly of the Complex Aegilops tauschii Genome**. *PloS one* 2013, **8**(2).

# Paper I

Belova et al. BMC Genomics 2013, 14:222
http://www.biomedcentral.com/1471-2164/14/222

BMC Genomics

RESEARCH ARTICLE Open Access

# Integration of mate pair sequences to improve shotgun assemblies of flow-sorted chromosome arms of hexaploid wheat

Tatiana Belova[1], Bujie Zhan[1], Jonathan Wright[2], Mario Caccamo[2], Torben Asp[3], Hana Šimková[4], Matthew Kent[5], Christian Bendixen[6], Frank Panitz[6], Sigbjørn Lien[5], Jaroslav Doležel[4], Odd-Arne Olsen[1] and Simen R Sandve[1*]

## Abstract

**Background:** The assembly of the bread wheat genome sequence is challenging due to allohexaploidy and extreme repeat content (>80%). Isolation of single chromosome arms by flow sorting can be used to overcome the polyploidy problem, but the repeat content cause extreme assembly fragmentation even at a single chromosome level. Long jump paired sequencing data (mate pairs) can help reduce assembly fragmentation by joining multiple contigs into single scaffolds. The aim of this work was to assess how mate pair data generated from multiple displacement amplified DNA of flow-sorted chromosomes affect assembly fragmentation of shotgun assemblies of the wheat chromosomes.

**Results:** Three mate pair (MP) libraries (2 Kb, 3 Kb, and 5 Kb) were sequenced to a total coverage of 89x and 64x for the short and long arm of chromosome 7B, respectively. Scaffolding using SSPACE improved the 7B assembly contiguity and decreased gene space fragmentation, but the degree of improvement was greatly affected by scaffolding stringency applied. At the lowest stringency the assembly N50 increased by ~7 fold, while at the highest stringency N50 was only increased by ~1.5 fold. Furthermore, a strong positive correlation between estimated scaffold reliability and scaffold assembly stringency was observed. A 7BS scaffold assembly with reduced MP coverage proved that assembly contiguity was affected only to a small degree down to ~50% of the original coverage.

**Conclusion:** The effect of MP data integration into pair end shotgun assemblies of wheat chromosome was moderate; possibly due to poor contig assembly contiguity, the extreme repeat content of wheat, and the use of amplified chromosomal DNA for MP library construction.

**Keywords:** Wheat, Assembly, Scaffold, Mate-pair, MDA, Improvement

## Background

Bread wheat is one of the most important food crops worldwide. However, present wheat production is far from the expected increased global demand in the near future [1,2]. Development of better yielding varieties with improved adaptation to the new climatic challenges is therefore important for global food security. A 'tool' with a great potential to revolutionize wheat breeding and production is a publicly available reference genome sequence. Genome sequences enable cost-effective identification of genomic variation which subsequently can be used to improve agricultural traits of interest through marker-assisted selection (MAS) and genomic selection programs [3]. A rapidly increasing number of genomes from important food crops are becoming available. In 2011 potato and cacao [4,5], in 2010 soybean [6], and in 2009 maize, sorghum and cucumber genomes were published [7-9]. However, even though wheat is one of the top five food commodities in the world, a wheat genome sequence is not yet available.

The main reason why the wheat genome sequencing is lagging behind is related to technical challenges due to

* Correspondence: simen.sandve@umb.no
[1]Department of Plant and Environmental Sciences, University of Life Sciences, Ås, Norway
Full list of author information is available at the end of the article

large size (17Gb) and the complexity of the hexaploid wheat genome. Bread wheat is allohexaploid and carries three distinct, but closely related homoelogous genomes ($2n = 6x = 42$, AABBDD) [10,11]. A distinction between homoeolog sequences in post sequencing processing of genomic sequence data is essentially impossible. Fortunately, the hexaploid wheat genome can be dissected to small parts by flow cytometric sorting of single chromosomes and chromosome arms [12,13]. This technological breakthrough has enabled production of wheat chromosome specific BAC-libraries [14] and facilitated construction of physical maps of hexaploid wheat chromosomes [15]. For some genomic applications, such as shotgun sequencing, large amount of DNA are required. In order to obtain sufficient DNA to sequence purified chromosome arms, millions of chromosomes must be sorted, a process, which is highly labor intensive [16]. Including an amplification step of flow-sorted DNA can significantly reduce the labor and consequently the cost of acquiring chromosome specific DNA for sequencing. Multiple displacement amplification (MDA) is the most common method for genome amplification for sequencing purposes as MDA generate relatively long amplification products (majority between 5-20 kb) [17]. However, MDA is known to give rise to chimeras, which can bring down the utility of the amplified DNA [18].

Shotgun sequencing of MDA DNA from flow-sorted chromosome arms, especially in combination with genetic maps and synteny information, has proven to be a highly cost effective way of gene discovery and construction of syntenic chromosome assemblies [19-21]. Unfortunately, the fragmentation level of the shotgun assemblies has been very high, which limits the information value of the assemblies. *De novo* assemblies of 7DS and 7BS using Illumina paired-end (PE) sequences with a chromosome arm coverage of 30-34×, resulted roughly in 600,000-1,000,000 contigs per chromosome arm, an N50 of ~500-1200 bp, and maximum contig sizes of just over 30,000 bp [21,22]. Consequently, many contigs do not contain complete gene sequences, and the relative order of genes can only be identified for a small subset of genes found on contigs containing multiple genes (i.e. multigene contigs).

High levels of DNA sequence assembly fragmentation is closely associated with the repeat content of the genome [23], and the wheat genome is extreme with respect to repeat content, having more than 80% repetitive DNA [24]. One way of reducing assembly fragmentation is to include additional sequencing libraries with large insert sizes, referred to as mate pair (MP) libraries [23]. MP reads can vary in insert sizes between 1-20 kb and the idea of these 'long jump' paired sequences is to span repetitive regions that cause assembly fragmentation, and thereby link multiple contigs into longer scaffolds. This will improve the information value of an assembly by (1) improving the assembly contiguity (2) increasing the proportion of full length genes contained in single sequences (i.e. link exons from different contigs), and (3) increase the number of linearly ordered genes.

A number of recent publications describe the effect of MP data on assemblies of plant genomes [4,9,25]. One example is the potato genome assembly, which had on average an N50 increase of 37 Kb for every 1 Kb increase in MP insert size [25]. Although the potato genome (1C = 865 Mbp) has a relatively high repeat content (total repeat content ≈ 62%, TE-derived repeats ≈ 32%), it does not compare to the hexaploid wheat genome (1C = 17,000 Mbp) that has >80% of TE-derived repetitive DNA [24]. It is thus not clear to what extent MP data may improve shotgun assemblies of genomes with extreme repeat content such as wheat. Additionally, the utility of MP data from MDA DNA from flow-sorted chromosomes is unknown. The aim of this paper is therefore to study the effects of MP from MDA DNA on assembly contiguousness and gene content in shotgun assemblies of a flow-sorted hexaploid wheat chromosome.

## Methods

### Preparation of DNA from chromosome arms 7BS and 7BL

A double ditelosomic line of wheat *Triticum aestivum* L. cv. Chinese Spring carrying both arms of chromosome 7B as telosomes (2n = 40 + 2t7BS + 2t7BL) was used to purify the 7BS and 7BL arms. The seeds were provided by Dr. Bikram Gill (Kansas State University, Manhattan, USA). The chromosome arms were purified by flow cytometry. 68,000 and 45,000 of 7BS and 7BL arms, respectively, corresponding to 50 ng of DNA, were isolated in several batches. In order to estimate contamination with other chromosomes, 1000 chromosomes were sorted onto a microscope slide and used for fluorescence *in situ* hybridization (FISH) with probes for *Afa* family and telomeric repeats. Batches with the highest purity of the sorted fraction (93 and 88% for 7BS and 7BL, respectively) were used for further processing. DNA was purified and subsequently amplified using Illustra GenomiPhi V2 DNA Amplification Kit (GE Healthcare, Chalfont St. Giles, United Kingdom) as previously described [17]. Three independent amplifications were performed for each arm to reduce amplification bias. Totally, 15.9 and 14 micrograms were prepared for 7BS and 7BL, respectively.

### Sequencing library construction

PE libraries with a mean insert size of ~350 bp (Illumina protocol) and 2 Kb MP libraries (in-house modified Roche MP protocol) were constructed and sequenced at Fasteris SA (Geneva, Switzerland). The PE reads were 100 bp, while the 2 Kb MP reads were 45 bp. 3 Kb and 5 Kb MP libraries were prepared according to "Mate Pair Library v2 Sample Preparation Guide" [26] at Aarhus University (Denmark).

The read length of the 3 and 5 Kb MP libraries were trimmed to 35 bp. All MP libraries were sequenced using HiSeq2000 technology (Illumina) according to manufacturer's recommendations.

### Contig assembly

Contigs were assembled with PE reads using ABySS [27] which is based on a *de Bruijn* graph approach. This method collects the information generated from fix-length words of *k –mers* shared by overlapping reads [28]. Initially, multiple assemblies were generated using different values of *k* and assessed using assembly quality statistics such as N50, maximum contig length, number of contigs in the assembly and the total amount of bases in the assembly. A *k-mer* length of 71 was chosen as the optimal value. A seed value of 150 was used (s parameter) and a minimum of 10 pairs were required to join contigs (n parameter). After assembly, contigs shorter than 200 bp were removed to generate a filtered dataset for scaffolding.

### Scaffold assembly

To accurately determine the mean insert size and insert size variation of each MP library, we mapped all mate pair reads back to the 7B contigs using BWA v0.6.0 [29] with the parameters BWA aln –t 10 -q 10. Based on the BWA results we identified the number of MP reads aligning to contigs, the proportion of MP read pairs mapping to the same or different contigs, and the orientation of the MP reads that mapped to the same contig. We also assessed if the genomic origin of MP reads were biased towards different fractions of the genome (i.e. repeat or conserved fraction). This was done by mapping reads to an in house repeat content database (TREP10 and the repeats identified in Choulet et al. 2011) and the NCBI nr database.

We initially tested three software packages for scaffolding of pre-assembled contigs: ABySS, SOAPdenovo and SSPACE. Unfortunately, we were not able to scaffold contigs using ABySS due to the large proportion of MP reads that mapped in forward-forward direction (see results and discussion for more details). SOAP and SSPACE both produced scaffolds, but as the N50 and gene space assembly statistics of SSPACE assemblies exceeded SOAPdenovo at all parameters tested, we chose to use SSPACE for further investigation of the effect of MP on shotgun assemblies. In SSPACE the key parameter that defines the stringency of the scaffolding is 'number of links' (k), i.e. number of independent read pairs that uniquely support a connection between two contigs. We performed SSPACE scaffolding with k equal to 3, 5, 7, 10, 15 and 20.

### Gene content

The protein annotation (v1.2) excluding splice variants of *Brachypodium distachyon* (referred to as Brachypodium) was used as query sequences in a TBLASTN search [30] to assess gene content in contigs and scaffold assemblies. Blast result filtering were carried out as follows: (1) Only query proteins having at least one exon hit with minimum 30 amino acid length and a minimum per cent identity of 70 were considered in the analyses. (2) Duplicated exon hits on one contig/scaffold were removed. Duplicate hits were defined as two or more query hits with identical query start and query end positions, identical mismatches, identical gap length, and identical hit identity. (3) For each query protein, the mean e-value of all hits were calculated and overlapping exon hits (overlapping >5 bp) from proteins with higher e-value were discarded. (4) Two types of gene coverage were calculated: 'total coverage' and 'adjusted coverage'. Total coverage was calculated as the total length of all the hits from a protein query relative to the query sequence length. Adjusted coverage was calculated as the number of unique query amino acid residues with a blast hit in the target sequence(s). To exclude gene hits from repetitive DNA (e.g. TE-associated coding regions) and spurious protein homology, genes with total coverage of >5 and genes with <10% adjusted coverage was not considered in any analyses.

A gene fragmentation index (GFI) was estimated to compare gene space fragmentation in different assemblies. The average blast hit coverage of Brachypodium gene homologs in the entire assembly, referred to as assembly coverage (AC) represents an approximation of the theoretically optimal situation, when each gene is contained within a single DNA sequence (i.e. no fragmentation). The AC estimate was then compared with the average Brachypodium gene coverage per contig or scaffolds, referred to as sequence coverage (SC), to calculate a gene fragmentation index (GFI) defined as (AC-SC)/AC. Hence, the GFI measures gene fragmentation as the difference in percent between SC and AC, and approaches 0 as SC and AC become similar.

### Evaluation of scaffold reliability

As we cannot directly measure the level of scaffolding errors due to the lack of any reference assembly, we estimated the level of scaffold errors by (1) utilizing information from synteny with Brachypodium and (2) comparing the 7BL scaffold assemblies with the sequence content of 50 random BAC clones from 7BL. Because the number of chimeric contigs is assumed to be very low, the level of errors introduced by scaffolding can be estimated by comparing the synteny levels in contigs with synteny in scaffolds of similar sizes. If homologs of two Brachypodium genes are present in a single wheat contig, these homologs have a probability of representing closely linked loci (referred to as neighbouring genes) on the Brachypodium chromosome. This probability depends on the synteny level between wheat and

Brachypodium in that exact region. If the scaffolding process does not introduce structural assembly errors, the proportion of neighbouring Brachypodium homologs should be similar in contigs and scaffolds of similar size. In our analyses we defined a neighbour gene pair as genes originating from Brachypodium loci with < =50 genes distance from each other. A bootstrap test was performed to test if the difference in proportions of neighbouring loci in contigs and scaffolds were likely to occur as a consequence of random sampling error. One thousand contig datasets were re-sampled (with replacement) and the *P*-value was calculated as the proportion of bootstrapped contig datasets with equal or lower proportions of neighbour genes as found in scaffolds.

In addition to the synteny approach we also utilized the sequence content of 50 BAC clones originating from 7BL to evaluate scaffold reliability (See Additional file 1 for assembly methods and Additional file 2 for sequence contigs). Raw sequencing reads from 7BL BACs are available upon request. We first identified scaffolds containing sequences derived from the BACs by BLASTN, using a threshold of >99% identity across minimum 2.5 Kb. With the assumption that identified scaffolds truly are derived from one of these 50 BACs, an estimate of scaffold reliability can therefore be defined as the proportion of contigs within a scaffold that originate from a certain BAC. To assess if contigs in scaffolds originate from the BAC we used BLASTN and defined a significant contig-to-BAC hit as having >99% identity across >50% of the contig length. Because longer scaffolds are more likely contain sequences belonging to multiple BACs (i.e. lower proportion of contigs originating from a single BAC) and scaffolding stringency affect scaffold length distribution, we normalized the scaffold reliability by dividing on scaffold length (i.e. proportion of contigs in a scaffold with a BLASTN hit to BAC/scaffold length). Normalized scaffold reliability is hereafter referred to as scaffold reliability index (SRI).

## Results

### Shotgun assembly of 7BS and 7BL

106 and 100 million 100 bp PE reads with an average insert size of 346 bp (7BS) and 362 bp (7BL) (Additional file 3) were generated from the MDA DNA from flow-sorted 7BS and 7BL chromosome arms, respectively (short read archive accession number: ERP002001). Of the mapped read pairs >99.8% were oriented in the assumed FR directions. This represents approximately 59x coverage of 7BS and 37× coverage of 7BL. The assembly with ABySS produced a total of 1,349,563 contigs for 7BS and 4,527,901 contigs for 7BL (Table 1) (contigs are available upon requests). After removing contigs of less than 200 bp, the assemblies were reduced to 178,789 7BS contigs and 328,725 7BL contigs, with an N50 of 2,428 and 1,556 bp, respectively (Table 1). The filtered datasets constituted 13.3% of 7BS and 7.3% of 7BL contigs, representing 57% and 48% of the two chromosome arms assuming a molecular size of 360 Mbp for 7BS and 540 Mbp for 7BL [14], respectively.

### Mate pair data

A total of 445 million 7BS and 478 million 7BL MP read pairs were generated (short read archive accession number: ERP002001), the coverage was estimated to be 88.9x and 63.9× for the short and long arm, respectively (Table 2). Seventy-one per cent of the MPs had both reads mapping to the assembly, 23% of the read pairs only had one read mapping to a contig (i.e. singleton), and about 5% of the MP data did not map to any of the 7B contigs. The between-library variation in the proportions of mapped reads were very low, however the 3 Kb and 5 Kb libraries had slightly smaller proportion of unmapped reads and singletons (2-5% less) compared to the 2 Kb library (data not shown).

MP reads, which map to the same contig, can be classified according to their orientation. In theory, MP should be oriented in a reverse/forward (RF) manner; however, of the MP reads that mapped to one contig, only 15% and 29% were classified as having a RF orientation on 7BS and 7BL, respectively. To better understand the nature of the non-MP oriented read pairs, we estimated insert size based on the mapping information. Figure 1 illustrates the variation and distribution of insert sizes for the RF, FR, and FF/RR oriented MP reads in the 3 Kb library of 7BS (Similar figures for all libraries can be found in Additional files 4 and 5). It is evident that the insert size distribution of the properly oriented MP read pairs represents a mix of the expected size range (a normally distributed peak) in addition to a relatively high proportion of reads with smaller and variable insert sizes. The insert size distribution of the non-MP oriented FF/RR and FR reads does not show the expected normal distribution, but is more similar to a log-normal distribution with a large proportion of reads from short insert size fragments of <1000 bp.

Mapping of MP reads classified as having RF, FR, and FF/RR orientation to the repeat database showed no apparent difference in repeat content (~36% mapped to repeat database, data not shown). It is not uncommon to have PE oriented reads (i.e. FR) in MP libraries (c.f. mate pair library sample preparation guide), but the occurrence of FF/RR reads is more difficult to explain. Hence, we specifically analysed the content of the FF/RR-mapped reads by BLASTN against the NCBI nr database to assess if the FF/RR reads originated from other sources than wheat DNA. All target hits with >80% identity in the NCBI nr nucleotide database were collected and the species information of each hit was extracted. More than 96% of the reads had a best BLAST hit to other grasses

**Table 1 Contig assembly summary statistics**

| Arm | Contig number | N50 (bp) | Mean length (bp) | Max length (bp) | Total (Mbp) |
|---|---|---|---|---|---|
| 7BS | 1,349,563 | 842 | 239 | 50,938 | 323 |
| | 178,789* | 2428 | 1152 | 50,938 | 206 |
| 7BL | 4,527,901 | 145 | 144 | 30,964 | 652 |
| | 328,725* | 1556 | 789 | 30,964 | 260 |

*Contigs > 200 bp.

(Additional file 6), implying that the FF/RR-mapping reads truly were derived from wheat DNA.

### Effect of MP integration on 7B assemblies

In the process of producing scaffolds, SSPACE enforces stringent criteria for incorporating MP information; MP reads used in SSPACE must have a unique perfect hit in the contig assembly and satisfy the *a priori* defined insert size range. Across all SSPACE assemblies only 1–1.8% of the MP read pairs satisfied the perfect match, read orientation, and insert size criteria (Table 3). Most of the discarded MP reads (63-71%) were classified as having unsatisfied pairing orientation, i.e. either FF/RR or as a pair end read (FR).

Even though only 1-2% of MP data was used for scaffolding the assembly N50 was improved substantially at low stringency levels; at k = 3 assembly N50 increased by 6 and 7.2 fold for 7BS and 7BL, respectively (scaffolds are available upon requests). However, as expected, the reduction in assembly fragmentation was dramatically affected when k was increased (Table 3). The number of contigs incorporated in scaffolds decreased from 42% to 8% and from 52% to 11% for 7BL and 7BS, respectively, and the total number of scaffolds decreased by ~70% when increasing k from 3 to 20. Furthermore the total number of residues included in scaffolds was reduced from ~40% to 18% of the total chromosome length (Table 3). The mean scaffold length however, was not affected much when k was increased due to a change in the distribution of length of contigs included in scaffolds; as k increased, the proportion of long contigs included in scaffolds also increased. Scaffold content was strongly biased towards gene containing contigs. Although a maximum of 40-50% of 7BS and 7BL contigs were incorporated into scaffolds (Table 3), as many as 75% (k = 20) and 95% (k = 3) of sequences containing full length genes (> = 70% Brachypodium homolog coverage) were included in scaffolds.

