Palacký University Olomouc

Faculty of Science

Department of Cell Biology and Genetics

and

Institute of Experimental Botany AS CR

Centre of Plant Structural and Functional Genomics

Olomouc

Zuzana Tulpová

Sequence and functional analysis of the short arm of wheat chromosome 7D

Ph.D. Thesis

Supervisor: Ing. Hana Šimková, CSc.

Olomouc 2019

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Ing. Hana Šimková, CSc., for professional guidance, invaluable advices, inspirational ideas, help and patience. Further, I would like to thank to prof. Ing. Jaroslav Doležel, DrSc., the head of the laboratory, for the opportunity to work in the Centre of Plant Structural and Functional Genomics, to my closest colleague Mgr. Helena Toegelová, Ph.D. for her help and friendship, to my colleague Mgr. Eva Hřibová, Ph.D. for her willingness to help me at any time and also to all members of the team for an amazing and motivating atmosphere. Last but not least, I would like to thank my family for all their love and support during my whole studies.

Declaration

I hereby declare that I have written the Ph.D. thesis independently under the supervision of Ing. Hana Šimková, CSc. using merely the information sources listed in the References.

……..

This work was supported by the Czech Science Foundation (award No. P501/12/2554) and the grant LO1204 from the National Program of Sustainability I.

BIBLIOGRAPHICAL IDENTIFICATION

Author's name: Zuzana Tulpová

Title: Sequence and functional analysis of the short arm of wheat chromosome 7D

Type of thesis: Ph.D. thesis

Department: Department of Cell Biology and Genetics

Supervisor: Ing. Hana Šimková, CSc.

The Year of Presentation: 2019

Abstract:

Bread wheat (*Triticum aestivum* L.) is a staple food for ~40 % of world's population and belongs, together with rice and corn, among the most important crop species. Until recently, any wheat genomics study was a challenge, mainly due to its huge genome size (~16 Gb), high content of repetitive sequences (85 %) and presence of three sub-genomes. Aiming to overcome these obstacles and obtain a complete wheat reference genome sequence, the International Wheat Genome Sequencing Consortium proposed a strategy involving separation of individual chromosomes and their arms by flow cytometry. These were used to construct chromosome-specific BAC libraries and BAC-based physical maps, which became the basic resources for sequencing the wheat genome.

Within a framework of this thesis, I focused on the short arm of wheat chromosome 7D (7DS). A previously constructed 7DS physical map was used to select the minimal set of BAC clones covering the arm, which were sequenced by Illumina pair-end and mate-pair sequencing and assembled. Contigs of the physical map were anchored on the chromosome using one radiation hybrid map and three high resolution genetic maps. Thus we assigned 73 % of the assembly to distinct genomic positions. The process of physical-map assembly and anchoring included the integration of the 7DS physical map with a whole-genome physical map of *Aegilops tauschii* and a 7DS Bionano genome (BNG) map, which together enabled efficient scaffolding of physical-map contigs even in the non-recombining region of the genetic centromere. Moreover, this approach facilitated a comparison of bread wheat and its ancestor at BAC-contig level and revealed

a reconstructed region in the 7DS pericentromere. The obtained 7DS physical map, BAC assemblies and the BNG map were then applied as supporting resources for assembling and validating the reference genome of bread wheat and for a gene cloning project.

The chromosome arm 7DS carries multiple genes underlying agronomically important traits, including a Russian wheat aphid resistance gene *Dn2401*, the cloning of which was the aim of the second part of the thesis. Previously, we mapped *Dn2401* into an interval of 0.83 cM and spanned it with five BAC clones. Within the framework of this thesis, we used a targeted strategy combining traditional approaches towards gene cloning, comprising genetic mapping and Illumina sequencing of BAC clones, with novel technologies, including optical mapping and long-read nanopore sequencing. The latter, with reads spanning the entire length of a BAC insert, enabled us to assemble the whole region, the task not achievable with short reads. Long-read optical mapping validated the DNA sequence in the interval and revealed a difference in the locus organization between resistant and susceptible genotypes. The complete and accurate sequence of the *Dn2401* region facilitated identification of new markers and precise annotation of the interval, revealing six high-confidence genes, including *Epoxide hydrolase 2* as the most likely *Dn2401* candidate.

Keywords: Bread wheat, *Triticum aestivum*, physical contig map, BAC, sequencing, Russian wheat aphid, *Diuraphis noxia*, resistance, gene cloning

Number of Pages/Appendices: 76/VII

Language: English

BIBLIOGRAFICKÉ IDENTIFIKACE

Jméno: Zuzana Tulpová

Název práce: Sekvence a funkční analýza krátkého ramene chromozómu 7D pšenice

Typ práce: Disertační práce

Katedra: Katedra buněčné biologie a genetiky

Školitel: Ing. Hana Šimková, CSc.