Next we assessed how MP data helped to join fragmented gene parts into more complete gene sequences by calculation of a gene fragmentation index (GFI) and counting full length genes contained in single sequences. After removing all BLAST hits with <10% coverage of a Brachypodium protein the AC and GFI were 0.54/0.17 and 0.49/0.21 in the 7BS and 7BL contig assemblies, respectively, while the scaffold assembly GFI ranged between 0.09-0.14 (Table 4). The MP integration also increased the number of full length genes in the range of 10–16% and 20-30%, depending on how a full length gene was defined (Table 4). In addition to aiding the joining of exons from fragmented gene sequences, MP information also helps to link genes belonging to different contigs together in multigene containing scaffolds (containing > =2 full length genes), and thereby helps ordering genes relative to each other. A modest effect on gene linking was observed after the MP integration in the 7BS and 7BL assemblies (Figure 2). The number of sequences containing 2 and 3 genes increased by 2–3 fold when applying k = 3 and by 1.5-2 fold when applying k = 20, compared to the contig assembly. However, virtually no changes was observed for sequences containing >3 genes. Moreover, the gene composition in scaffolds containing multiple genes were not random with respect to the length of the gene, but showed a clear bias towards shorter genes. For example, for the SSPACE k = 5 assembly the mean gene length in scaffolds and contigs with 3 or more genes were much shorter (contigs 462 bp/scaffolds 722 bp) than the mean CDS in sequences with 2 genes (contigs = 1,354 bp/scaffolds = 1,604 bp).

### Scaffold content reliability

Integration of mate pair data can lead to misassemblies due to erroneous coupling of contigs. We took advantage of the Brachypodium model genome to estimate the scaffolding error levels based on synteny. Scaffolds and

**Table 2 Summary table of mate pair sequence data**

| Arm | Mate pair library | | | Total pairs | Read class | | |
|---|---|---|---|---|---|---|---|
| | 2 Kb° | 3 Kb† | 5 Kb† | | Pairs | Singletons | Unaligned |
| 7BS | 2.60*107 | 2.23*108 | 1.97*108 | 4.46*108 | 71.8% | 22.4% | 5.9% |
| 7BL | 3.13*107 | 2.32*108 | 2.16*108 | 4.79*108 | 71.2% | 23.7% | 5.1% |

Total numbers of read pairs are given for each MP library. The read pair classification is based on mapping of MP data back to assembled contigs from PE data.
°Roche library, †Illumina library.

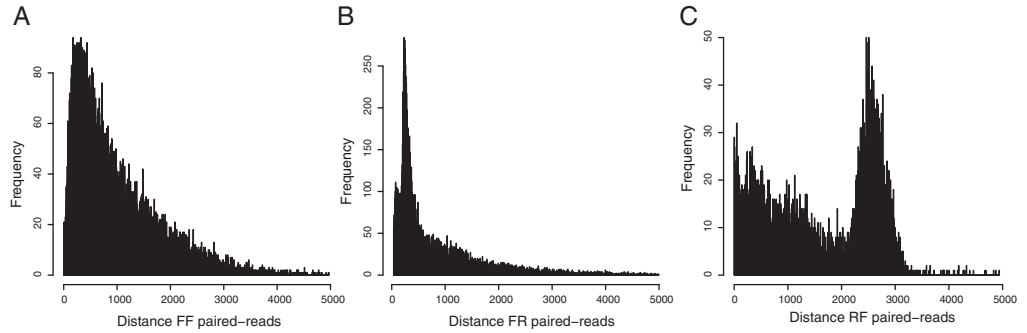


**Figure 1 MP read classes and the estimated insert size distributions for the 3 KB library of 7BS. A**) distribution of insert sizes for the FF reads **B**) distribution of insert sizes for the FR reads **C**) distribution of insert sizes for the RF reads.

contigs containing 2 full length Brachypodium homologs were identified and the proportion of neighbour genes based on the location in the Brachypodium genome was calculated. The frequency of neighbouring genes in contigs were 0.48 and in scaffolds between 0.4 (k = 3) and 0.49 (k = 7) (Figure 3A). Furthermore, the bootstrap tests did not reject the null hypotheses that contigs have higher proportion of neighbouring genes at $\alpha = 0.05$, even for the scaffolds produced at the lowest stringency (k = 3, $P = 0.15$). Taken together, our synteny error rate estimates do not indicate high rates of random contig joining at any level of SSPACE stringency. Scaffold reliability estimation based on sequence content in BAC clones reflected a slightly different and more pronounced effect of changing scaffolding stringency. The median scaffold reliability index increased progressively from the k = 3 (0.38) to k = 20 (0.86) assemblies (Figure 3B), indicating a higher scaffold correctness in SSPACE assemblies with high k-values.

### Effect of MP coverage

Increasing sequence coverage of MP data has impact on assembly statistics, but upon reaching certain coverage, added value of additional MP sequencing may not justify the cost of data generation. It is therefore important to evaluate the effect of MP coverage on our assembly metrics. To assess the relationship between MP coverage and assembly improvement, we generated randomly reduced datasets of our 7BS MP libraries with 1, 10, 25, 40, 50, 60 and 75% of original MP coverage and generated scaffolds with the number of links parameter k = 5. Three random sub-sets of MP data were generated for each reduced level of coverage. Interestingly, little change was observed in assembly statistics until the coverage was

**Table 3 Scaffold assembly summary statistics**

| k | Arm | MP used (%) | No. scaffolds | Contigs in scaffolds | | | Mean length (min-max) | Sum scaffolds (Mbp) | Assembly N50* (Kb) |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Mean | Max | % | (Kb) | | |
| 3 | 7BS | 0.96 | 20,654 | 4.51 | 38 | 52 | 11.2 (1.7-143.1) | 168 | 14.49 |
| | 7BL | 1.56 | 31,582 | 4.33 | 43 | 42 | 9.6 (2.0-117.6) | 192 | 11.15 |
| 5 | 7BS | 1.06 | 17,481 | 3.81 | 27 | 37 | 10.7 (1.7-129.4) | 148 | 11.03 |
| | 7BL | 1.41 | 23,365 | 3.91 | 32 | 28 | 9.5 (2.2-122.1) | 166 | 8.31 |
| 7 | 7BS | 1.14 | 15,230 | 3.4 | 20 | 29 | 10.5 (1.72-109.6) | 133 | 9 |
| | 7BL | 1.48 | 19,610 | 3.56 | 25 | 21 | 9.3 (2.3-81.9) | 148 | 6.33 |
| 10 | 7BS | 1.24 | 12,750 | 3.04 | 15 | 22 | 10.5 (1.8-108.9) | 115 | 7.04 |
| | 7BL | 1.58 | 15,896 | 3.22 | 20 | 16 | 9.3 (2.3-77.7) | 128 | 4.49 |
| 15 | 7BS | 1.35 | 9,733 | 2,73 | 12 | 15 | 10.7 (2.0-102.4) | 92 | 5.2 |
| | 7BL | 1.7 | 12,052 | 2.89 | 17 | 11 | 9.4 (2.54-67.4) | 103 | 2.84 |
| 20 | 7BS | 1.42 | 7,618 | 2,55 | 10 | 11 | 10.9 (2.1-73.3) | 76 | 4.2 |
| | 7BL | 1.79 | 9,458 | 2.68 | 14 | 8 | 9.6 (2.8-69.2) | 84 | 1.97 |

* Including all sequences (contigs + scaffolds).

**Table 4 Gene content in ABySS and SSPACE assemblies**

| Assembly | Arm | Brachypodium homologs (>30 aa, >70 pident) | Brachypodium homolog coverage[1] (mean) | GFI[2] | Full length genes | |
|---|---|---|---|---|---|---|
| | | | | | Coverage* | Start-stop† |
| SSPACE k3 | 7BS | 1029 | 0.49 | 0.09 | 449 | 193 |
| | 7BL | 1539 | 0.44 | 0.10 | 551 | 224 |
| SSPACE k5 | 7BS | 1038 | 0.49 | 0.09 | 445 | 193 |
| | 7BL | 1545 | 0.43 | 0.12 | 547 | 227 |
| SSPACE k7 | 7BS | 1032 | 0.49 | 0.09 | 449 | 196 |
| | 7BL | 1551 | 0.43 | 0.12 | 535 | 221 |
| SSPACE k10 | 7BS | 1038 | 0.49 | 0.09 | 447 | 195 |
| | 7BL | 1555 | 0.43 | 0.12 | 533 | 215 |
| SSPACE k15 | 7BS | 1040 | 0.48 | 0.12 | 436 | 186 |
| | 7BL | 1576 | 0.42 | 0.14 | 529 | 217 |
| SSPACE k20 | 7BS | 1048 | 0.47 | 0.13 | 433 | 183 |
| | 7BL | 1574 | 0.42 | 0.14 | 516 | 205 |
| Contigs | 7BS | 1071 | 0.45 | 0.17 | 403 | 160 |
| | 7BL | 1621 | 0.39 | 0.21 | 457 | 162 |

[1]Mean coverage per sequence (contig/scaffold) of Brachypodium homologs based on blast analyses (see methods).
[2] Gene Fragmentation Index (GFI) is defined in the methods section.
* TBLASTN hits covering > =70% of a homologous Brachypodium protein in a single contig/scaffold.
† TBLASTN hits covering an entire Brachypodium protein (+ − 10aa) in a single contig/scaffold.

reduced with 50% (Figure 4), and a corresponding 22% reduction in N50 was observed. Even less effect of decreasing MP coverage was seen in statistics for gene content information in reduced MP coverage assemblies. For example, reducing the MP coverage by 90% only produced a 25% reduction in the number of full length genes (427) while the 50% reduced coverage assembly contained 7% fewer full length genes (435).

## Discussion

### Modest effect of MP integration in 7B shotgun assemblies

*De novo* assembly of shotgun sequences from large plant genomes like wheat remains a challenging task, mainly due to prevalence of repetitive DNA [31]. Its presence can lead to complex, misassembled rearrangements and the collapse of reads coming from distinct copies of repetitive DNA into single assembled sequences [23]

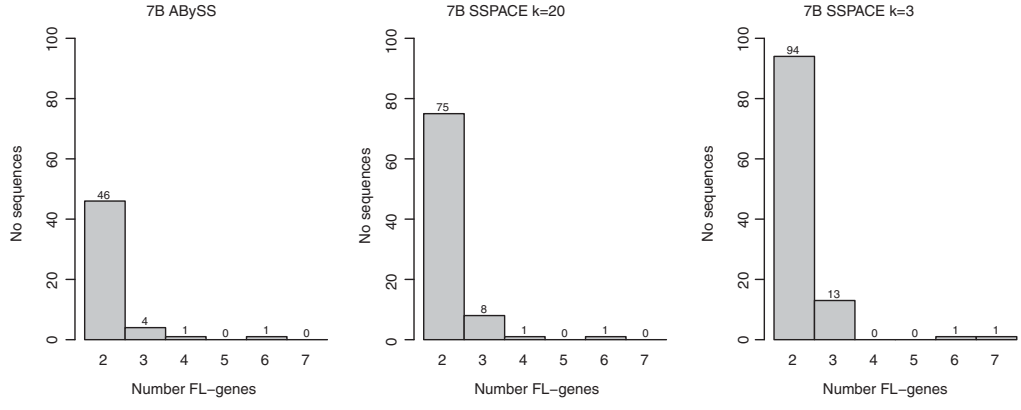


**Figure 2 MP effect on sequences containing multiple full length (FL) genes.**

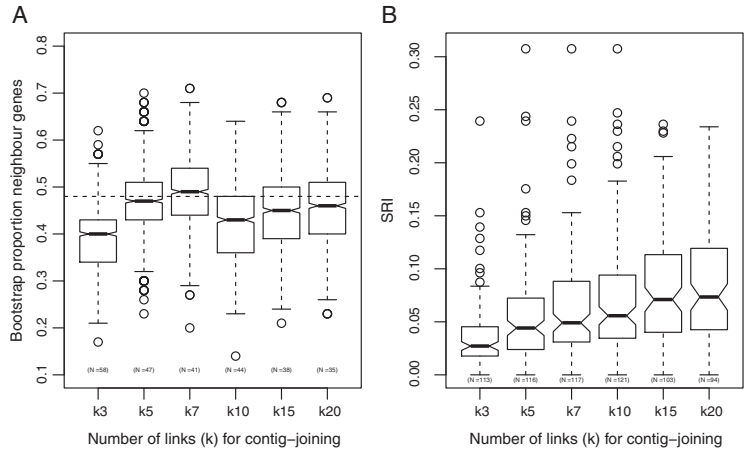


**Figure 3 Evaluation of scaffold reliability. A**) Scaffold reliability evaluated by the proportion of conserved syntenic relationship to the Brachypodium genome. **B**) Scaffold reliability evaluated by comparisons of sequence content in BACs and scaffolds. SRI = scaffold reliability index (defined in method section).

which results in contracted and fragmented assemblies. Ideally, genomes with high repeat content should therefore be assembled using reads longer then the length of the repeats. Wheat consists of >80% TE-derived repetitive DNA with a mean length of ~4600 bp [24], far longer than the read length of sequence reads from next generation sequencing data. One assembly strategy to reduce the fragmentation generated by repetitive DNA is therefore to link neighboring chromosome regions belonging to different sequence contigs with long insert size MP reads that bridge the repetitive segments.

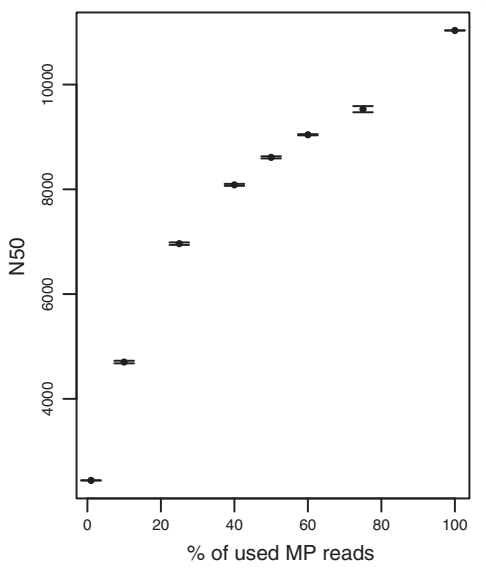


**Figure 4 Effect of MP sequencing depth on assembly N50.** Only 7BS was used to estimate the MP coverage effect.

Our results show that integration of MP data in shotgun assemblies of flow-sorted wheat chromosomes improves assembly contiguity and decrease gene space fragmentation, but that the degree of assembly statistics improvement is greatly affected by scaffolding stringency (Table 1, Table 3). For example, at low stringency the assembly N50 increased by 6–7.5-fold, while at the highest stringency (k = 20) N50 was only increased by 1.3-1.8 compared to contig assemblies. Although a negative correlation between stringency and assembly improvements also was observed for the gene space statistics, the MP effect on gene space were less affected by scaffold assembly stringency compared to the N50 statistic (Table 4). This is likely explained by the fact that genes are more often found in longer contigs, hence even at high stringency a large proportion of the gene containing contigs were joined into scaffolds.

Even for the least stringent scaffold assembly, the assembly improvement for 7B does not seem to be in the same magnitude as reported for some recently sequenced plant genomes. For the cucumber (1C = 367 Mbp) and cacao (1C = 430 Mbp) genomes, addition of long insert libraries improved N50 by 14-fold (172 Kb) and 60-fold (473.8 Kb), respectively [4,9]. Assembly metrics from these shotgun assemblies is difficult to compare directly due to the use of different sequencing platforms, different PE and MP libraries, and different sequencing coverage. However,

it is possible to estimate the MP-effectiveness based on the total gain in N50 per Kb of MP insert size length (N50 gain/max MP insert size). For cucumber and cacao these numbers are 14/2 Kb and 60/8 Kb compared to for example 5/5 Kb in wheat k = 5.

Potato is another recently published plant genome [25] for which different MP libraries were added to improve assembly. Integration of a 2 Kb MP library produced a 3-fold increase in N50, while using both 2 Kb + 5Kb MP libraries increased the N50 by 8-fold. Comparable metrics were obtained in 7B SSPACE assemblies using k = 5; the 2 Kb MP libraries produced 1.9- and 2.6-fold changes in N50 for 7BS and 7BL, respectively, while the final N50 fold change was ~5 after addition of the 5 Kb MP data. Thus, even though the actual scaffold N50 was much higher in potato after adding 2 + 5 Kb MP libraries (173 Kb) compared to the 7B assemblies with 2 + 3 + 5 Kb MP data (7BS = 11 Kb/7BL = 8.3 Kb), the relative N50 gain was not that different.

The modest impact of MP data in the chromosome 7B assemblies compared to other plant genomes could be explained by the inherent repeat characteristics of the wheat genome. While the wheat genome consists of >80% TE-derived repetitive DNA [24], potato, cacao and cucumber genomes are much smaller and all have <35% TE-derived repetitive DNA [32]. This difference will undoubtedly cause large differences in the effect of MP data on assembly contiguity. Another reason for the relatively low impact of MP data on the scaffold N50 could be related to the quality of the MP data. Only a small fraction (~1%) of the MP reads from MDA chromosomal DNA satisfied requirements of SSPACE for being included in scaffold construction, mostly due to a very high portion of reads having a different orientation or discrepancy between expected and estimated insert size for MP reads (Table 2). Lastly, the fragmentation level of the contig assembly is important for the MP effect. It is evident that small contigs have less chance of being put into scaffolds due to the fact that small sequences will have few MP reads originating from them. An improved contig assembly N50, for example by increasing PE sequencing coverage or adding additional PE libraries with different insert sizes, could therefore be a good strategy to be able to include a larger proportion of the contig assembly into scaffolds, and hence increase scaffold N50.

### Scaffold reliability and assembly stringency

Assembly errors can be introduced at the scaffolding stage when the software has to choose between two similar solutions and falsely connects contigs from non-adjacent chromosome regions or links two adjacent contigs in the wrong orientation. Our synteny- and BAC-based scaffold reliability estimates provides measures of reliability at two types of different genomic landscapes. Our synteny approach did not detect signatures of erroneous contig joining in small scaffolds from gene dense regions in the assemblies; however when using sequence contents from 50 BACs to assess scaffold reliability a strong correlation between estimated scaffold reliability and scaffold assembly stringency was observed (Figure 3). We interpret these differences in test conclusions to reflect that scaffolds from non-genic genomic regions are more prone to contain errors (especially at low stringency parameters), likely due to higher content of repetitive DNA in the intragenic space.

### The origin of erroneous MP orientation

The MP data contained a high percentage of forward-reverse reads (i.e. PE) as well as contamination of read pairs that map in the same direction (FF/RR) (Table 2). The high proportion of FR reads in our MP data is most likely explained by contamination with PE reads, which represent non-biotinylated fragments that were not removed during the wash step in library preparation (c.f. mate pair library sample preparation guide). This is supported by the fact that these PE oriented reads have a smaller estimated insert size of around 500 bp (Figure 1, Additional files 4 and 5). The origin of MP reads oriented in FF/RR direction, which make up ~38% of the total MP data, is less obvious. There is no evidence for FF/RR reads containing non-wheat DNA contamination, nor do the FF/RR reads have increased proportions of reads from TE-repetitive DNA. Moreover, since the 2 Kb and 3/5 Kb libraries were produced and sequenced by different labs using different protocols it is highly unlikely that systematic technical errors have been introduced. One possible explanation to the high FF/RR fraction is that they originate from rearranged DNA generated in the multiple displacement amplification (MDA) step, which was used to increase DNA amount after chromosome flow-sorting. It has been shown that MDA generates genomic rearrangement in the amplified DNA with a frequency of 1 rearrangement per 10 Kb, and majority of chimeras are inverted sequences [18]. In a *de-novo* assembly of a single bacteria cell MDA, >50% of the MP pairs were chimeric pairs [33]. Hence, even though MDA has proven to be very useful to prepare DNA from flow sorted chromosomes for single-end and short insert size PE sequencing [20,21,34], the use of MDA DNA in long insert size MP library construction and scaffolding might not be an optimal strategy for wheat genome scaffolding. Another limitation due to the high proportion of FF/RR pairs is that it restricted us from using any type of scaffold-assembler. For example, when trying to integrate MP data using ABySS, the software did not handle the large proportion of the MP reads with non-MP orientation.

## Conclusion

The wheat chromosome 7B was sequenced and assembled using PE reads with an insert size of ~350 bp in combination with 2, 3 and 5 Kb MP libraries. MP integration improved both assembly contiguity and reduced fragmentation of the gene space, but only to a modest extent. Scaffold reliability increased with increasing assembly stringency, emphasizing the need to use high stringency scaffolding parameters to avoid scaffolding errors. Scaffold assemblies of 7BS with reduced MP coverage showed that MP sequence coverage of ~40-50× would be sufficient to produce assemblies with slightly reduced N50 but comparable results for gene space improvement compared to the full coverage assembly (89×). In conclusion, MP assembly improvements was lower than for other recently assembled plant genomes, possibly due to the extreme repeat content of wheat, high fragmentation of contig assemblies, and the use of MDA DNA to construct MP libraries.

## Additional files

**Additional file 1: 7BL_BAC_assemblies.**

**Additional file 2: BAC_contigs_from_50_BACs.**

**Additional file 3: PE_insert_size_distributions.**

**Additional file 4: 7BS_libraries.**

**Additional file 5: 7BL_libraries.**

**Additional file 6: Distribution_of_ff_reads_genus_hits.**

#### Competing interest

The authors declare that they have no competing interests.

#### Authors' contributions

TB carried out bioinformatics on scaffolding and participated in writing the manuscript. BZ estimated insert sizes of MP libraries, participated in data analyses and helped draft the manuscript. JW and MC performed assembly of PE reads, and participated in drafting the manuscript. TA, CB and FP coordinated and carried out the MP sequencing and helped to draft the manuscript. HS- carried out isolation and preparation of MDA DNA, and helped draft the manuscript. JD led the work on flow-sorting of 7B chromosome, and was involved in drafting the manuscript. MK was responsible for the sequencing of one MP library, carried out data analyses and helped draft the manuscript. SL and OAO helped coordinate the study, participated in data analyses, and helped draft the manuscript. SRS carried out bioinformatics analyses of gene content, helped to draft the manuscript, and was responsible for the final version of the manuscript. All authors read and approved the final manuscript.

#### Acknowledgements

The project was funded by grants from the Norwegian Research Council (project no. 199387/I99) and Graminor A/S to Odd-Arne Olsen. Hana Šimková and Jaroslav Doležel were supported by the Czech Science Foundation (award no. P501/12/2554) and by Ministry of Education, Youth and Sports of the Czech Republic and the European Regional Development Fund (Operational Programme Research and Development for Innovations No. ED0007/01/01).

#### Author details

[1]Department of Plant and Environmental Sciences, University of Life Sciences, Ås, Norway. [2]The Genome Analysis Centre (TGAC), Norwich Research Park, Norwich NR4 7UH, UK. [3]Department of Molecular Biology and Genetics, Aarhus University, Forsøgsvej 1, 4200, Slagelse, Denmark. [4]Centre of the Region Haná, Institute of Experimental Botany, 77200, Olomouc, Czech Republic. [5]Centre for Integrative Genetics (CIGENE) and Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, Ås N-1432, Norway. [6]Department of Genetics and Biotechnology, Faculty of Agricultural Sciences, Aarhus University, Tjele 8830, Denmark.

Received: 26 July 2012 Accepted: 22 March 2013
Published: 4 April 2013

#### References

1. Global Perspective Studies Unit - Food and Agriculture Organization of the United Nation: *World agriculturae: towards 2030/2050 - Interim report - Prospects for food, nutrition, agriculture and major commodity groups.* Rome: Food and Agriculture Organization of the United Nation; 2006.
2. The Government Office for Science: *Foresight: The Future of Food and Farming - Final Project Report.* London, United Kingdom: Government Office for Science; 2011.
3. Jannink J-L, Lorenz AJ, Iwata H: **Genomic selection in plant breeding: from theory to practice.** *Brief Funct Genomics* 2010, **9**(2):166–177.
4. Argout X, Salse J, Aury JM, Guiltinan MJ, Droc G, Gouzy J, Allegre M, Chaparro C, Legavre T, Maximova SN, *et al*: **The genome of Theobroma cacao.** *Nat Genet* 2011, **43**(2):101–108.
5. The Potato Sequencing Consortium: **Genome sequence and analysis of the tuber crop potato.** *Nature* 2011, **475**(7355):189–195.
6. Schmutz J, Cannon SB, Schlueter J, Ma JX, Mitros T, Nelson W, Hyten DL, Song QJ, Thelen JJ, Cheng JL, *et al*: **Genome sequence of the palaeopolyploid soybean (vol 463, pg 178, 2010).** *Nature* 2010, **465**(7294):120.
7. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, *et al*: **The B73 Maize Genome: Complexity, Diversity, and Dynamics.** *Science* 2009, **326**(5956):1112–1115.
8. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, *et al*: **The Sorghum bicolor genome and the diversification of grasses.** *Nature* 2009, **457**(7229):551–556.
9. Huang S, Li R, Zhang Z, Li L, Gu X, Fan W, Lucas WJ, Wang X, Xie B, Ni P, *et al*: **The genome of the cucumber, Cucumis sativus L.** *Nat Genet* 2009, **41**(12):1275–1281.
10. Dvorak J, Terlizzi P, Zhang HB, Resta P: **The evolution of polyploid wheats: identification of the A genome donor species.** *Genome* 1993, **36**(1):21–31.
11. Dvorak J, Zhang HB: **Variation in repeated nucleotide sequences sheds light on the phylogeny of the wheat B and G genomes.** *Proc Natl Acad Sci USA* 1990, **87**(24):9640–9644.
12. Doležel J, Kubalakova M, Paux E, Bartos J, Feuillet C: **Chromosome-based genomics in the cereals.** *Chromosome Res* 2007, **15**(1):51–66.
13. Vraná J, Kubalakova M, Simkova H, Cihalikova J, Lysak MA, Dolezel J: **Flow sorting of mitotic chromosomes in common wheat (Triticum aestivum L.).** *Genetics* 2000, **156**(4):2033–2041.
14. Šafář J, Simkova H, Kubalakova M, Cihalikova J, Suchankova P, Bartos J, Dolezel J: **Development of chromosome-specific BAC resources for genomics of bread wheat.** *Cytogenet Genome Res* 2010, **129**(1–3):211–223.
15. Paux E, Sourdille P, Salse J, Saintenac C, Choulet F, Leroy P, Korol A, Michalak M, Kianian S, Spielmeyer W, *et al*: **A physical map of the 1-gigabase bread wheat chromosome 3B.** *Science* 2008, **322**(5898):101–104.
16. Doležel JŠH, Kubaláková M, Šafář J, Suchánková P, Číhalíková J, Bartoš J, Valárik M: **Chromosome genomics in the Triticeae**. In *Genetics and Genomics of the Triticeae.* Edited by Feuillet C, Muehlbauer G. New York: Springer; 2009:285–316.
17. Simková H, Svensson JT, Condamine P, Hribova E, Suchankova P, Bhat PR, Bartos J, Safar J, Close TJ, Dolezel J: **Coupling amplified DNA from flow-sorted chromosomes to high-density SNP mapping in barley.** *BMC Genomics* 2008, **9**:294.
18. Lasken RS, Stockwell TB: **Mechanism of chimera formation during the Multiple Displacement Amplification reaction.** *BMC Biotechnol* 2007, **7**:19.
19. Mayer KF, Martis M, Hedley PE, Simkova H, Liu H, Morris JA, Steuernagel B, Taudien S, Roessner S, Gundlach H, *et al*: **Unlocking the barley genome by chromosomal and comparative genomics.** *Plant Cell* 2011, **23**(4):1249–1263.
20. Hernandez P, Martis M, Dorado G, Pfeifer M, Galvez S, Schaaf S, Jouve N, Simkova H, Valarik M, Dolezel J, *et al*: **Next-generation sequencing and syntenic integration of flow-sorted arms of wheat chromosome 4A**

exposes the chromosome structure and gene content. *Plant J* 2012, **69**(3):377–386.
21. Berkman PJ, Skarshewski A, Lorenc MT, Lai K, Duran C, Ling EY, Stiller J, Smits L, Imelfort M, Manoli S, *et al*: **Sequencing and assembly of low copy and genic regions of isolated Triticum aestivum chromosome arm 7DS.** *Plant Biotechnol J* 2011, **9**(7):768–775.
22. Berkman PJ, Skarshewski A, Manoli S, Lorenc MT, Stiller J, Smits L, Lai K, Campbell E, Kubalakova M, Simkova H, *et al*: **Sequencing wheat chromosome arm 7BS delimits the 7BS/4AL translocation and reveals homoeologous gene conservation.** *Theor Appl Genet* 2012, **124**(3):423–432.
23. Treangen TJ, Salzberg SL: **Repetitive DNA and next-generation sequencing: computational challenges and solutions.** *Nat Rev Genet* 2012, **13**(1):36–46.
24. Choulet F, Wicker T, Rustenholz C, Paux E, Salse J, Leroy P, Schlub S, Le Paslier MC, Magdelenat G, Gonthier C, *et al*: **Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces.** *Plant Cell* 2010, **22**(6):1686–1701.
25. Xu X, Pan S, Cheng S, Zhang B, Mu D, Ni P, Zhang G, Yang S, Li R, Wang J, *et al*: **Genome sequence and analysis of the tuber crop potato.** *Nature* 2011, **475**(7355):189–195.
26. *Mate Pair Library v2 Sample Preparation Guide.* [https://shell.cgrb.oregonstate.edu/sites/default/files/Files/Docs/Illumina/prep/MatePair_v2_2-5kb_SamplePrep_Guide_15008135_A.pdf].
27. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I: **ABySS: a parallel assembler for short read sequence data.** *Genome Res* 2009, **19**(6):1117–1123.
28. Compeau PE, Pevzner PA, Tesler G: **How to apply de Bruijn graphs to genome assembly.** *Nat Biotechnol* 2011, **29**(11):987–991.
29. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25**(14):1754–1760.
30. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**(3):403–410.
31. Li WL, Zhang P, Fellers JP, Friebe B, Gill BS: **Sequence composition, organization, and evolution of the core Triticeae genome.** *Plant J* 2004, **40**(4):500–511.
32. Zhu W, Ouyang S, Iovene M, O'Brien K, Vuong H, Jiang J, Buell CR: **Analysis of 90 Mb of the potato genome reveals conservation of gene structures and order with tomato but divergence in repetitive sequence composition.** *BMC Genomics* 2008, **9**:286.
33. Rodrigue S, Malmstrom RR, Berlin AM, Birren BW, Henn MR, Chisholm SW: **Whole genome amplification and de novo assembly of single bacterial cells.** *PLoS One* 2009, **4**(9):e6864.
34. Vitulo N, Albiero A, Forcato C, Campagna D, Dal Pero F, Bagnaresi P, Colaiacovo M, Faccioli P, Lamontanara A, Simkova H, *et al*: **First Survey of the Wheat Chromosome 5A Composition through a Next Generation Sequencing Approach.** *PLoS One* 2011, **6**(10):e26421.

doi:10.1186/1471-2164-14-222
**Cite this article as:** Belova *et al.*: Integration of mate pair sequences to improve shotgun assemblies of flow-sorted chromosome arms of hexaploid wheat. *BMC Genomics* 2013 **14**:222.

# Paper II

Theor Appl Genet (2014) 127:2029–2040
DOI 10.1007/s00122-014-2358-z

ORIGINAL PAPER

# Utilization of deletion bins to anchor and order sequences along the wheat 7B chromosome

**Tatiana Belova · Lars Grønvold · Ajay Kumar · Shahryar Kianian · Xinyao He · Morten Lillemo · Nathan M. Springer · Sigbjørn Lien · Odd-Arne Olsen · Simen R. Sandve**

Received: 22 March 2014 / Accepted: 13 July 2014 / Published online: 19 August 2014


**Abstract**

***Key message*** **A total of 3,671 sequence contigs and scaffolds were mapped to deletion bins on wheat chromosome 7B providing a foundation for developing high-resolution integrated physical map for this chromosome.**

*Abstract* Bread wheat (*Triticum aestivum* L.) has a large, complex and highly repetitive genome which is challenging to assemble into high quality pseudo-chromosomes. As part of the international effort to sequence the hexaploid bread wheat genome by the international wheat genome sequencing consortium (IWGSC) we are focused on assembling a reference sequence for chromosome 7B. The successful completion of the reference chromosome sequence is highly dependent on the integration of genetic and physical maps. To aid the integration of these two types of maps, we have constructed a high-density deletion bin map of chromosome 7B. Using the 270 K Nimblegen comparative genomic hybridization (CGH) array on a set of cv. Chinese spring deletion lines, a total of 3,671 sequence contigs and scaffolds (~7.8 % of chromosome 7B physical length) were mapped into nine deletion bins. Our method of genotyping deletions on chromosome 7B relied on a model-based clustering algorithm (Mclust) to accurately predict the presence or absence of a given genomic sequence in a deletion line. The bin mapping results were validated using three different approaches, viz. (a) PCR-based amplification of randomly selected bin mapped sequences (b) comparison with previously mapped ESTs and (c) comparison with a 7B genetic map developed

Communicated by Hong-Qing Ling.

**Electronic supplementary material** The online version of this article (doi:10.1007/s00122-014-2358-z) contains supplementary material, which is available to authorized users.

T. Belova · L. Grønvold · M. Lillemo · O.-A. Olsen · S. R. Sandve (✉)
Department of Plant Sciences, Norwegian University of Life Sciences, Ås, Norway
e-mail: simen.sandve@nmbu.no

T. Belova
e-mail: tatiana.belova@nmbu.no

L. Grønvold
e-mail: lars.gronvold@nmbu.no

M. Lillemo
e-mail: morten.lillemo@nmbu.no

O.-A. Olsen
e-mail: odd-arne.olsen@nmbu.no

A. Kumar · S. Kianian
Department of Plant Sciences, North Dakota State University, Fargo, ND, USA
e-mail: Ajay.Kumar.2@ndsu.edu

S. Kianian
e-mail: S.Kianian@ndsu.edu

X. He
International Maize and Wheat Improvement Center (CIMMYT), Apdo.Postal 6-641, 06600 Mexico, DF, Mexico
e-mail: x.he@cgiar.org

N. M. Springer
Department of Plant Biology, Microbial and Plant Genomics Institute, University of Minnesota, Saint Paul, MN 55108, USA
e-mail: springer@umn.edu

S. Lien
Centre for Integrative Genetics (CIGENE), Norwegian University of Life Sciences, Ås, Norway
e-mail: sigbjorn.lien@nmbu.no

in the present study. Validation of the bin mapping results suggested a high accuracy of the assignment of 7B sequence contigs and scaffolds to the 7B deletion bins.

**Abbreviations**

| | |
|---|---|
| CGH | Comparative genomic hybridization |
| IWGSC | International wheat genome sequencing consortium |
| CSS | Chromosome survey sequencing |
| cv CS | Cultivar chinese spring |
| LDN | Langdon |
| LDN-DS 7D(7B) | Langdon 7B substitution line |
| ISBP | Insertion site-based polymorphism |
| RG | Random genomic probes |
| FL | Fraction length |
| RIL | Recombinant inbred lines |

## Introduction

The recent chromosome survey sequence (CSS) assembly of the hexaploid bread wheat genome (*Triticum aestivum* L.; $2n = 6 \times = 42$; AABBDD) by the international wheat genome sequencing consortium (IWGSC) (IWGSC, 2014) serves as an important first step towards a wheat reference genome sequence (www.wheatgenome.org). This chromosome-specific assembly allow for a deeper understanding of the wheat genome composition, organization, and evolution, as well as providing a resource for future research- and breeding efforts. However, due to the large chromosome size and extreme repeat content (>80 %), the wheat chromosome sequence assemblies are highly fragmented compared to for example barley (The International Barley Genome Sequencing Consortium 2012), rice (The International Rice Genome Sequencing Consortium 2005), potato (The Potato Genome Sequencing Consortium 2011) and sorghum (Paterson et al. 2009).

In order to move towards a complete genome assembly, physical contigs and scaffolds must be integrated with genetic maps at high density and high resolution. A major constraint for the genetic mapping in wheat is the non-uniform distribution of recombination events along the chromosomes, with recombination rates dropping dramatically towards the centromere (Devos et al. 1995; Werner et al. 1992; Akhunov et al. 2003). For instance, detailed analyses of recombination frequencies in bread wheat chromosome 3B show that 90 % of crossing overs occur in only 40 % of the chromosome (Saintenac et al. 2009). The same study also observed >85-fold differences for crossover frequency per physical distance (cM/Mb) for a centrometic bin (C-3BS1-0.33) compared to a sub-telomeric bin (3BS8-0.78-0.87) on chromosome 3B. This "recombination stiffness" makes it very difficult to place and order sequence contigs along a chromosome. One approach has therefore been to combine several independent and complementary mapping approaches with meiotic mapping, such as synteny-based mapping using closely related species and deletion bin mapping (e.g. 3B and 1BL) (Paux et al. 2008; Philippe et al. 2013). Although synteny-based mapping approaches can be powerful, inversions and translocations of genes and gene blocks in wheat relative to other grass genomes (like *Brachypodium*, rice and sorghum) is common (Kumar et al. 2012). Synteny-based mapping is therefore more reliable within smaller chromosomal blocks. Hence, assigning sequence contigs to smaller bins along the chromosome is of high value for the downstream synteny-based sequence ordering, but also an important source for independent verification of the meiotic mapping results.

Deletion bin mapping is a recombination independent mapping strategy and involves the use of a series of overlapping deletions to map markers to relatively short (range 20–155 Mb in size) chromosomal segments (deletion bins) (Qi et al. 2004). In bread wheat, aneuploid stocks have been extensively used to assign markers to chromosomes, chromosome arms, and bins within chromosome arms. Sears (1954) was the first to study and produce bread wheat aneuploids (cv. Chinese spring, CS), including 21 monosomics, 21 nullisomics and 21 tetrasomics (Sears 1954). In addition, more recently, using gametocidal genes to induce chromosome breaks, a set of 436 terminal chromosome deletions were identified in hexaploid wheat (Endo and Gill 1996). Later, using set of wheat aneuploids and deletion stocks, 16,000 ESTs were bin mapped (Qi et al. 2004), of which 549 ESTs (corresponding to ~0.08 % of the chromosome 7B physical length) were assigned to six bins on chromosome 7B (Hossain et al. 2004).

In the present study we describe the development of a high-density deletion bin map of wheat chromosome 7B, placing ~7.8 % of the chromosome 7B physical length into nine bins using Nimblegen comparative genome hybridization (CGH). In addition, an $F_6$ recombinant inbred line (RIL) population containing 131 lines was assayed with the 90 K iSelect SNP chip (Wang et al. 2014), resulting in incorporation of 629 SNP markers into the 7B genetic map. This work is part of the IWGSC Norwegian 7B sequencing project and aid in anchoring and ordering of physical sequence contigs from MTP (Minimal Tiling Path) BAC sequencing, a critical step towards a complete 7B reference sequence.

## Materials and methods

### Oligonucleotide probe design

Two types of oligonucleotide probes were extracted from the shotgun sequence assembly of chromosome 7B (Belova

et al. 2013): random genomic (RG) probes and insertion site-based polymorphism (ISBP) probes (Fig. 1). In order to develop the RG probes, assemblies were first masked for repeats with RepeatMasker (Smit et al. 1996–2010) against an in-house repeat content database [TREP ten combined with the repeats identified in Choulet et al. (2010)]. Masked contigs were fragmented in non-overlapping sequences of 50 bp located ≥50 bp apart. ISBP finder (Paux et al. 2010) was used to identify ISBP sites with high and medium confidence levels from which sequences of 50 bp, 25 bp from each side of the junction, were selected as ISBP probes.

Subsequent to the identification of RG and ISBP probes, we used BLASTN (Altschul et al. 1990) to identify and remove probe sequences with high similarity (hit length >45 bp and identity >95 %) to contigs in the 7A and 7D assemblies (IWGSC; http://www.wheatgenome.org/). Probes carrying homopolymers longer than 8 bp were excluded from the analysis. We also excluded probes that did not pass the 'Cycle script' designed by Nimblegen or had a calculated oligonucleotide melting temperature outside the 66–86 °C range. A collection of wheat ESTs (Lazo et al. 2004) was used to design random control probes (50 bp long) that were not overrepresented with 7B sequences.

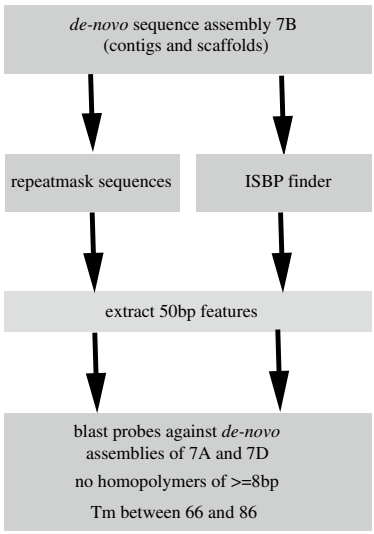


**Fig. 1** Overview of the CGH Nimblegen probe design pipeline

**Table 1** Set of deletion lines with their corresponding fragment length (FL), showing the percent of the chromosome arm present

| Deletion stock | Fragment length | Nomenclature |
|---|---|---|
| Del7BS-2 | 0.27 | FL-0.27 |
| Del7BL-14 | 0.14 | FL-0.14 |
| Del7BL-2 | 0.33 | FL-0.33 |
| Del7BL-1 | 0.40 | FL-0.40 |
| Del7BL-9 | 0.45 | FL-0.45 |
| Del7BL-7Del1DS-3 | 0.63 | FL-0.63 |
| Del7BL-5 | 0.69 | FL-0.69 |
| Del7BL-13 | 0.79 | FL-0.79 |
| Del7BL-3 | 0.86 | FL-0.86 |

## Plant material

Two tetraploid wheat lines, Langdon (LDN; AABB; $2n = 4\times = 28{:}13'' + 7B''$) and Langdon chromosome substitution line (LDN-DS 7D (7B), $2n = 4\times = 28{:}13'' + 7D''$, in which chromosomes 7B is substituted by 7D chromosomes of the hexaploid cultivar CS (Joppa and Williams 1977), were used initially for screening and identification of 7B specific probes and later as reference samples to estimate absence/presence (i.e. *M*-values) of probes in CS 7B deletion stocks (see sections below for details).

Among the deletion stocks of the hexaploid wheat cultivar CS (*T. aestivum*) (Endo and Gill 1996), lines with terminal deletions in chromosome 7B and its ditelosomic lines (DT7BL and DT7BS) were used in the CGH assays. Details of 7B deletion stocks used in this study are provided in Table 1. The fraction length (FL) reflects the position of the breakpoint from the centromere relative to the length of the complete arm. Seeds for deletion lines were kindly provided by Dr. Bikram S. Gill, Department of Plant Pathology, Kansas State University, Manhattan, KS, USA.

## CGH sample preparation and hybridization

DNA from leaf tissue was isolated by the CTAB method (Springer 2010). Labeling and hybridization of samples were performed according to the Nimblegen protocol. Half a µg DNA of each sample was labeled using either Cy3 or Cy5-labeled Random Nonamers. Samples were denaturated at 98 °C for 10 min and chilled on ice for 2 min. The DNA was incubated for 2 h at 37 °C with 100 units Klenow Fragment (5′-3′ exo-) and dNTP mix (10 mM each). After adding stop solution (0.5 M EDTA), samples were precipitated with NaCl and isopropanol and centrifuged at 12,000 × *g* for 10 min. The pellets were re-suspended in 25 µl of $H_20$. Twenty µg of Cy3 and Cy5 labeled samples were mixed in a 1.5 ml tube and dried in a vacuum concentrator on low heat. Each sample pair was then re-suspended in unique sample tracking control and added to 8.7 µl of the hybridization solution mix. Tubes were first incubated at 95 °C for 5 min, and then at 42 °C for 5 min. Samples were hybridized to CGH array for 60–72 h at 42 °C. Slides were washed and immediately scanned using the MS 200

microarray scanner according to the array manufacturer's protocol. Probe fluorescence intensities were extracted with the NimbleScan 2.1 software. Raw data was normalized by two-dimensional loess spatial normalization followed by M-A loess normalization for each sample comparison using the control probes as training set (GEO submission GSE57461).

### Selection of 7B chromosome specific probes

In order to select a subset of 7B specific probes as well as a set of control probes which do not hybridize to the 7B genomic sequence we first performed CGH between two tetraploid Langdon wheat lines that differ only by the presence of the 7B chromosome in the genome (LDN contains 7B, while LDN-DS 7D (7B) lacks 7B). The experiment was carried out with a 3*720 K CGH microarray using a dye swap design where each sample was labeled with both Cy5 and Cy3. The selected set of 7B specific probes and control probes was then printed on a 12*270 K CGH chip (Roche, NimbleGen Inc.) and hybridized with CS deletion lines.

### Genotyping presence absence variation in CS deletion lines

The CS deletion lines have various sized terminal overlapping deletions, usually >10 % of the chromosome arm (Endo and Gill 1996). The distribution of $\log_2$ ratios of hybridization signal intensities between deletion lines and wild type (referred to as *M*-values) is therefore expected to be a combination of two underlying distributions, representing probes being deleted (i.e. absent) and those that are present (Fig. 2).

To determine the probability for a probe to belong either to the "present" or "absent" classes, we used Gaussian mixture model clustering [$P_{abs}$, $(1 - P_{abs}) = P_{pres}$] as implemented in the R package *Mclust* (Fraley and Raftery 2007). The parameter '*G*' (number of groups) was fixed to 2, while all other parameters were estimated by the Mclust software. The number of absent probes for each deletion line was estimated by intersecting results from two different Mclust analyses using different LDN line hybridizations as a reference. A sequence was only assigned as absent or present if both Mclust analyses supported the same classification with ≥80 % probability. Probes that did not meet this criterion were assigned to the NA class (i.e. not possible to classify). Long *M*-value distribution tails in combination with a limited separation of the absent and present distribution peaks sometimes lead to erroneously assignment of probes with high *M*-values to the absent class. Probes with *M*-values higher than the mean *M*-value in the present class were therefore given a probability of 0 for belonging to the absent class.

### Assignment of sequences to deletion bins

Assignment of sequences to deletion bins was based on a two-step strategy using the absence/presence classifications from each deletion line. In the first step, we compared each deletion line (X) to the deletion line with an incrementally smaller deletion (Y) and identified sequences present in Y but absent in X. To assign sequences to the most distal deletion bins on the two 7B arms, we compared lines with the smallest deletions to ditelosomic lines carrying a complete copy of that chromosome arm. In the second step, we used this initial bin assignment and confirmed that each bin mapped sequence was present in all other deletion lines with smaller deletions.

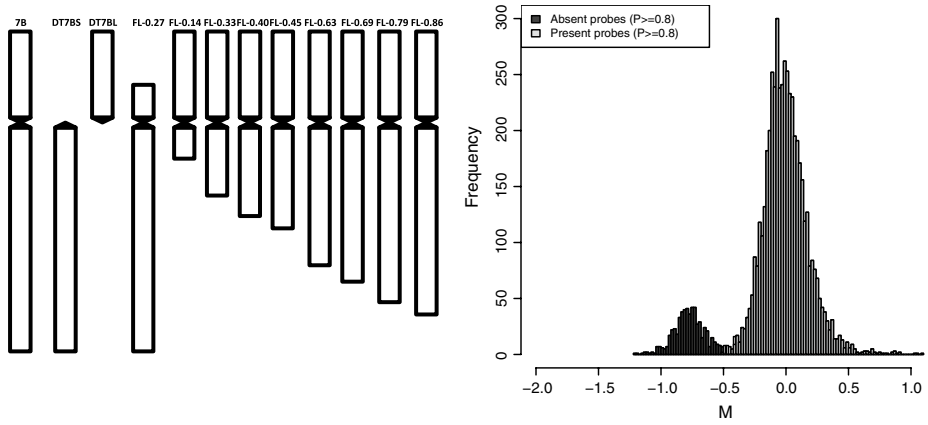


**Fig. 2** Schematic explanation of the two groups of probes (present and absent) when comparing *deletion line* and *reference line*

### SNP-based genetic map of chromosome 7B

A mapping population of 131 RILs was developed from a cross between the CIMMYT breeding line 'SABUF/5/BCN/4/RABI//GS/CRA/3/AE.SQUARROSA (190)' (selection history CASS94Y00042S-32PR-1B-0 M-0Y) and the German spring wheat cv. 'Naxos' (pedigree Tordo/St.Mir808-Bastion//Miranet). The population was advanced from $F_2$ to $F_6$ through the single seed descent (SSD) method. DNA was extracted from $F_6$ plants using DNeasy plant DNA extraction kit (Qiagen). The population was genotyped with the iSelect 90 K wheat chip from Illumina, which contains a total of 81,587 SNP markers (Wang et al. 2014). Genotypes were called using Genome Studio V2011.1. Due to the hexaploid genome constitution of wheat, the automatic clustering algorithm identified only 3,117 polymorphic markers that fit the expected segregation ratio for a diploid locus in the $F_6$ population. An additional 7,255 polymorphic markers with skewed clustering patterns due to signal noise from the other two genomes were called manually giving a total of 10,372 SNP markers for further analysis. Genetic linkage groups were created using the program MST map (Wu et al. 2008) with a cutoff $p$ value of 1e-6, maximum distance of 15 cM between markers, minimum size of linkage group being 2 cM. MST map linkage groups were then assigned to chromosomes based on the BLASTN results of SNP sequences against survey sequences of A-, B- and D genomes (IWGSC data repository at http://wheat-urgi.versailles.inra.fr/). Only markers giving a hit to a single chromosome with ≥99 % sequence identity and 100 % coverage were assigned to a chromosome. Finally, the JoinMap v.4 Maximum Likelihood algorithm was used to estimate marker order for the 7B linkage group.

### Verification of the CGH bin mapping results

We used three independent methods to verify the bin mapping approach and estimate the error rate. In the first approach, we performed PCR-based verification of the mapped sequences. Primers were designed from bin mapped sequences with ISBP finder, tested for 7B specificity using the 7B CS ditelosomic lines and then used for PCR amplification in CS deletion lines to identify the bin location of the markers. PCR reactions were carried out in 10 µl total reaction using 60 ng of genomic DNA containing 1 µl 10 × PCR buffer, 0.2 µl 10 mM dNTPs and 0.1 µl of 5 units/µl of AmpliTaq DNA polymerase (Applied Biosystems). The PCR conditions used were as follows: 94 °C for 3 min, 45 cycles of: 45 s at 94 °C, 45 s at 59 °C, 90 s at 72 °C, followed by a final extension at 72 °C for 10 min. The PCR products were separated on a 1.5 % agarose gel and visualized using ethidium bromide staining.

In the second verification approach, we took advantage of the fact that some of the bin mapped sequences in this study have sequence homology with the previously bin mapped ESTs (Hossain et al. 2004). The sequences that we bin map in the present study were used in a BLASTN search against sequences of the previously 7B bin mapped ESTs. BLAST hits were filtered based on ≥99 % identity and 100 % coverage. Redundant ESTs were not considered in this analysis. Redundant ESTs are defined as ESTs giving a hit to the same bin mapped sequence with the identical start and end position, identical mismatches, identical gap length and identical hit length.

The final validation of deletion bin mapping results was done by integrating genetically mapped SNPs into the deletion bin map. In order to assign SNP markers to deletion bins, BLASTN search of SNP sequences against bin mapped sequences was performed. Only hits with ≥99 % identity and 100 % coverage of the marker locus were considered in this study.

### Distribution of genes along deletion bins

The 7B gene calls from the wheat CSS (IWGSC data repository at http://wheat-urgi.versailles.inra.fr/Seq-Repository/Genes-annotations) were used in a BLASTX (Altschul et al. 1990) search to estimate gene content of the bin mapped sequences. BLAST result filtering was carried out in the following way: (a) Only query hits with a minimum sequence identity of ≥99 % and a minimum length of 30 amino acid were considered in the analyses (b) Duplicated gene hits in one scaffold were removed from the analyses. Duplicated hits were defined as hits belonging to the same gene ID. The gene density in a bin was calculated by dividing the number of gene hits with the total scaffold length in that bin.

## Results

### CGH and selection of chromosome 7B specific CGH probes

In order to identify probe sequences which detect presence/absence variation (PAV) between LDN and LDN 7D (7B) genotypes we conducted a pilot experiment using a 720 K CGH microarray chip. The $M$-values of LDN versus LDN-DS 7D (7B) comparisons, was used as probe selection criterium. Probes with large difference in hybridization intensity ($M$-values >0.35) and high reproducibility between replicates were classified as chromosome 7B specific. Non-polymorphic control probes were selected from the subset of probes with an $M$-value close to zero ($-0.02 < M < +0.02$). From this experiment, a set

of 49,500 7B probes (11 % ISBP and 89 % RG probes) and 18,000 control probes were selected and printed on a 12*270 K CGH chip with each probe replicated four times per chip. Using BLASTN against 7B IWGSC gene calls we estimate that 0.9 % of RG probes on the 270 K CGH chip are derived from coding genes. This is comparable to the total percentage of coding sequence in the 7B CSS assembly (0.7 %). Low quality CGH hybridizations were excluded from the dataset based on the experimental metrics reports (NimbleScan 2.1 software). In total, we hybridized 17 CS cytogenetic stocks out of which 11 yielded high quality CGH results and were used for the bin mapping (Table 1).

### Effect of probe type on *M*-value distribution

In hexaploid wheat, ISBP markers have provided high level of sub-genome specificity compared to DNA probes designed from the coding regions (Choulet et al. 2010). Generally, probes will have a better signal to background ratio when there is less cross hybridization to other regions of the genome. In order to investigate the relationship between the type of the probe and its hybridization properties, we first generated ten *M*-value distributions between different deletion lines, calculated the proportion of ISBP and RG probes in the 10 % lowest range of $\log_2$ distributions, and then compared this with the total proportions of ISBP and RG probes on the array. Mean proportion of ISBP probes in the lowest range of $\log_2$ distributions were 10 % (range 7–14 %) (data not shown), comparable to the proportion of ISBPs among the total number of the probes (11 %) indicating similar hybridization properties of both types of probes.

### Effect of combining signals from multiple probes in presence/absence genotyping

Our approach to genotyping presence/absence variation on chromosome 7B relies on the ability to accurately predict which sub-distribution of *M*-values a particular probe belongs to (Fig. 2). Thus, a good separation of the two underlying *M*-values distributions is expected to result in more robust probe classification. We have used 7B assembly sequence to design 7B specific probes. In many cases multiple probes were derived from the same contig or scaffold (combination of several contigs). In total, we designed 49,500 probes from 33,286 contigs, giving an average of 1.49 probes per contig (range 1–17). Furthermore, many contigs belong to larger scaffolds (1.43 contigs per scaffold, range 1–18). We therefore investigated the relationship between the number of probes used to estimate *M*-values and the peak separation in the bimodal *M*-value distribution. This showed that the separation between *M*-value distributions significantly improved as the number of probes used for *M*-value calculation increased (Fig. 3a). However, merging signals from several probes comes at a cost, since it leads to fewer data points. We chose to conduct all following analyses using *M*-values from sequences containing at least 3 probes, giving a total of 5,577 sequences for the analysis. The ratio of sequences derived from 7BS to 7BL was 40:60 % (2267/3310) and a total of 5,177 probe sequences could be classified as either absent or present in at least one deletion line. Ninety-two percent of the sequences were assigned to a class in at least 50 % of all CGH analyses of deletion lines. Among these, a strong positive correlation was observed between the number of probe observations and the power to classify a sequence as present/absent (Fig. 3b). Thus, the parameters chosen represent a reasonable trade-off between the number of probes used per sequence for presence/absence calling and mapping accuracy.

### Determining PAV in the deletion lines

The estimated number of deleted sequences in the CS deletion lines is shown in Table 2. With a few exceptions, our results fit well with the expected deletion sizes based on cytological evidence (Endo and Gill 1996). Mean discrepancy between expected and observed deletion size was 5.2 %. Three lines with deletions in the 7BL (FL-0.86, FL-0.79 and FL-0.69) showed significant deviations from the expectations. The FL-0.86 had a 7 % higher proportion of absent probes than expected, while FL-0.79 and FL-0.69 had fewer absent probes than expected (Table 2).

To further describe the nature of deletions in FL-0.86, FL-0.79 and FL-0.69 lines, we plotted $\log_2$ distributions of FL-0.86 vs. FL-0.79, FL-0.79 vs. FL-0.69, and FL-0.86 vs. FL-0.69 (Fig. 4). If FL-0.86, FL-0.79 and FL-0.69 represent incrementally larger deletions, we expect set of *M*-values to occur on the lower right side of the plot (i.e. present probes in the smaller deletion which are absent in the larger deletion). From the plots it appears that FL-0.79 has presence and absence variation relative to both FL-0.86 and FL-0.69. Moreover, FL-0.86 and FL-0.69 have virtually identical *M*-values across the 7BL sequences (Fig. 4). This result shows that cytological estimation of the deletion length and type in the FL-0.86, FL-0.79 and FL-0.69 lines-most likely is wrong.

### Bin mapping of 7B sequences

In total we bin mapped ~74,130 Kbp (3,671 sequence contigs and scaffolds), representing ~7.8 % of 7B chromosome sequence. Using the 7B gene models generated in the CSS project we estimated the gene density (genes/Kbp) along

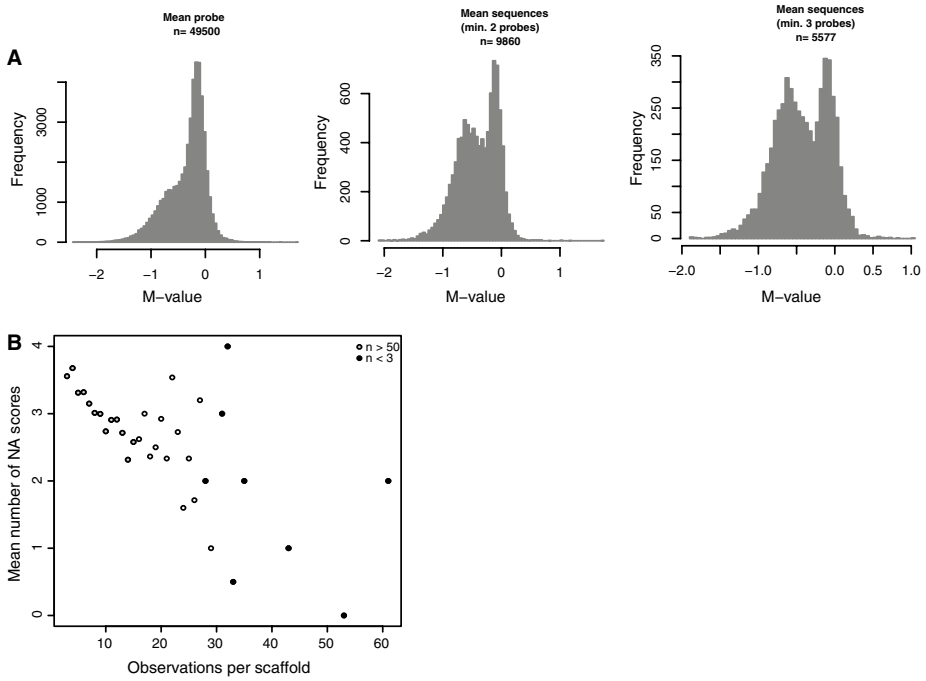


**Fig. 3** **a** *M*-value distributions with *M*-values averaged per probe and per sequence containing 2 and 3 probes. **b** Relationship between the propensity of sequences to be assigned a class (present or absent) and the number of probes contained in sequences

**Table 2** Deleted sequences in 7B deletion lines

| Arm | Deletion line | Expected proportion deleted | Absent probes | | |
|---|---|---|---|---|---|
| | | | 7BS | 7BL | Proportion[a] |
| 7BS | DT7BS | – | 1,680 | 7 | – |
| | FL-0.27 | 0.73 | 1,294 | 4 | 0.57 |
| 7BL | FL-0.14 | 0.86 | 1 | 2,227 | 0.67 |
| | FL-0.33 | 0.67 | 2 | 2,257 | 0.68 |
| | FL-0.40 | 0.6 | 1 | 1,855 | 0.56 |
| | FL-0.45 | 0.55 | 0 | 1,770 | 0.53 |
| | FL-0.63 | 0.37 | 2 | 1,453 | 0.44 |
| | FL-0.69 | 0.31 | 0 | 658 | 0.20 |
| | FL-0.79 | 0.21 | 0 | 563 | 0.17 |
| | FL-0.86 | 0.14 | 8 | 709 | 0.21 |
| | DT7BL | – | 7 | 2,902 | – |

[a] Proportion is calculated based on predicted absent scaffolds on the correct arm only

the bins on 7B (Table 3). Gene density was distributed unevenly along the 7B chromosome with an increase in gene density from the centromere to the telomere (from 0.01 to 0.02 for 7BL and from 0.01 to 0.02 for 7BS). The average gene density for the centromeric region (bins 7BL_0 - 0.14 and 7BS_0 - 0.27) was 1 gene per 107 Kb. The gene density increased by ~2-fold for distal bins.

In constructing a precise bin map, single terminal deletions are preferred over multiple or interstitial deletions (Hohmann et al. 1995). Since our hybridization data support an aberrant nature of deletions types (i.e. not single terminal deletions) in the FL-0.69, FL-0.79 and FL-0.86 lines, all sequences deleted in any of these lines were grouped into one pseudo bin (7BL_0.69* − 1.00). For each of these three lines bin mapping was first performed by comparison with ditelosomic line carrying a complete 7BL arm. Three hundred and nine, 283 and 351 sequences were mapped to 7BL_0.69 − 1.00, 7BL_0.79 − 1.00 and 7BL_0.86 − 1.00, respectively. Among them, 253 sequences (4,429 Kbp) were mapped to all three bins. Bin 7BL_0.79 − 1.00 had 25 unique mapped sequences compared to 7BL_0.86 − 1.00. Eight out of 25 were mapped to the bin BL_0.69 − 1.00 as well. Bins 7BL_0.69 − 1.00 and 7BL_0.86 − 1.00 shared additional 63 mapped sequences. 5, 17 and 30 sequences were uniquely mapped to 7BL_0.69 − 1.00, 7BL_0.79 − 1.00 and 7BL_0.86 − 1.00. In total, 381 sequences were mapped to 7BL_0.69* − 1.00.

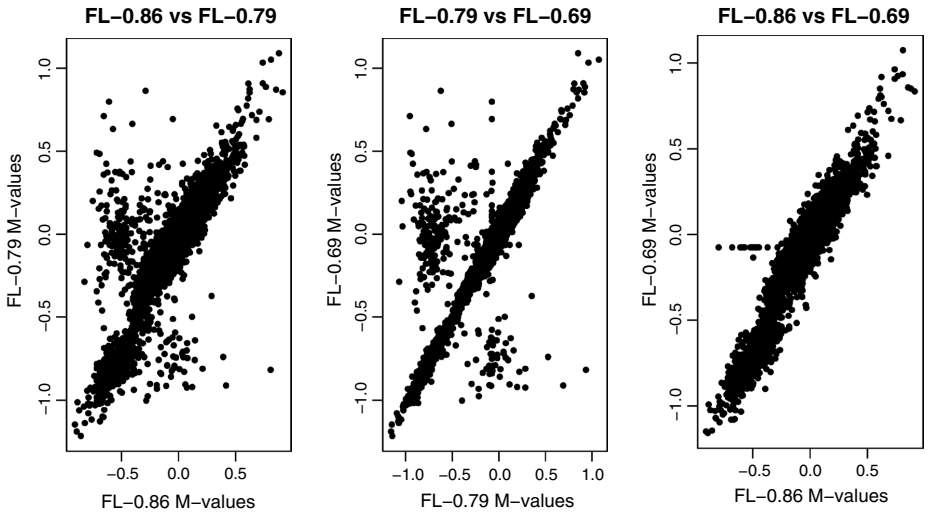


**Fig. 4** *M*-value correlations between different deletions lines

**Table 3** Bin mapping of sequences with proportion of gene hits in deletion bins

| Arm | Bin | No of sequences | | Mapped Kbp | Gene hits per Kbp | Number of unique gene hits |
|---|---|---|---|---|---|---|
| | | 7BS | 7BL | | | |
| 7BS | BS | 1,643 | 0 | | | |
| | BS_0.27 − 1.00 | 1,262 | 0 | 26,807 | 0.023 | 607 |
| | BS_0 − 0.27 | 211 | 0 | 5,318 | 0.011 | 60 |
| 7BL | BL_0 − 0.14 | 1 | 407 | 8,222 | 0.010 | 83 |
| | BL_0.14 − 0.33 | 0 | 64 | 1,118 | 0.016 | 18 |
| | BL_0.33 − 0.4 | 0 | 233 | 4,494 | 0.019 | 85 |
| | BL_0.4 − 0.45 | 0 | 127 | 2,268 | 0.020 | 46 |
| | BL_0.45 − 0.63 | 0 | 312 | 6,420 | 0.022 | 138 |
| | BL_0.63 - 0.69* | 0 | 674 | 12,909 | 0.018 | 241 |
| | BL_0.69* - 1.00 | 0 | 381 | 6,574 | 0.021 | 136 |
| | BL | 0 | 1,906 | | | |

BL_0.69* - 1.00 contains sequences mapped to BL_0.69 − 1.00, BL_0.79 - 1.00 and BL_0.86 - 1.00

Validation of the bin mapping results

In order to verify the accuracy of our bin mapping results we used three different approaches. First, PCR-based ISBP markers were designed from randomly selected bin mapped 7B sequences and used to validate the bin assignment. Out of 12 markers screened, ten were absent in relevant deletion lines. Two markers were mapped to all deletion lines tested, however, there were clear band intensity difference (Supplementary material 1). In the second validation approach we compared our bin mapping results with previously mapped ESTs (Hossain et al. 2004). Seventy-one percent of the sequences were assigned to the same bin as reported earlier (Hossain et al. 2004). The remaining 29 % of sequences (20 sequences out of 69 tested) were mapped to different deletion bin than previously reported. In order to determine if this discrepancy represented error in our CGH-based bin mapping, PCR-based ISBP markers were designed from sequence scaffolds from which the mapped ESTs were derived (determined by BLASTN). Twelve out of the 20 ISBP markers were 7B specific and used in PCR reactions with CS deletion lines. Eleven of these assays supported our CGH-based results, while only one supported the results from Hossain et al. (2004) (Fig. 5). The third approach to confirm the results of CGH-based bin mapping involved a comparison of the positions of marker sequences of bin maps with a genetic map. A RIL mapping population was used to construct genetic map of

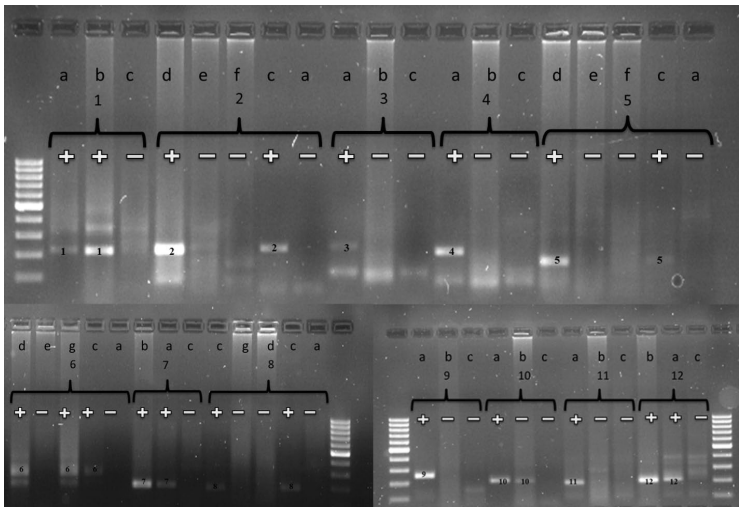

**Fig. 5** Agarose gel profile of PCR products amplified by ISBP primers designed from sequences assigned to deletion bins. *Deletion lines* used are: *a*-DT7BS, *b*-FL-0.27, *c*-DT7BL, *d*-FL-0.63, *e*-FL-0.45, *f*-FL-0.33, *g*- FL-0.79. *Numbers* indicate the primer pairs (supplementary material 3). Plus and minus indicate expected presence or absence of bands in the corresponding CS deletion lines according to the deletion bin mapping results

chromosome 7B consisting of 629 SNP markers representing 225 unique loci (total map length 180 cM). A total of 116 markers on the genetic map could be assigned to 70 sequences mapped across five deletion bins of the wheat 7B (Fig. 6). An almost perfect relationship between the order of markers on the genetic map and the order of deletion bin assignment was identified by visual inspection. The only exception was marker wsnp_BE443010B_Ta_2_2 (see dotted line with aestrix in Fig. 6), which was mapped to the deletion bin BL_0.45 − 0.63 (Supplementary material 2), while the surrounding markers were assigned to BL_0.63 − 0.69*. Therefore, the comparative analysis with previously bin mapped ESTs, 7B genetic map, and PCR verification suggest high reliability of our approach of the contig and scaffold sequence assignation to deletion bins.

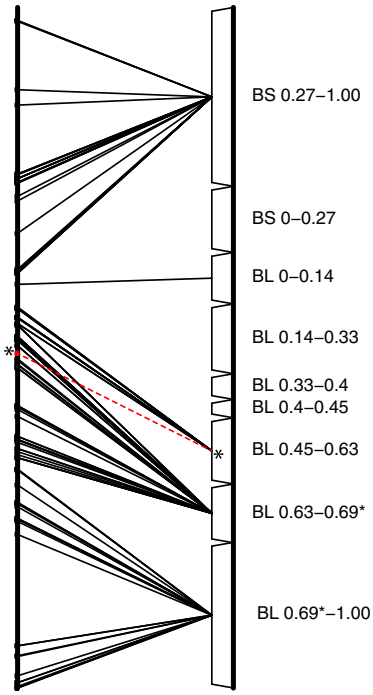


**Fig. 6** Comparison of the 7B genetic map and deletion bin map

## Discussion

The development of a high quality and high resolution integrated physical and genetic map for the allohexaploid genome of bread wheat represents a significant challenge. Various mapping data, including recombination based mapping, radiation hybrid mapping, synteny-based mapping and deletion bin mapping is deemed necessary to anchor and order physical contigs (Paux et al. 2008; Philippe et al. 2013). In this study a high-density deletion bin map of chromosome 7B placing ~7.8 % of the 7B sequence into nine deletion bins was developed using high-throughput Nimblegen CGH microarray platform. This work represents an important step towards a physically ordered 7B reference sequence. In addition, high-density deletion bin map of 7B may serve as an excellent resource of new markers for fine mapping and map based cloning of genes/QTL located on chromosome 7B.

### CGH design, PAV genotyping, and error rates

Deletion calling in a hexaploid genome with hybridization based methods is inherently difficult due to background signals from highly similar homeologous DNA sequences. It is therefore critical to maximize the specificity of probe hybridization. As a first step we used the CSS assemblies to perform a BLAST-based filtering step to remove probe sequences with high sequence similarity with 7A and 7D chromosomes. Furthermore, we included different classes of genomic probes on the CGH array with potential differences in sub-genome specificity. Interestingly, the ISBP-based probes known to be highly sub-genome specific (Choulet et al. 2010) did not differ significantly in hybridization properties compared to RG probes. Thus both probe types can be successfully used in CGH experiments in hexaploid wheat.

*Mclust* method was used to classify probes based on their *M*-values as a DNA sequence that was either "present" or "absent" (i.e. deleted). It was evident that this method had low power when applied to single probes (i.e. 50 bp replicated four times per CGH array) (Fig. 3). However, when multiple co-localized probes were used to estimate *M*-values for whole sequence contigs and/or scaffolds we could successfully assign DNA sequences to chromosome arms and deletion bins (Table 3) with low error rates. Frequencies of incorrect assignment of sequences to chromosome arms were very low, but slightly higher when we used two (0.9 %) compared to three probes per estimated *M*-value (0.03 %) (data not shown). Furthermore, three different CGH-independent verification approaches also supported a high level of accuracy for bin assignment. In the PCR-based validation experiments, 83 % of the markers were unambiguously mapped to the expected deletion bins. The remaining markers mapped to all deletion lines tested, however, a clear band intensity difference between deletion lines indicates a difference in DNA content (i.e. deletion) as predicted from the CGH results (Supplementary material 1). Next, a comparison between CGH results and previous bin mapped ESTs revealed 29 % discordance; however, additional PCR based assays revealed that only 8 % of this discordance suggested error in our CGH results. The comparison with previous EST bin mapping therefore suggests an error rate of approximately 2.5 % ($0.29 \times 0.08 = 0.023$). Another validation of the accuracy of our results was provided by comparison with a genetic map of 7B. The comparison of sequence order from a genetic map with the bin map showed that only 1 out of 116 (0.8 %) sequences in the genetic map did not concur with the bin map order. In conclusion, the results from validation experiment suggest an error rate of <2.5 %.

### Discrepancy between cytological and genetic estimates of deletion sizes

A strong overall correlation between the observed and expected proportion of deleted probes was found for the aneuploid lines (Table 2). However, three deletion lines (FL_0.69, FL_0.79 and FL_0.86) showed discrepancies in the ranking of deletion size and that these lines have both presence and absence variation relative to each other (Table 2; Fig. 4). These results agree with earlier reports that these three lines contain terminal deletions combined with interstitial deletions rather than single terminal deletions (Hohmann et al. 1995).

## Conclusion

The high-density deletion bin map of wheat chromosome 7B was successfully constructed by genotyping aneuploid wheat stocks using 270 K CGH Nimblegen microarray. Using the most recently published chromosome survey sequences of bread wheat A, B and D sub-genomes (IWGSC, 2014) we could design 7B specific CGH probes, and accurately assign a total of 3,671 sequence contigs and scaffolds (~8 % of the chromosome 7B) to nine chromosomal bins. This map is the highest density deletion bin map for 7B so far, representing a ~100× increase in the bin mapped 7B sequence compared to previous studies (Hossain et al. 2004) and represents an important step towards high-resolution physical map of 7B.

**Author contributions** TB was responsible for carrying out experimental work, participated in data analysis and writing the manuscript, LG performed the normalization of hybridization data, AK and SK participated in drafting the manuscript, XH performed the clustering of SNP data in the RIL population, ML developed the RIL mapping population, worked on clustering SNP data in the RIL population, NS participated in the design of the experiment, data analysis and drafting the manuscript, SL and OA helped coordinate the study and draft the manuscript, SS performed data analysis, helped to draft the manuscript and was responsible for final version of the manuscript.

**Acknowledgments** The project was funded by grants from the Norwegian Research Council (project no. 199387/I99) and Graminor A/S to Odd-Arne Olsen.

**Conflict of interest** The authors declare they have no conflict interests.

## References


Akhunov ED, Goodyear AW, Geng S, Qi LL, Echalier B, Gill BS, Miftahudin, Gustafson JP, Lazo G, Chao SM, Anderson OD, Linkiewicz AM, Dubcovsky J, La Rota M, Sorrells ME, Zhang DS, Nguyen HT, Kalavacharla V, Hossain K, Kianian SF, Peng JH, Lapitan NLV, Gonzalez-Hernandeiz JL, Anderson JA, Choi DW, Close TJ, Dilbirligi M, Gill KS, Walker-Simmons MK, Steber C, McGuire PE, Qualset CO, Dvorak J (2003) The organization and rate of evolution of wheat genomes are correlated with recombination rates along chromosome arms. Genome Res 13(5):753–763. doi:10.1101/Gr.808603

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215(3):403–410. doi:10.1016/S0022-2836(05)80360-2

Belova T, Zhan B, Wright J, Caccamo M, Asp T, Simkova H, Kent M, Bendixen C, Panitz F, Lien S, Dolezel J, Olsen OA, Sandve SR (2013) Integration of mate pair sequences to improve shotgun assemblies of flow-sorted chromosome arms of hexaploid wheat. BMC Genom 14:222. doi:10.1186/1471-2164-14-222

Choulet F, Wicker T, Rustenholz C, Paux E, Salse J, Leroy P, Schlub S, Le Paslier MC, Magdelenat G, Gonthier C, Couloux A, Budak H, Breen J, Pumphrey M, Liu S, Kong X, Jia J, Gut M, Brunel D, Anderson JA, Gill BS, Appels R, Keller B, Feuillet C (2010) Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. Plant Cell 22(6):1686–1701. doi:10.1105/tpc.110.074187

Devos KM, Dubcovsky J, Dvorak J, Chinoy CN, Gale MD (1995) Structural evolution of wheat chromosomes 4a, 5a, and 7b and its impact on recombination. Theor Appl Genet 91(2):282–288. doi:10.1007/Bf00220890

Endo TR, Gill BS (1996) The deletion stocks of common wheat. J Hered 87(4):295–307

Fraley C, Raftery AE (2007) Bayesian regularization for normal mixture estimation and model-based clustering. J Classif 24(2):155–181. doi:10.1007/S00357-007-0004-5

Hohmann U, Endo TR, Herrmann RG, Gill BS (1995) Characterization of deletions in common wheat induced by an Aegilops cylindrica Chromosome: detection of multiple chromosome rearrangements. Theor Appl Genet 91(4):611–617

Hossain KG, Kalavacharla V, Lazo GR, Hegstad J, Wentz MJ, Kianian PM, Simons K, Gehlhar S, Rust JL, Syamala RR, Obeori K, Bhamidimarri S, Karunadharma P, Chao S, Anderson OD, Qi LL, Echalier B, Gill BS, Linkiewicz AM, Ratnasiri A, Dubcovsky J, Akhunov ED, Dvorak J, Miftahudin, Ross K, Gustafson JP, Radhawa HS, Dilbirligi M, Gill KS, Peng JH, Lapitan NL, Greene RA, Bermudez-Kandianis CE, Sorrells ME, Feril O, Pathan MS, Nguyen HT, Gonzalez-Hernandez JL, Conley EJ, Anderson JA, Choi DW, Fenton D, Close TJ, McGuire PE, Qualset CO, Kianian SF (2004) A chromosome bin map of 2148 expressed sequence tag loci of wheat homoeologous group 7. Genetics 168(2):687–699. doi:10.1534/genetics.104.034850

International Barley Sequencing Consortium (2012) A physical, genetic and functional sequence assembly of the barley genome. Nature 491(7426):711–716. doi:10.1038/nature11543

International Rice Genome Sequencing Consortium (2005) The map-based sequence of the rice genome. Nature 436(7052):793–800. doi:10.1038/nature03895

International Wheat Genome Sequencing Consortium (IWGSC) (2014) A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. Science 345(6194):1251788

Joppa LR, Williams ND (1977) D-genome substitution-monosomics of durum-wheat. Crop Sci 17(5):772–776

Kumar S, Balyan HS, Gupta PK (2012) Comparative DNA sequence analysis involving wheat, brachypodium and rice genomes using mapped wheat ESTs. Triticeae Genomic Genetic 3(3):25–37

Lazo GR, Chao S, Hummel DD, Edwards H, Crossman CC, Lui N, Matthews DE, Carollo VL, Hane DL, You FM, Butler GE, Miller RE, Close TJ, Peng JH, Lapitan NL, Gustafson JP, Qi LL, Echalier B, Gill BS, Dilbirligi M, Randhawa HS, Gill KS, Greene RA, Sorrells ME, Akhunov ED, Dvorak J, Linkiewicz AM, Dubcovsky J, Hossain KG, Kalavacharla V, Kianian SF, Mahmoud AA, Miftahudin, Ma XF, Conley EJ, Anderson JA, Pathan MS, Nguyen HT, McGuire PE, Qualset CO, Anderson OD (2004) Development of an expressed sequence tag (EST) resource for wheat (*Triticum aestivum* L.): EST generation, unigene analysis, probe selection and bioinformatics for a 16,000-locus bin-delineated map. Genetics 168(2):585–593. doi:10.1534/genetics.104.034777

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, Schmutz J, Spannagl M, Tang H, Wang X, Wicker T, Bharti AK, Chapman J, Feltus FA, Gowik U, Grigoriev IV, Lyons E, Maher CA, Martis M, Narechania A, Otillar RP, Penning BW, Salamov AA, Wang Y, Zhang L, Carpita NC, Freeling M, Gingle AR, Hash CT, Keller B, Klein P, Kresovich S, McCann MC, Ming R, Peterson DG, Mehboob ur R, Ware D, Westhoff P, Mayer KF, Messing J, Rokhsar DS (2009) The Sorghum bicolor genome and the diversification of grasses. Nature 457(7229):551–556. doi:10.1038/nature07723

Paux E, Sourdille P, Salse J, Saintenac C, Choulet F, Leroy P, Korol A, Michalak M, Kianian S, Spielmeyer W, Lagudah E, Somers D, Kilian A, Alaux M, Vautrin S, Berges H, Eversole K, Appels R, Safar J, Simkova H, Dolezel J, Bernard M, Feuillet C (2008) A physical map of the 1-gigabase bread wheat chromosome 3B. Science 322(5898):101–104. doi:10.1126/science.1161847

Paux E, Faure S, Choulet F, Roger D, Gauthier V, Martinant JP, Sourdille P, Balfourier F, Le Paslier MC, Chauveau A, Cakir M, Gandon B, Feuillet C (2010) Insertion site-based polymorphism markers open new perspectives for genome saturation and marker-assisted selection in wheat. Plant Biotechnol J 8(2):196–210. doi:10.1111/j.1467-7652.2009.00477.x

Philippe R, Paux E, Bertin I, Sourdille P, Choulet F, Laugier C, Simkova H, Safar J, Bellec A, Vautrin S, Frenkel Z, Cattonaro F, Magni F, Scalabrin S, Martis MM, Mayer KF, Korol A, Berges H, Dolezel J, Feuillet C (2013) A high density physical map of chromosome 1BL supports evolutionary studies, map-based cloning and sequencing in wheat. Genome Biol 14(6):R64. doi:10.1186/gb-2013-14-6-r64

The Potato Genome Sequencing Consortium (2011) Genome sequence and analysis of the tuber crop potato. Nature 475(7355):189–195. doi:10.1038/nature10158

Qi LL, Echalier B, Chao S, Lazo GR, Butler GE, Anderson OD, Akhunov ED, Dvorak J, Linkiewicz AM, Ratnasiri A, Dubcovsky J, Bermudez-Kandianis CE, Greene RA, Kantety R, La Rota CM, Munkvold JD, Sorrells SF, Sorrells ME, Dilbirligi M, Sidhu D, Erayman M, Randhawa HS, Sandhu D, Bondareva SN, Gill KS, Mahmoud AA, Ma XF, Miftahudin, Gustafson JP, Conley EJ, Nduati V, Gonzalez-Hernandez JL, Anderson JA, Peng JH, Lapitan NLV, Hossain KG, Kalavacharla V, Kianian SF, Pathan MS, Zhang DS, Nguyen HT, Choi DW, Fenton RD, Close TJ, McGuire PE, Qualset CO, Gill BS (2004) A chromosome bin map of 16,000 expressed sequence tag loci and distribution of genes among the three genomes of polyploid wheat. Genetics 168(2):701–712. doi:10.1534/genetics.104.034868

Saintenac C, Falque M, Martin OC, Paux E, Feuillet C, Sourdille P (2009) Detailed recombination studies along chromosome 3B provide new insights on crossover distribution in wheat (*Triticum aestivum* L.). Genetics 181(2):393–403. doi:10.1534/genetics.108.097469

Sears ER (1954) The aneuploids of common wheat. College of Agriculture, Agricultural Experimental Station. Res Bull 572:1–58

Smit AFA, Hubley R, Green P (1996–2010) RepeatMasker Open-3.0. http://www.repeatmasker.org.

Springer NM (2010) Isolation of plant DNA for PCR and genotyping using organic extraction and CTAB. Cold Spring Harb Protoc 2010(11):pdb prot5515. doi:10.1101/pdb.prot5515

Wang S, Wong D, Forrest K, Allen A, Chao S, Huang BE, Maccaferri M, Salvi S, Milner SG, Cattivelli L, Mastrangelo AM, Whan A, Stephen S, Barker G, Wieseke R, Plieske J, International Wheat Genome Sequencing C, Lillemo M, Mather D, Appels R, Dolferus R, Brown-Guedira G, Korol A, Akhunova AR, Feuillet C, Salse J, Morgante M, Pozniak C, Luo MC, Dvorak J, Morell M, Dubcovsky J, Ganal M, Tuberosa R, Lawley C, Mikoulitch I, Cavanagh C, Edwards KJ, Hayden M, Akhunov E (2014) Characterization of polyploid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array. Plant Biotechnol J 12:787–796. doi:10.1111/pbi.12183

Werner JE, Endo TR, Gill BS (1992) Toward a cytogenetically based physical map of the wheat genome. Proc Natl Acad Sci USA 89(23):11307–11311

Wu Y, Bhat PR, Close TJ, Lonardi S (2008) Efficient and accurate construction of genetic linkage maps from the minimum spanning tree of a graph. PLoS Genet 4(10):e1000212. doi:10.1371/journal.pgen.1000212

# Paper III

# Anchoring physical contigs of bread wheat chromosome 7B long arm.

Belova, T[1]., Frenkel, Z.[2], Zhan, B.[1], Lillemo, M.[1], Korol, A.[2], Paux, E.[3] , Balfourier, F.[3], Sourdille, P.[3], Simkova, H[4]., Kubalakova, M[4], Dolezel, J.[4] , Cattonaro, F.[5], Li, L[6]., Min, J[6]., Chen, J[6]., Yang,Y[6]., Xu, X.[6], Kent, M[1]., Lien, S[1]., Sandve, S.R. and Olsen, O.-A.[1]*

[1] Norwegian University of Life Sciences, CIGENE/IPV, Norway.
[2] University of Haifa, Institute of Evolution, Haifa, Israel
[3] Institut National de la recherché agronomique, INRA, Clermont-Ferrand, France
[4] Institute of Experimental Botany, Laboratory of Molecular Cytogenetics and Cytometry, Olomouc, Czech Republic.
[5]Istituto di Genomica Applicata and IGA Technology Services, Udine, Italy
[6]Beijing Genomics Institute, China

* corresponding author
Email: odd-arne.olsen@nmbu.no

**Abstract**

Bread wheat is one of the world's most important cereals, yet wheat sequencing and genomic research remain challenging because of its complex polyploid genome and high repeat content. To obtain a high-quality reference sequence for wheat, a BAC-based physical map for each chromosome is constructed prior to sequencing. In this study we present the first draft of a genetically anchored and ordered physical map of the long arm of bread wheat chromosome 7B. First, fingerprinting generated a 7BL MTP consisting of 5,229 BACs. Sequencing of these BACs resulted in 105,445 contigs with a contig N50 of 17,5 Kbp. Further scaffolding with 10Kb- and 20Kb- MP libraries improved N50 by approximately 7 fold. Finally, the integration of the 7BL physical map with the generated sequences resulted in 125 physical contigs covering the entire chromosome arm. To anchor these physical contigs to the chromosome map, we applied a combination of several mapping resources including high-density deletion bin mapping, genetic mapping utilizing three crosses and synteny mapping. By performing *in silico* integration of the physical map, 109 7BL physical contigs spanning ~514 Mbp or ~98% of the 7B sequence scaffold were assigned chromosomal positions. Among them 96 physical contigs were placed in a linear order with 61% assigned with a high confidence. In the current study, 16 of 7BL physical contigs failed to be anchored to the 7B map used, representing only 1.7% of the 7BL sequence.

*Keywords:* wheat, physical map, genetic map, synteny-based mapping, deletion mapping, contig

*Abbreviations:* MTP: Minimum Tiling Path, MP: mate pair, CS: Chinese Spring

## Introduction

Interest for unraveling the genomes of cereals is mainly governed by the critical role of cereals in world food supplies. Wheat is one of the most nutritionally important cereals, providing around 20% of the calories in human diet and its consumption is growing rapidly. Over the past years the growth in grain yields has stagnated at around 0.9% per year (Wheat lag 2014). To meet future demands for food, it is estimated that wheat yields must grow by 1.7% each year. One important resource expected to contribute significantly to this increase is the genome sequence of bread wheat.

Bread wheat, *Triticum aestivum. L.*, is an allohexaploid (A, B and D genomes) with a genome size of 17 Gbp containing more than 80% repeat sequences (Choulet et al. 2010; Dvorak and Zhang 1990; Dvorak et al. 1993). The genome of *T. aestivum* arose from three separate hybridization events. According to recent findings, the first hybridization occurred between species carrying the A and the B genomes 1-2 million years after their divergence from a common ancestor ~7 million years ago giving rise to the D-genome through homoploid hybrid speciation (Marcussen et al. 2014). The second hybridization is estimated to have occurred approximately 500,000 years ago between the two grass species *T. urartu* (the A genome donor), and *T. speltoides* (the B genome donor) giving rise to tetraploid species. The third hybridization is believed to have occurred approximately 10,000 years ago in cereal fields between cultivated tetraploid wheat (AABB) and the wild grass *Ae. taushii* (D genome) (Petersen et al. 2006). The two last hybridizations were followed by chromosome doubling in the new hybrid, enabling normal bivalent formation at meiosis and thus the production of fertile plants.

One commonly used strategy for sequencing large genomes is the BAC-by-BAC approach, in which BAC libraries containing DNA from the species under study are first constructed, then the BACs are assembled into physical contigs using BAC finger printing, followed by anchoring of the derived physical contigs to chromosomes using various methods. BACs that cover the entire chromosome with a minimum sequence overlap are then identified, referred to as a Minimal Tiling Path (MTP). These MTP BACs are then sequenced and assembled into the sequence of the physical contigs. One approach to perform fingerprinting is to use the SNaPshot HICF technology in which individual BAC clones are digested with restriction nucleases producing DNA fragments that are separated by capillary electrophoresis (Luo et al.

2010). Physical contigs are made by assembling fingerprinted BAC fragments based on the overlaps of sets of bands (using the FPC (Soderlund et al. 1997) or LTC (Frenkel et al. 2010) software). It should be noted that erroneous BAC contig formation can occur due to several reasons, including genome repeats, locally low information content in finger printing, and clone-by-clone DNA contamination.

Integration of fingerprint-based physical maps with other molecular maps such as genetic maps, deletion bin maps, radiation hybrid maps reduces the number of chimeric physical contigs and increases overall confidence in the final assembly. Successful anchoring of physical contigs using genetic maps depends on a high marker density and evenly distributed markers, as well as sufficient map resolution. In wheat, ordering of physical contigs along the chromosomes solely based on genetic maps is challenging, if not impossible, due to suppression of recombination in the (peri) centromeric region (Saintenac et al. 2009). To augment the anchoring process in wheat, other approaches have been utilized, including deletion bin mapping, synteny-based mapping using closely related species (Kumar et al. 2012b; Philippe et al. 2013) and radiation hybrid mapping. Deletion mapping utilizes a set of wheat deletion stocks and allows anchoring of markers to relatively short (ranging from 20 to 155 Mb in size) chromosomal segments (deletion bins) (Qi et al. 2004). The limitation of this approach is a lack of order of sequences within the bins. Local synteny between wheat and other related sequenced genomes of species, e.g. *Brachypodium*, rice and sorghum, can help in the ordering and the orientation of contigs within bins (Kumar et al. 2012b) (Kumar et al. 2012b). One of the limitations of this approach is the presence of inversions and transposition of genes and/or gene blocks (Kumar et al. 2012b). Recent publications suggest that radiation-hybrid mapping can enable high-resolution mapping in wheat. Mapping resolution reported for wheat chromosomes have been reported to be in the range 140-200Kb (Kumar et al. 2012a; Kalavacharla et al. 2006). However, even though the reported resolution is extremely high, the experiments were performed on few BAC contigs, with no evidence for high-resolution mapping and ordering of all BAC contigs along wheat chromosomes. Altogether, it is apparent that combined anchoring strategies must be applied to complex genomes such as wheat to successfully order BAC contigs along the chromosomes.

The international wheat genome sequencing consortium (IWGSC) was established to sequence the bread wheat genome and develop physical and genetic maps (www.wheatgenome.org). To reduce the complexity of the genome, the task of sequencing the entire wheat genome has been subdivided into the sequencing of individual chromosome

arms. Sufficient amount of chromosome arm-specific DNA for BAC library construction was obtained by flow-sorting (Safar et al. 2010). This manuscript focuses on the ordering of the 7BL physical contigs produced by the BAC sequencing effort by the Norwegian IWGSC sub-project "Expanding the technology base for Norwegian wheat breeding; Sequencing wheat chromosome 7B". The purpose of the work described here is to anchor the 7BL physical contigs to the 7B genetic and molecular maps. The anchoring is carried out by a combination of 7BL deletion bin mapping, genetic and synteny-based mapping to produce the first draft version of the 7BL chromosome arm.

## Material and methods

*Physical contig sequencing and assembly*

The sequence of the physical contigs from the long arm of bread wheat chromosome 7B was produced by the Norwegian IWGSC sub-project "Expanding the technology base for Norwegian wheat breeding "Sequencing wheat chromosome 7B". The work leading to the physical contigs included DNA isolation from chromosome 7BL by flow-sorting, construction of BAC libraries and finger printing of the BACs to construct MTP. As a supplement to MTP clones, 579 randomly selected clones were selected for sequencing, giving a total of 5,808 sequenced 7BL BAC clones. In addition to Illumina pair end sequencing, mate pair libraries of 10Kb and 20Kb inserts were constructed for pools of 12 BACs of the entire 7BL MTP. The physical contigs for the 7BL-specific BAC library were assembled using two methods: (a) FPC software (Soderlund et al. 1997) according to the standard recommendation of the IWGSC (Scalabrin et al. 2010) and (b) the LTC software (Frenkel et al. 2010). The MTP used was selected based on a LTC contig assembly. LTC-based contigs were manually elongated and merged into the longest possible physical scaffolds. Based on the anchoring results presented below, some of these physical contigs were subjected to further editing.

*Ordering and anchoring of the 7BL physical contigs using molecular and genetic maps*

Anchoring of the physical contigs was performed using a combination of different mapping approaches: (a) deletion bin mapping, (b) genetic mapping and (c) synteny-based mapping. The sequences of the individual BACs were used to anchor BAC physical contigs *in silico* to selected marker sequences with BLASTN (Altschul et al. 1990). Subsequent to the initial anchoring, we have improved the accuracy of marker positions within each BAC contig and the positioning of BAC contigs by manual curation. Physical contigs mapping to conflicting marker positions were identified and reexamined. Putatively chimeric contigs were manually edited using LTC.

*a. Deletion-bin mapping of chromosome 7BL*

In order to place the physical contigs into deletion bins along chromosome 7BL, sequences previously assigned to 7B chromosomal bins as part of this thesis (Belova et al. 2014) were

used in BLASTN searches to identify these sequences in the BACs of the physical contigs. The BLASTN results were parsed to keep hits with at least 99% of sequence identity and covering at least 30% of the query length.

*b. Genetic mapping of chromosome 7B*

Three genetic populations were used to genetically anchor the physical contigs of 7BL. The first population, comprising ~282 F8 individuals derived from the cross between Chinese Spring and Renan (later referred as CS*Renan) were genotyped with an Axiom high-density genotyping chip carrying 423,385 SNPs (420K SNP chip). The 7B linkage map was provided by P. Sourdille (Institut National de la recherché agronomique, INRA). The INRA group kindly provided access to the IWGSC chromosome survey sequences from which SNP marker sequences originated. To assign SNP markers of the 7B genetic map to the BAC sequences, BLASTN searches of the chromosome-survey sequences (Consortium 2014) against the BAC sequences were performed. BLASTN hits were filtered based on >=99% identity and >=10% coverage of the query length (the length of the survey sequence contig). The second population consisting of 131 recombinant inbred lines (RILs) was developed from a cross between the CIMMYT breeding line 'SABUF/5/BCN/4/RABI//GS/CRA/3/AE.SQUARROSA (190)' (selection history CASS94Y00042S-32PR-1B-0M-0Y) and the German spring wheat cv. 'Naxos' (pedigree Tordo/St.Mir808-Bastion//Miranet), and is referred to as SY*Naxos in the text below. The third population consisted of 181 F6 RIL lines and was developed by single descent pedigree from the cross SHA3/CBRD*Naxos. SHA3/CBRD is a spring type breeding line from CIMMYT with pedigree 'Shanghai-3//Chuanmai 18/Bagula' and selection history "-0SHG-6GH-0FGR-0FGR-0Y" (later referred as SHA3/CBRD*Naxos). Populations 2 and 3 were genotyped with iSelect 90K wheat chip from Illumina, which contains a total of 81,587 SNP markers (Wang et al. 2014). The genotypes were called using the Genome Studio V2011.1. Genetic linkage groups were created using the MultiPoint Ultradense software, with a cutoff of maximum missing data 18, minimum size of bound together markers 3, recombination fraction 0.3, LOD threshold 2.0. Anchoring of markers from the 90K SNP chip was based on BLASTN searches against BAC sequences. Matches with >=99% identity and 100% coverage of marker locus were accepted for anchoring BAC sequences.

*c. Utilizing the syntenic conservation between bread wheat and Brachypodium, rice, and sorghum to align 7BL physical contigs.*

The third approach to arrange physical contigs along the chromosome is the use of the GenomeZipper approach which is based on the syntenic conservation of local gene order in grasses. In the recent work of (Pfeifer et al. 2014) the linear order of 57,903 bread wheat genes has been predicted for all wheat chromosomes on the basis of transciptome data and integration of syntenic gene content information from rice, *Brachypodium* and sorghum and gene order information from barley (Pfeifer et al. 2014). We have used transcripts placed on the 7B *Triticeae* prototype chromosome (Pfeifer et al. 2014) in BLASTN search against the 7B BAC sequences. Blast hits with >=99% identity and covering at least 10% of the query sequence were accepted for anchoring a BAC sequence.

## Results

*Building 7BL physical contigs and BAC-by-BAC sequencing*

In total 72,960 7BL BAC clones were finger printed, representing >12x coverage of the long arm of bread wheat chromosome 7B. Only clones with high fingerprint quality (~80%) were used in the analysis (Table 1). The final LTC-based physical contig assembly obtained after manual editing and end-to-end merging included 47,013 of the fingerprinted 7BL BAC clones, excluding singleton BAC clones (Table 1). These contigs covered ~470 Mbp of the sequence length of 7BL based on estimates of virtual band length (~1.2 kbp) and number of bands for each clone. In the first round of paired end sequencing of 7BL MTP BAC clones, sequences were obtained for 5,808 BACs with an average coverage of ~48X. Within this set of BACs, 1,183 had low sequencing coverage (<20X) and were re-sequenced together with BACs for which sequencing had failed. The average MTP-BAC overlap was estimated to about 30% based on the sequence data (not shown), in concordance with the expected MTP BAC overlap. After the second round of sequencing, more than 95% of MTP BACs had a coverage of > 20X. The assembly statistics are presented in table 2. The number of contigs obtained was 105,445 summing up to ~538 Mbp with a N50 of 17.5Kb. After including the sequences from the 10Kb and 20Kb mate-pair libraries for scaffolding, assembly N50 was improved by 6.2-fold (Table 2). The mean number of sequences per BAC clone decreased from 19 to 7. The final MTP for 7BL (after inclusion of sequence data) consists of 125

unordered physical contigs (Table 3, supplementary table 1). For a list of the 125 7BL physical contigs with corresponding BAC clones, please see supplementary table 1.

*Anchoring of the 7BL MTP-contigs to genetic and molecular maps*

In order to anchor the 7BL physical contigs we proceeded to integrate the 7BL physical map with the 7B deletion bin map, genetic maps and 7B *Triticeae* map.

*a) Deletion bin mapping*

Using the sequences of the 3,671 bin mapped contigs/scaffolds from (Belova et al. 2014) we assigned 105 7BL physical contigs to seven deletion bins along the long arm of chromosome 7B (Fig.1, Table 5). The number of physical contigs assigned to bins ranged from 5 to 33. The cumulative length of bin mapped contigs was estimated to represent ~97% of the 7BL sequence scaffold length. By the nature of deletion bin mapping, the physical contigs within bins are unordered. An example of the map of bins 7BL_0.4-0.45 and 7BL_0.45_0.63 is shown in Figure 2, each containing 6 physical contigs.

*b) Recombination mapping*

A physical BAC contig can be assigned unambiguously to a specific genetic location if it contains markers that do not hit sequences present in BACs of another BAC contig. The core resource for ordering physical contigs along the 7B chromosome in our study was high-density genetic map produced from CS*Renan population, because cv. CS is the reference for wheat chromosome genome sequencing and physical mapping. For anchoring SNP sequences from 420K Axiom SNP chip we performed BLAST filtering with relaxed criteria (i.e. CSS vs BACs we have retained hits with minimum coverage of 10% and >=99% identity). To increase the reliability of physical contig anchoring, only markers from CS*Renan map that are associated with BACs in single physical contig were considered for further analysis. The CS*Renan map contained 4438 markers distributed over 308 unique loci spanning 127.3 cM. In total, 52% of the SNP markers on the genetic map were anchored to 7B BAC clones of which 98.6 % belonged to specific BAC contigs. An example of genetic mapping within the bins 7BL_0.4-0.45 and 7BL_0.45-0.63 is shown in Figure 2. After selecting markers which (i) belong to unique IWGSC survey sequence contigs, (ii) have unique genetic position and (iii)

hit BACs in not more than one physical contig, 1211 markers were identified. These provided anchor points for 76 of the 7BL physical contigs (Table 4, Table 5).

Among these, 289 were anchored to 7BL BAC clones belonging to 52 7BL BAC physical contigs. Thirdly, the 7B linkage map built from the SHA3/CBRD*Naxos population contained 594 markers with a total genetic map length of 100,67 cM. Of these, 321 markers anchored to 247 unique 7BL BAC clones providing anchoring information for 46 7BL physical contigs (Table 4, Supplementary table 2). The 7B linkage map obtained from genotyping the SY*Naxos population contained 623 markers with a total length of 220.2 cM (Supplementary table 3). For SY*Naxos and SHA3/CBRD*Naxos populations we have included markers associated with single physical contig and also have reported positions for contigs which have evidence from CS*Renan and 7B *Triticea* prototype map.

To assess the power of high-density genetic mapping in resolving BAC contig ordering, we investigated the distribution of 7BL physical scaffolds per genetic position in the Ren*CS population (Fig.3). Only markers that are associated with BACs in single physical scaffolds were considered in this analysis. From the results it is evident that the resolution of the genetic map is not uniform along the length of the chromosome and that the resolution is particularly low in the centromeric region. The number of anchored physical contigs per genetic position varied between the distal, middle and centromeric parts of 7BL chromosome arm. The highest resolution was observed for the middle part of 7BL. The number of anchored contigs did not exceed 2 for the distal parts while up to 5 physical contigs were anchored to single genetic position for the centromeric region of 7BL chromosome arm (Fig.3).

*Ordering of the 7B physical contigs based on collinearity to other Triticeae genomes*

An additional anchoring strategy applied in this study is the so-called GenomeZipper approach. This method exploits a set of genes which are highly conserved among the wheat, rice, *Brachypodium* and sorghum genomes to deduce the virtual order of genes along the wheat chromosomes. In total, 583 syntenic genes were included in this analysis. Among them 97% (563/583) were orthologs from *Brachypodium*, 77% (449/583) from rice and 84% (492/583) from sorghum. Regions with conserved gene content to wheat chromosome 7B encompassed *Brachypodium* chromosomes 1 and 3, rice chromosomes 6 and 8 and regions on sorghum chromosomes 7 and 10 (Fig. 4, Supplementary table 4). Figure 4 illustrates the synteny relationship between 7BL and chromosomes 1 and 3 of *Brachypodium*. The subcentromeric region on 7BL showed synteny to *Brachypodium* chromosome 3, and the

distal part of the arm was syntenic to Brachypodium chromosome 1. An example of synteny mapping within the bins 7BL_0.4-0.45 and 7BL_0.45-0.63 is shown in Figure 2. The number of zipper-based markers per physical scaffold ranged from 1 to 31. The density of genic markers per physical scaffold correlated well the length scaffold length; short physical scaffolds were anchored with fewer markers than larger scaffolds. The mean length of the anchored and unanchored physical scaffolds was 49 and 9 clones, respectively.

*A draft version of chromosome 7BL*

In total, among the 125 7BL physical contigs 109 7BL physical contigs representing ~514 Mbp of the 7B sequence scaffold is assigned positional information. Among these, 4 lack a bin position. Nighty six physical contigs (~503 Mbp of the 7B sequence scaffolds) were assigned either a genetic and/or a 7B *Triticeae* map position, whereas 13 were assigned only a bin position. Among these 96 anchored physical contigs, 96% could be oriented relative to each other. Figure 2 illustrates an example of ordering physical contigs within two 7BL deletion bins (7BL_0.4_0.45 and 7BL_0.45_0.63) as well as the contribution of different maps to ordering. In total, there are 12 physical scaffolds placed into these two bins. Of these, 10 were ordered based on the genetic maps and synteny map whereas the placement of one physical contig (7BLctg61) was inferred solely from synteny based map. In total, among 125 7BL physical contigs, only 16 of them, covering only ~9Mb (1.7%) of the 7B sequence scaffolds, remained without positional information.

## Discussion

The generation of the integrated physical and genetic map is challenging in hexaploid wheat and it is apparent from recent reports on wheat physical mapping (Philippe et al. 2013; Paux et al. 2008; Raats et al. 2013) that multiple resources are required to build a sequential order of physical contigs along the chromosome. The present study is part of the Norwegian participation in the International Wheat Genome Sequencing Consortium (IWGSC) which aims to sequence bread wheat chromosome arms using the strategy of constructing BAC-based physical maps prior to sequencing. The estimated size of 7BL is 540 Mb and the LTC-based physical map is comprised of 125 physical contigs that have on the average 360 BAC clones per physical contig. The sequence assembly of the MTP BACs of 7B covers ~94% of the estimated chromosome 7B size (Table 2). Based on the sequences of the MTP BACs, the anchoring of the physical contigs was performed *in silico* by DNA sequence homology searches of BACs against marker sequences. One complication using this approach for

anchoring is that a single marker may hits several locations in the MTP. One possible explanation for such behavior is duplicated regions of the genome. In this case, markers cannot be unambiguously placed on the genetic maps. Another explanation for multiple hits is that a marker is located in the overlap-region of two MTP-contigs and that these contigs were not merged because of poor clone overlap at BAC-fingerprint level. The latter situation may be resolved by a detailed examination of the fingerprints and sequences of the region in question. A third possibility is that a marker mapping to different locations in the MTP contig is caused by errors in BAC contig assembly or in BAC sequencing. To increase the accuracy of physical contig mapping in our analysis, we used markers assigned to single physical contig. However, if physical contig had stong evidence from one map we report its position from other maps if the positions are not inconsistent. Generally we manually examined the consistency of marker positions within contigs.

In this study we have determined the order of 87 % of 7BL physical contigs, using a three step strategy. In the first step using bin-mapping, we assigned 105 7BL physical scaffolds to 7 deletion bins (Table 5, Fig. 1). Even though the size of the deletion bins are quite large (ranging from the 20 to 155Mb), making it impossible to access the order of contigs within the bin, it provided essential information for the initial assignment of the physical contigs along the chromosome (Table 5).

In the second step of building the integrated 7BL physical map we established the order of physical contigs within each deletion bin using genetic maps from three crosses (Table 5). Comparing the bin map assignment with the genetic map revealed a high level of accuracy between bin map position and Renan*CS genetic position.

It is well known that the main limitation of genetic mapping in wheat, as well as in other plant species including maize, barley, Arabidopsis and rice is the failure to fully resolve the order of the contigs in the (peri) centromeric compartments. The effect of this is that high numbers of physical contigs map to the same genetic position. In our study this was reflected by the mapping of up to 5 physical contigs to a single genetic position in the centromeric region (Fig. 3). In total, genetic information provided anchoring for 83 of 7BL physical scaffolds (Table 5). The order for most of the 7BL physical contigs within distal bins 7BL_0.63-0.69* and 7BL_0.69*_1.00 were derived from the 7B genetic maps.

In the third step of 7B map building, the order of physical contigs unresolved by genetic mapping was obtained from the 7B *Triticeae* prototype map. Even though there are

rearrangements between the genetic map and the synteny map that presumably correspond to true genome rearrangements (Fig. 4), there are identifiable blocks of genes which are collinear between wheat and *Brachypodium* allowing us to deduce the order of physical scaffolds. For example, the region covering bin 7BL_0.4-0.45 and bin 7BL_0.45_063 represents a *Brachypodium* inversion compared to wheat (Fig. 2). Synteny-based mapping provided the framework for integration and deduction of the virtual order of 55 physical scaffolds on 7BL. The order of 13 physical scaffolds within bins was established solely upon synteny information (Table 5). The syntenic integration based on information from rice, sorghum and *Brachypodium* proved to be especially valuable for regions with limited genetic resolution, i.e. centromeric regions (Table 5, Supplementary table 5). As shown in table 5 the order of contigs within the centromeric bins 7BL_0-0.14 and 7BL_0.14_033 was obtained mainly based on the synteny map. The combination of genetic and synteny-based mapping data allowed us to allocate 96 of the 7BL physical contigs into a proposed linear order with 96% having a unique anchoring position (table 5). Fifty nine 7BL physical contigs were anchored in very reliable way based on evidence from bin map, genetic map(s) and synteny map (Table 5). Placement of 37 7BL physical contigs was supported by bin map and genetic or synteny maps, or synteny and genetic maps (Table 5). In the current study, 16 (12.8%) of 7BL physical contigs remained unanchored to the 7B map, representing only 1.7% of the 7B sequence scaffolds. Most of unanchored contigs contained a small number of BACs.

One of the further strategies to anchor unplaced physical contigs is to develop molecular markers specific to the unanchored contigs and screen BAC pools, deletion stocks and genetic populations to determine their chromosomal positions. Additionally, genome mapping on nanochannel arrays (Hastie et al. 2013) and optical mapping (Zhou et al. 2009) successfully applied in other genome projects can provide an additional layer of mapping information in future studies.

## Acknowledgements

The project was funded by grants from the Norwegian Research Council (project no.199387/I99) and Graminor A/S to Odd-Arne Olsen.

## References


Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. Journal of molecular biology 215 (3):403-410. doi:10.1016/S0022-2836(05)80360-2

Belova T, Grønvold L, Kumar A, Kianian S, He X, Lillemo M, Springer NM, Lien S, Olsen O-A, Sandve SR (2014) High-density deletion bin map of wheat chromosome 7B. accepted, TAG

Choulet F, Wicker T, Rustenholz C, Paux E, Salse J, Leroy P, Schlub S, Le Paslier MC, Magdelenat G, Gonthier C, Couloux A, Budak H, Breen J, Pumphrey M, Liu SX, Kong XY, Jia JZ, Gut M, Brunel D, Anderson JA, Gill BS, Appels R, Keller B, Feuillet C (2010) Megabase Level Sequencing Reveals Contrasted Organization and Evolution Patterns of the Wheat Gene and Transposable Element Spaces. Plant Cell 22 (6):1686-1701. doi:DOI 10.1105/tpc.110.074187

Consortium IWGS (2014) A chromosome-based draft sequence of the hexaploid bread wheat genome. accepted, Science

Dvorak J, Terlizzi P, Zhang HB, Resta P (1993) The evolution of polyploid wheats: identification of the A genome donor species. Genome / National Research Council Canada = Genome / Conseil national de recherches Canada 36 (1):21-31

Dvorak J, Zhang HB (1990) Variation in repeated nucleotide sequences sheds light on the phylogeny of the wheat B and G genomes. Proceedings of the National Academy of Sciences of the United States of America 87 (24):9640-9644

Frenkel Z, Paux E, Mester D, Feuillet C, Korol A (2010) LTC: a novel algorithm to improve the efficiency of contig assembly for physical mapping in complex genomes. BMC bioinformatics 11:584. doi:10.1186/1471-2105-11-584

Hastie AR, Dong L, Smith A, Finklestein J, Lam ET, Huo N, Cao H, Kwok PY, Deal KR, Dvorak J, Luo MC, Gu Y, Xiao M (2013) Rapid genome mapping in nanochannel arrays for highly complete and accurate de novo sequence assembly of the complex Aegilops tauschii genome. PloS one 8 (2):e55864. doi:10.1371/journal.pone.0055864

Kalavacharla V, Hossain K, Gu Y, Riera-Lizarazu O, Vales MI, Bhamidimarri S, Gonzalez-Hernandez JL, Maan SS, Kianian SF (2006) High-resolution radiation hybrid map of wheat chromosome 1D. Genetics 173 (2):1089-1099. doi:10.1534/genetics.106.056481

Kumar A, Simons K, Iqbal MJ, de Jimenez MM, Bassi FM, Ghavami F, Al-Azzam O, Drader T, Wang Y, Luo MC, Gu YQ, Denton A, Lazo GR, Xu SS, Dvorak J, Kianian PM, Kianian SF (2012a) Physical mapping resources for large plant genomes: radiation hybrids for wheat D-genome progenitor Aegilops tauschii. BMC genomics 13:597. doi:10.1186/1471-2164-13-597

Kumar S, Balyan HS, Gupta PK (2012b) Comparative DNA sequence snalysis involving wheat, brachypodium and rice genomes using mapped wheat ESTs. Triticeae Genomics and Genetics 3 (3):25-37

Luo MC, Ma Y, You FM, Anderson OD, Kopecky D, Simkova H, Safar J, Dolezel J, Gill B, McGuire PE, Dvorak J (2010) Feasibility of physical map construction from fingerprinted bacterial artificial chromosome libraries of polyploid plant species. BMC genomics 11:122. doi:10.1186/1471-2164-11-122

Marcussen T, Sandve SR, Heier L, Spannagl M, Pfeifer M, IWGSC, Jakobsen KS, Wulff B, Steuernagel B, Mayer K, Olsen O-A (2014) Ancient Hybridizations Among the Ancestral Genomes of Bread Wheat. accepted, Science

Paux E, Sourdille P, Salse J, Saintenac C, Choulet F, Leroy P, Korol A, Michalak M, Kianian S, Spielmeyer W, Lagudah E, Somers D, Kilian A, Alaux M, Vautrin S, Berges H,

Eversole K, Appels R, Safar J, Simkova H, Dolezel J, Bernard M, Feuillet C (2008) A physical map of the 1-gigabase bread wheat chromosome 3B. Science 322 (5898):101-104. doi:10.1126/science.1161847

Petersen G, Seberg O, Yde M, Berthelsen K (2006) Phylogenetic relationships of Triticum and Aegilops and evidence for the origin of the A, B, and D genomes of common wheat (Triticum aestivum). Molecular phylogenetics and evolution 39 (1):70-82. doi:10.1016/j.ympev.2006.01.023

Pfeifer M, Kugler K, Sandve SR, Zhan B, Rudi H, Hvidsten TR, IWGSC, Mayer KF, Olsen O-A (2014) Genome interplay in the grain transcriptome of hexaploid bread wheat. accepted, Science

Philippe R, Paux E, Bertin I, Sourdille P, Choulet F, Laugier C, Simkova H, Safar J, Bellec A, Vautrin S, Frenkel Z, Cattonaro F, Magni F, Scalabrin S, Martis MM, Mayer KF, Korol A, Berges H, Dolezel J, Feuillet C (2013) A high density physical map of chromosome 1BL supports evolutionary studies, map-based cloning and sequencing in wheat. Genome biology 14 (6):R64. doi:10.1186/gb-2013-14-6-r64

Qi LL, Echalier B, Chao S, Lazo GR, Butler GE, Anderson OD, Akhunov ED, Dvorak J, Linkiewicz AM, Ratnasiri A, Dubcovsky J, Bermudez-Kandianis CE, Greene RA, Kantety R, La Rota CM, Munkvold JD, Sorrells SF, Sorrells ME, Dilbirligi M, Sidhu D, Erayman M, Randhawa HS, Sandhu D, Bondareva SN, Gill KS, Mahmoud AA, Ma XF, Miftahudin, Gustafson JP, Conley EJ, Nduati V, Gonzalez-Hernandez JL, Anderson JA, Peng JH, Lapitan NL, Hossain KG, Kalavacharla V, Kianian SF, Pathan MS, Zhang DS, Nguyen HT, Choi DW, Fenton RD, Close TJ, McGuire PE, Qualset CO, Gill BS (2004) A chromosome bin map of 16,000 expressed sequence tag loci and distribution of genes among the three genomes of polyploid wheat. Genetics 168 (2):701-712. doi:10.1534/genetics.104.034868

Raats D, Frenkel Z, Krugman T, Dodek I, Sela H, Simkova H, Magni F, Cattonaro F, Vautrin S, Berges H, Wicker T, Keller B, Leroy P, Philippe R, Paux E, Dolezel J, Feuillet C, Korol A, Fahima T (2013) The physical map of wheat chromosome 1BS provides insights into its gene space organization and evolution. Genome biology 14 (12). doi:Artn R138 Doi 10.1186/Gb-2013-14-12-R138

Safar J, Simkova H, Kubalakova M, Cihalikova J, Suchankova P, Bartos J, Dolezel J (2010) Development of chromosome-specific BAC resources for genomics of bread wheat. Cytogenetic and genome research 129 (1-3):211-223. doi:10.1159/000313072

Saintenac C, Falque M, Martin OC, Paux E, Feuillet C, Sourdille P (2009) Detailed Recombination Studies Along Chromosome 3B Provide New Insights on Crossover Distribution in Wheat (Triticum aestivum L.). Genetics 181 (2):393-403. doi:DOI 10.1534/genetics.108.097469

Scalabrin S, Bartos J, Febrer M, Schulte D, Paux E (2010) Guideline for physical map assembly http://www.wheatgenome.org/content/download/379/4740/file/PhysicalMapAssembly_guideline.pdf.

Soderlund C, Longden I, Mott R (1997) FPC: a system for building contigs from restriction fingerprinted clones. Computer applications in the biosciences : CABIOS 13 (5):523-535

Wang S, Wong D, Forrest K, Allen A, Chao S, Huang BE, Maccaferri M, Salvi S, Milner SG, Cattivelli L, Mastrangelo AM, Whan A, Stephen S, Barker G, Wieseke R, Plieske J, International Wheat Genome Sequencing C, Lillemo M, Mather D, Appels R, Dolferus R, Brown-Guedira G, Korol A, Akhunova AR, Feuillet C, Salse J, Morgante M, Pozniak C, Luo MC, Dvorak J, Morell M, Dubcovsky J, Ganal M, Tuberosa R, Lawley C, Mikoulitch I, Cavanagh C, Edwards KJ, Hayden M, Akhunov E (2014)

Characterization of polyploid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array. Plant biotechnology journal. doi:10.1111/pbi.12183

Wheat lag (2014). Nature 507 (7493):399-340

Zhou S, Wei F, Nguyen J, Bechner M, Potamousis K, Goldstein S, Pape L, Mehan MR, Churas C, Pasternak S, Forrest DK, Wise R, Ware D, Wing RA, Waterman MS, Livny M, Schwartz DC (2009) A single molecule scaffold for the maize genome. PLoS genetics 5 (11):e1000711. doi:10.1371/journal.pgen.1000711

## Figure legends

**Fig.1** Distribution of physical contigs along sevem 7BL deletion bins

**Fig.2** Contig ordering in 7BL_0.4-0.45 and 7BL_0.45-0.63 deletion bins showing the contribution of each map to the placement of 12 7BL physical contigs

**Fig.3** Distribution of 7BL physical contigs per genetic position in the CS*Renan population

**Fig.4** Synteny relationship between 7BL and chromosomes 1 and 3 of Brachypodium

## Supplementary material description/legends

**Supplementary table 1.** List of 7BL physical contigs with corresponding BAC clones

**Supplementary table 2**. 7B genetic map obtained from SHA3/CBRD*Naxos population

**Supplementary table 3**. 7B genetic map obtained from SY*Naxos population

**Supplementary table 4.** 7B *Triticeae* prototype map

**Supplementary table 5.** Ordering of 7BL physical contigs based on bin map, genetic maps and synteny-based map

# Tables

**Table 1. Summary statistics for the final version of LTC-based BAC contigs and 7BL MTP.**

| Statistics | **7BL** |
|---|---|
| Assumed arm size (Mbp) | 540 |
| BAC insert size | 136Kb |
| BAC fingerprints total | 72,960 |
| BACs fingerprints filtered | 60,798 |
| Virtual band size (kbp) | 1.2 |
| Number of singletons | 13,785 (23%) |
| Clones in contigs (total) | 47,013 |
| BACs in MTP | 5,229 |
| Clones in contigs with ≥6 clones | 45,087 |
| Total number of contigs | 716 |
| Number of contigs with ≥6 clones | 125 |
| estimated coverage (%)* | 87% |
| Contigs N50 (clones/Kb*) | 469/ 4,458 |
| Contigs L50 | 34 |
| Number of contigs ≥ 5 clones (MTP) | 252 |

* Based on the assumptions of arm sizes and virtual band size

**Table 2. Summary statistics for the final 7BL BAC contigs and scaffolds**

| Statistics | Contigs | Scaffolds |
|---|---|---|
| Number of sequences | 105,445 | 40,677 |
| Total size | 538 | 523 |
| N50 | 17,525 | 110,347 |
| N80 | 5,815 | 22,776 |
| average | 5,107 | 12,873 |
| No.Seq/BAC | 19 | 7 |

**Table 3. Summary statistics for 7BL MTP-contig assembly.**

| Statistics | 7BL |
|---|---|
| No. MTP-contig clones | 5561 |
| No. MTP contig | 125 |
| Clones per MTP contig/min/max | 44.8 / 1 / 219 |
| Total Kbp | 529407 |
| Arm coverage* | 98 % |

* Assuming estimated size 7BL of 540Mb

**Table 4. Chromosome7B genetic maps**

| Statistics | Re*CS | SY*Nax | SHA3/CBRD*Nax |
|---|---|---|---|
| number of markers on the linkage group | 4438 | 623 | 594 |
| number unique loci | 308 | 71 | 38 |
| genetic length cM | 127.3 | 220 | 100.67 |
| number of markers mapped to 7BL BACs | 1259 | 289 | 247 |
| 7BL supercontigs | 76 | 52 | 46 |

Table 5. Contribution of different maps to the first draft order of physical contigs along 7BL

| bin | contigs | nclones | CS*Renan | Sh3CBRDxNaxos | Sy*Naxos | *Triticeae* map | final order |
|---|---|---|---|---|---|---|---|
| 7BL_0-0.14 | ctg25 | 1082 | ctg25,ctg15, ctg18,ctg24, ctg2,ctg_13_90 | ctg42 | | ctg25 | ctg25 |
| 7BL_0-0.14 | ctg15 | 1429 | ctg42, ctg10 | ctg10 | ctg42,ctg10 | ctg15 | ctg15 |
| 7BL_0-0.14 | ctg18 | 1046 | | | | ctg18 | ctg18 |
| 7BL_0-0.14 | ctg55 | 231 | | | | ctg55 | ctg55 |
| 7BL_0-0.14 | ctg24 | 391 | | | | ctg24 | ctg24 |
| 7BL_0-0.14 | ctg2 | 1139 | | | | ctg2 | ctg2, ctg_13_90 |
| 7BL_0-0.14 | ctg_13_90 | 679 | | | | ctg42 | ctg42 |
| 7BL_0-0.14 | ctg42 | 473 | | | | ctg115 | ctg115 |
| 7BL_0-0.14 | ctg115 | 47 | | | | ctg71 | ctg71 |
| 7BL_0-0.14 | ctg114 | 48 | | | | ctg77 | ctg77 |
| 7BL_0-0.14 | ctg123 | 26 | | | | ctg10 | ctg10 |
| 7BL_0-0.14 | ctg52 | 208 | | | | | |
| 7BL_0-0.14 | ctg71 | 185 | | | | | |
| 7BL_0-0.14 | ctg77 | 181 | | | | | |
| 7BL_0-0.14 | ctg10 | 675 | | | | | |
| 7BL_0.14-0.33 | ctg69 | 198 | ctg46, ctg23 | | | ctg69 | ctg69 |
| 7BL_0.14-0.33 | ctg26 | 371 | ctg50 | | | ctg26 | ctg26 |
| 7BL_0.14-0.33 | ctg46 | 589 | | | | ctg46 | ctg46 |
| 7BL_0.14-0.33 | ctg65 | 356 | | | | ctg65 | ctg65 |
| 7BL_0.14-0.33 | ctg23 | 433 | | | | ctg23 | ctg23 |
| 7BL_0.14-0.33 | ctg50 | 290 | | | | | ctg50 |
| 7BL_0.33-0.4 | ctg83 | 136 | ctg83 | ctg_41_16_143 | ctg_41_16_143,ctg_6_88 | ctg_41_16_143 | ctg83 |
| 7BL_0.33-0.4 | ctg_41_16_143 | 887 | ctg_6_88 | ctg_6_88 | | ctg_6_88 | ctg_41_16_143 |
| 7BL_0.33-0.4 | ctg_6_88 | 898 | ctg47, ctg7 | | | ctg5 | ctg_6_88 |
| 7BL_0.33-0.4 | ctg5 | 500 | | | | ctg_17_67 | ctg5 |
| 7BL_0.33-0.4 | ctg_17_67 | 386 | | | | ctg38 | ctg_17_67 |
| 7BL_0.33-0.4 | ctg38 | 170 | | | | ctg47 | ctg38 |
| 7BL_0.33-0.4 | ctg47 | 771 | | | | ctg7 | ctg47 |
| 7BL_0.33-0.4 | ctg101 | 66 | | | | | ctg7 |
| 7BL_0.33-0.4 | ctg113 | 104 | | | | | |
| 7BL_0.33-0.4 | ctg133 | 9 | | | | | |
| 7BL_0.33-0.4 | ctg4 | 91 | | | | | |
| 7BL_0.33-0.4 | ctg40 | 113 | | | | | |
| 7BL_0.33-0.4 | ctg7* | 966 | | | | | |
| 7BL_0.4-0.45 | ctg9 | 871 | ctg9 | ctg_34_53 | ctg9, ctg_30_74 | ctg9 | ctg9 |
| 7BL_0.4-0.45 | ctg84 | 136 | ctg84 | ctg_30_74 | | ctg21 | ctg84 |
| 7BL_0.4-0.45 | ctg21 | 433 | ctg21 | | | ctg_34_53 | ctg21 |
| 7BL_0.4-0.45 | ctg_34_53 | 879 | ctg_34_53 | | | ctg_30_74 | ctg_34_53 |
| 7BL_0.4-0.45 | ctg_30_74* | 1474 | ctg_30_74 | | | | ctg_30_74 |
| 7BL_0.45-0.63 | ctg8 | 1112 | ctg8 | ctg8,ctg_12_33 | ctg8, ctg_12_33 | ctg8 | ctg8 |
| 7BL_0.45-0.63 | ctg_12_33 | 1585 | ctg_12_33 | ctg39 | ctg39, ctg100 | ctg_12_33 | ctg_12_33 |
| 7BL_0.45-0.63 | ctg61 | 298 | ctg39 | ctg100 | | ctg61 | ctg61 |
| 7BL_0.45-0.63 | ctg39 | 732 | | | | ctg39 | ctg39 |
| 7BL_0.45-0.63 | ctg100 | 72 | | | | ctg100 | ctg100 |
| 7BL_0.45-0.63 | ctg96 | 91 | | | | | |
| 7BL_0.63-0.69* | ctg_43_116 | 638 | ctg_43_116 | ctg_43_116 | ctg_43_116 | ctg_43_116 | ctg_43_116 |
| 7BL_0.63-0.69* | ctg87 | 222 | ctg22 | ctg22 | ctg22 | ctg87 | ctg87 |
| 7BL_0.63-0.69* | ctg22 | 1413 | ctg85 | ctg11 | ctg11 | ctg22 | ctg22 |
| 7BL_0.63-0.69* | ctg85 | 190 | ctg11 | ctg117, ctg99 | ctg117, ctg99 | ctg85 | ctg85 |
| 7BL_0.63-0.69* | ctg11 | 643 | ctg1 | ctg54, ctg45,ctg49 | ctg45, ctg49, ctg51,ctg80 | ctg11 | ctg11 |
| 7BL_0.63-0.69* | ctg117 | 49 | ctg28 | ctg51 | ctg_103_cl697 | ctg110 | ctg117 |
| 7BL_0.63-0.69* | ctg110 | 97 | ctg_78 | ctg_103_cl697 | ctg_72_36_37,ctg93,ctg68 | ctg99 | ctg110 |
| 7BL_0.63-0.69* | ctg99 | 75 | ctg54 | ctg35 | ctg35 | ctg1 | ctg99 |
| 7BL_0.63-0.69* | ctg1 | 920 | ctg45 | ctg_73_109,ctg27,ctg44 | ctg44 | ctg28 | ctg1 |
| 7BL_0.63-0.69* | ctg28 | 1945 | ctg128 | | ctg31, ctg29 | ctg_78 | ctg28 |
| 7BL_0.63-0.69* | ctg_78 | 174 | ctg49 | | | ctg54 | ctg_78 |
| 7BL_0.63-0.69* | ctg54 | 452 | ctg112 | | | ctg45 | ctg54 |
| 7BL_0.63-0.69* | ctg45 | 272 | ctg70 | | | ctg51 | ctg45 |
| 7BL_0.63-0.69* | ctg128 | 20 | ctg51 | | | ctg_103_cl697 | ctg128 |
| 7BL_0.63-0.69* | ctg49 | 494 | ctg_103_cl697, ctg76, ctg80 | | | ctg76 | ctg49 |
| 7BL_0.63-0.69* | ctg112 | 54 | ctg_72_36_37 | | | ctg80 | ctg112 |
| 7BL_0.63-0.69* | ctg70 | 230 | ctg68 | | | ctg_72_36_37 | ctg70 |
| 7BL_0.63-0.69* | ctg51 | 300 | ctg35 | | | ctg35 | ctg51 |
| 7BL_0.63-0.69* | ctg_103_cl697 | 246 | ctg_73_109 | | | ctg_73_109 | ctg_103_cl697 |
| 7BL_0.63-0.69* | ctg76 | 178 | ctg27 | | | ctg27 | ctg76 |
| 7BL_0.63-0.69* | ctg80 | 147 | ctg44 | | | ctg44 | ctg80 |
| 7BL_0.63-0.69* | ctg93 | 65 | ctg31 | | | ctg31 | ctg_72_36_37 |
| 7BL_0.63-0.69* | ctg_72_36_37 | 963 | ctg29,ctg_79a_129 | | | ctg29 | ctg93,ctg68 |
| 7BL_0.63-0.69* | ctg68 | 359 | | | | ctg_79a_129 | ctg35 |
| 7BL_0.63-0.69* | ctg35 | 1454 | | | | | ctg_73_109 |
| 7BL_0.63-0.69* | ctg_73_109 | 307 | | | | | ctg27 |
| 7BL_0.63-0.69* | ctg27 | 760 | | | | | ctg44 |
| 7BL_0.63-0.69* | ctg44 | 272 | | | | | ctg31 |
| 7BL_0.63-0.69* | ctg31 | 404 | | | | | ctg29 |
| 7BL_0.63-0.69* | ctg29 | 436 | | | | | ctg_79a_129 |
| 7BL_0.63-0.69* | ctg_79a_129 | 101 | | | | | |
| 7BL_0.63-0.69* | ctg122 | 27 | | | | | |
| 7BL_0.63-0.69* | ctg127 | 23 | | | | | |
| 7BL_0.63-0.69* | ctg135 | 8 | | | | | |
| 7BL_0.69*-1.00 | ctg_79b | 502 | ctg_79b | ctg_79b | ctg_79b | ctg_79b | ctg_79b |
| 7BL_0.69*-1.00 | ctg66 | 205 | ctg66 | ctg66 | ctg66 | ctg66 | ctg66 |
| 7BL_0.69*-1.00 | ctg119 | 40 | ctg19 | ctg119 | ctg48 | ctg19 | ctg119 |
| 7BL_0.69*-1.00 | ctg19 | 449 | ctg_81_86_cl4550 | ctg_81_86_cl4550,ctg48 | ctg62 | ctg48 | ctg19 |
| 7BL_0.69*-1.00 | ctg_81_86_cl4550 | 346 | ctg48 | ctg62,ctg20 | ctg_75_98a | ctg62 | ctg_81_86_cl4550 |
| 7BL_0.69*-1.00 | ctg48 | 451 | ctg124 | ctg_75_98a,ctg_82_89 | ctg_82_89 | ctg20 | ctg48 |
| 7BL_0.69*-1.00 | ctg124 | 53 | ctg3 | ctg_32_121_106_60 | ctg_32_121_106_60 | ctg_75_98a | ctg124 |
| 7BL_0.69*-1.00 | ctg3 | 75 | ctg62 | ctg59 | ctg59,ctg131,ctg63 | ctg_82_89 | ctg3 |
| 7BL_0.69*-1.00 | ctg62 | 683 | ctg20 | ctg64,ctg58,ctg118,ctg97 | ctg118 | ctg_32_121_106_60 | ctg62 |
| 7BL_0.69*-1.00 | ctg20 | 728 | ctg_75_98a | ctg_98b_57_108 | ctg97 | ctg59 | ctg20 |
| 7BL_0.69*-1.00 | ctg_75_98a | 460 | ctg_82_89 | | ctg_98b_57_108 | ctg63 | ctg_75_98a |
| 7BL_0.69*-1.00 | ctg_82_89 | 501 | ctg94 | | | ctg64 | ctg_82_89 ctg94 |
| 7BL_0.69*-1.00 | ctg94 | 107 | ctg_32_121_106_60 | | | ctg58 | ctg_32_121_106_60 |
| 7BL_0.69*-1.00 | ctg_32_121_106_60 | 681 | ctg111,ctg126 | | | ctg118 | ctg111,ctg126 |
| 7BL_0.69*-1.00 | ctg111 | 70 | ctg59 | | | ctg97 | ctg59 |
| no | ctg126 | 23 | ctg63 | | | ctg_98b_57_108 | ctg63 ctg131 |
| 7BL_0.69*-1.00 | ctg59 | 377 | ctg56 | | | | ctg56 |
| no | ctg131 | 11 | ctg64 | | | | ctg64 |
| 7BL_0.69*-1.00 | ctg63 | 206 | ctg58 | | | | ctg58 |
| no | ctg56 | 298 | ctg118 | | | | ctg118 |
| 7BL_0.69*-1.00 | ctg64 | 203 | ctg132 | | | | ctg132 |
| 7BL_0.69*-1.00 | ctg58 | 47 | ctg97 | | | | ctg97 |
| 7BL_0.69*-1.00 | ctg118 | 44 | ctg107 | | | | ctg107 |
| no | ctg132 | 9 | ctg_98b_57_108 | | | | ctg_98b_57_108 |
| 7BL_0.69*-1.00 | ctg97 | 153 | ctg105 | | | | ctg105 |
| 7BL_0.69*-1.00 | ctg107 | 56 | ctg102 | | | | ctg102 |
| 7BL_0.69*-1.00 | ctg_98b_57_108 | 350 | | | | | |
| 7BL_0.69*-1.00 | ctg105 | 63 | | | | | |
| 7BL_0.69*-1.00 | ctg102 | 66 | | | | | |
| 7BL_0.69*-1.00 | ctg95 | 93 | | | | | |

physical contigs anchored based on evidence from bin map, genetic map(s) and synteny map

physical contigs anchored based on evidence from two resources (bin map and genetic or synteny map, or synteny and genetic maps)

physical contigs have only bin position

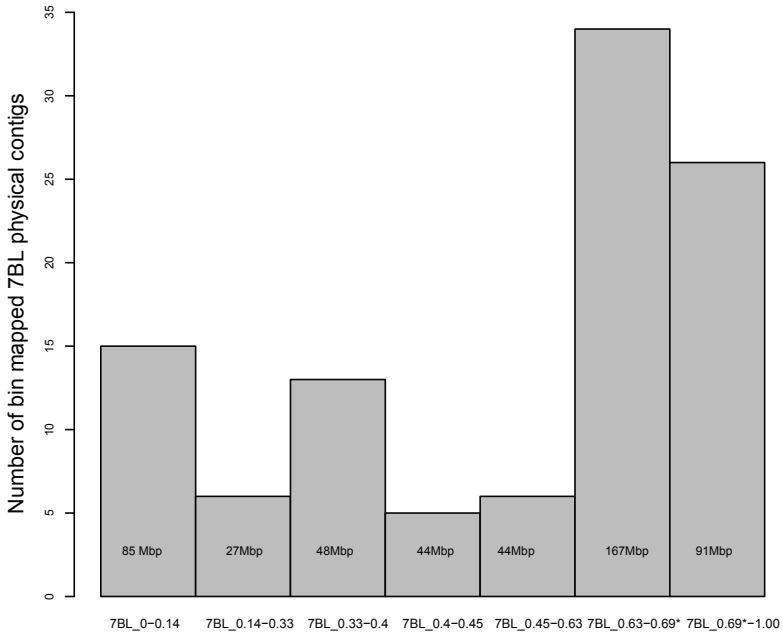
Number of bin mapped 7BL physical contigs
0
5
10
15
20
25
30
35
85 Mbp
27Mbp
48Mbp
44Mbp
44Mbp
167Mbp
91Mbp
7BL_0−0.14
7BL_0.14−0.33
7BL_0.33−0.4
7BL_0.4−0.45
7BL_0.45−0.63
7BL_0.63−0.69*
7BL_0.69*−1.00

**Fig.1**

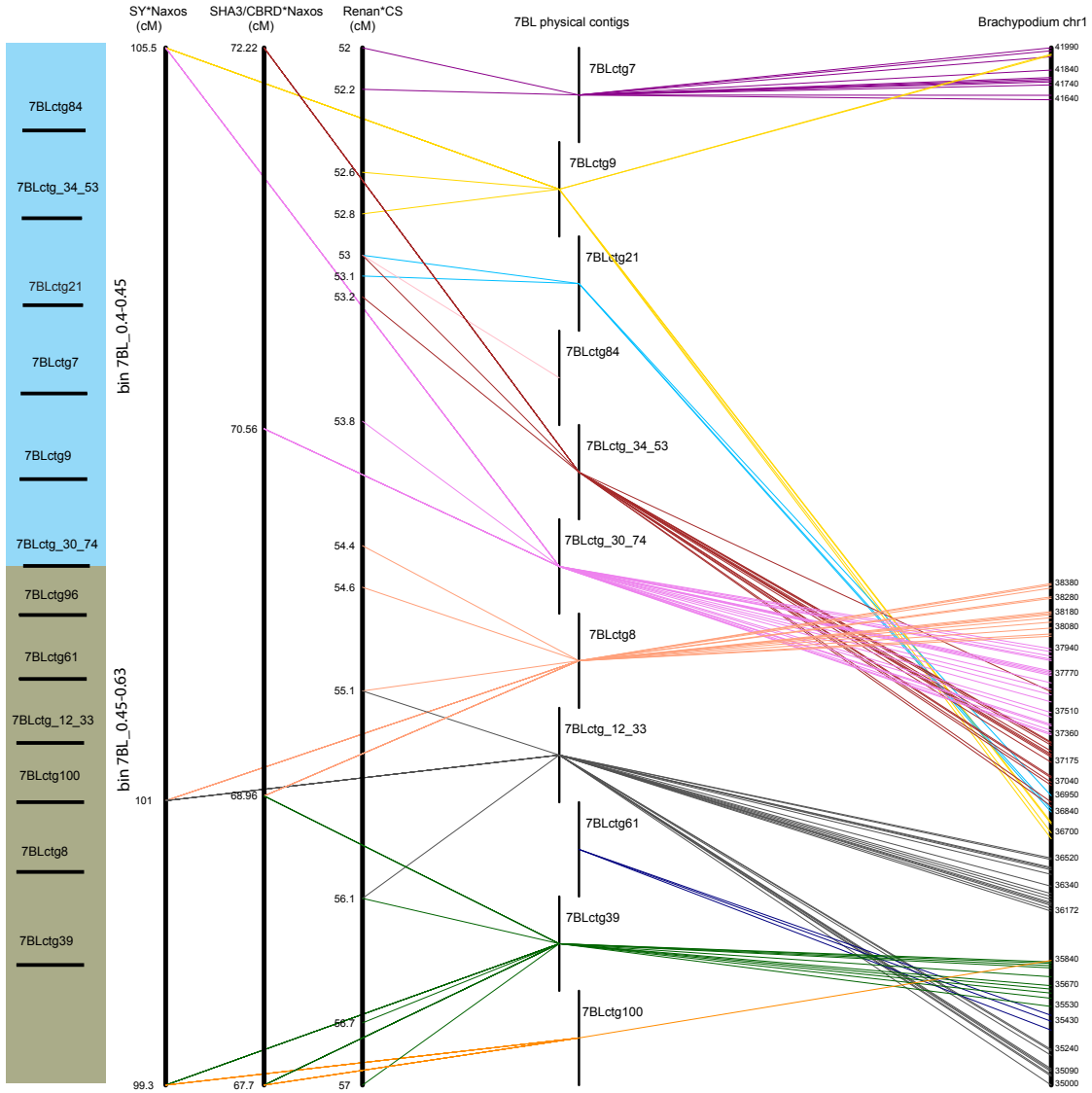
SY*Naxos (cM)
SHA3/CBRD*Naxos (cM)
Renan*CS (cM)
7BL physical contigs
Brachypodium chr1
7BLctg84
7BLctg_34_53
7BLctg21
7BLctg7
7BLctg9
7BLctg_30_74
7BLctg96
7BLctg61
7BLctg_12_33
7BLctg100
7BLctg8
7BLctg39
bin 7BL_0.4-0.45
bin 7BL_0.45-0.63
105.5
101
99.3
72.22
68.96
67.7
70.56
52
52.2
52.6
52.8
53
53.1
53.2
53.8
54.4
54.6
55.1
56.1
56.7
57
7BLctg7
7BLctg9
7BLctg21
7BLctg84
7BLctg_34_53
7BLctg_30_74
7BLctg8
7BLctg_12_33
7BLctg61
7BLctg39
7BLctg100
41990
41840
41740
41640
38380
38280
38180
38080
37940
37770
37510
37360
37175
37040
36950
36840
36700
36520
36340
36172
35840
35670
35530
35430
35240
35090
35000

**Fig.2**

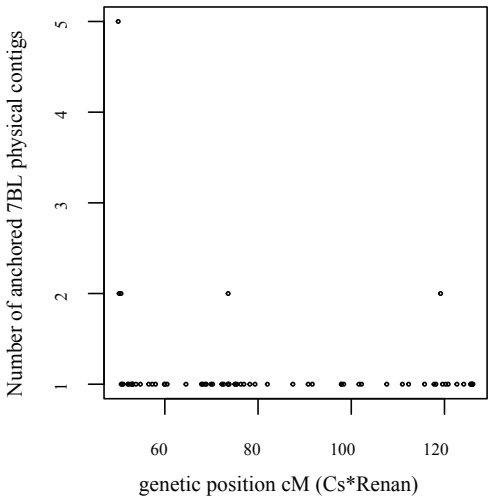
Number of anchored 7BL physical contigs
1
2
3
4
5
60
80
100
120
genetic position cM (Cs*Renan)

**Fig.3**

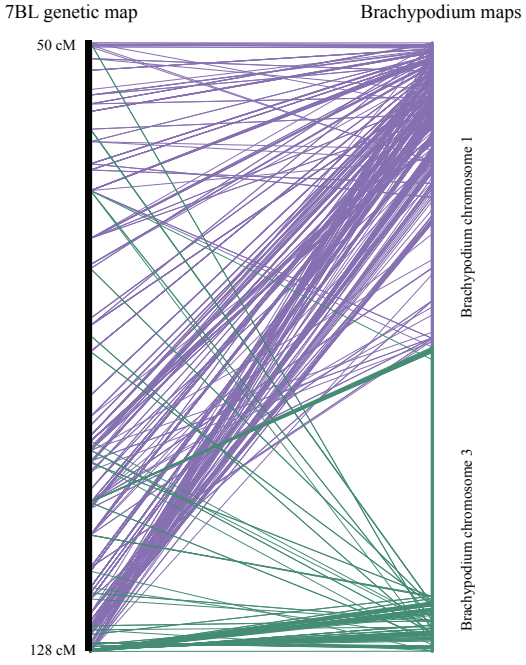
7BL genetic map
Brachypodium maps
50 cM
128 cM
Brachypodium chromosome 1
Brachypodium chromosome 3

**Fig.4**