Rok obhajoby: 2019

Abstrakt:

Pšenice setá (*Triticum aestivum* L.) je zdrojem potravy pro přibližně 40 % světové populace a patří tak společně s rýží a kukuřicí k nejvýznamnějším zemědělským plodinám. V nedávné minulosti představovalo jakékoli studium pšeničného genomu velkou výzvu, a to především díky značné velikosti genomu (~16 Gb), vysokému obsahu repetitivních sekvencí (85 %) a přítomnosti tří homeologních subgenomů. Řešení těchto problémů přinesla strategie založená na třídění jednotlivých chromozómů a jejich ramen pomocí průtokové cytometrie, z nichž byly následně tvořeny knihovny dlouhých insertů a fyzické mapy, jež představovaly základní genomové zdroje pro získání kompletní referenční sekvence.

Předkládaná práce se zabývá studiem krátkého ramene chromozómu 7D pšenice. S pomocí již dříve sestavené fyzické mapy 7DS byla vybrána minimální sestava BAC klonů, která reprezentuje celé rameno 7DS (tzv. *minimal tilling path*, MTP). Klony MTP pak byly osekvenovány pomocí platformy Illumina a sestaveny do sekvenčních kontigů. Pro ukotvení kontigů fyzické mapy byly využity tři genetické mapy s vysokým rozlišením a mapa radiačních hybridů. Díky tomu se podařilo určit konkrétní genomovou pozici pro 73 % sestavené mapy. Jedním z přístupů pro ukotvení fyzické mapy bylo i využití integrace fyzické mapy ramene 7DS s celogenomovou fyzickou mapou *Aegilops tauschii* a Bionano optickou mapou ramene 7DS. Kombinace těchto zdrojů umožnila efektivní ukotvování kontigů fyzické mapy, včetně nerekombinující oblasti genetické centromery, ale i bezprostřední porovnání pšenice s jejím předchůdcem. Díky tomu se podařilo identifikovat přestavby na úrovni BAC kontigů v pericentromerické oblasti ramene 7DS. Kromě toho byly ukotvená fyzická mapa, sekvence BAC klonů a optická mapa ramene

7DS využity jako podpůrné zdroje pro sestavení a kontrolu referenční sekvence pšenice a navazující projekt pozičního klonování.

Rameno 7DS je nositelem celé řady genů pro agronomicky významné znaky, mezi něž patří i gen *Dn2401* podmiňující rezistenci ke mšici zhoubné. Závěrečná část dizertační práce je věnována projektu pozičního klonování tohoto genu. V předchozí studii byla pomocí genetického mapování stanovena velikost intervalu, v němž se gen *Dn2401* nachází, na 0.83 cM. Současně bylo identifikováno pět BAC klonů fyzické mapy, které daný interval překlenují. V rámci předkládané práce byla zvolena strategie kombinující tradiční přístupy klonování genů, jako jsou genetické mapování a sekvenování BAC klonů na platformě Illumina, s novými technologiemi zahrnujícími optické mapování nebo sekvenování pomocí nanoporů. Takto získaná dlouhá čtení, která překlenula celý insert sekvenovaného BAC klonu, pak umožnila sestavení kontinuální sekvence celého intervalu. Optické mapování následně potvrdilo správnost sekvence a odhalilo rozdíly mezi rezistentním a citlivým genotypem. Kompletní a přesná sekvence zájmové oblasti umožnila identifikaci nových markerů a detailní anotaci, která odhalila šest genů kódujích proteiny, včetně *Epoxid hydrolázy 2* jakožto nejpravděpodobnějšího kandidáta genu *Dn2401*.

Klíčová slova: Pšenice setá, *Triticum aestivum*, fyzická kontigová mapa, BAC, mšice zhoubná, *Diuraphis noxia*, rezistence, poziční klonování genů

Počet stran/příloh: 76/VII

Jazyk: Anglický

CONTENT

1 LITERATURE OVERVIEW

1.1 Bread wheat (*Triticum aestivum* L.)

Bread wheat (*Triticum aestivum* L.) is a monocotyledonous species belonging to the family *Poaceae*, subfamily *Pooideae* and the tribe *Triticeae*. It is an allohexaploid species (2n = 6x = 42), whose genome arose through spontaneous hybridization of three progenitors resulting into a highly complex genome of ~16 Gb (Doležel *et al*., 2018) consisting of three homoeologous sub-genomes A, B and D and comprising over 85 % repetitive DNA (IWGSC, 2018).

About 10 thousand years ago, early farming practices made use of wild diploid wheat species from *Aegilops* and *Triticum* genera, but as agriculture advanced, wild wheats were gradually substituted with domesticated diploid and polyploid wheat species, including hexaploid bread wheat (Marcussen *et al*., 2014). Thus bread wheat represents one of the first domesticated crops, which has spread worldwide due to its adaptability to a wide range of climatic conditions. Nowadays, wheat is the most widely cultivated cereal in the world with more than 220 million ha of harvested area (http://www.fao.org/faostat/en/#data/FBS). The rapid growth of the productivity of wheat, sometimes designated as the "miracle crop" of the last century (Shiferaw *et al*., 2013), has helped to overcome famines and thus wheat has become the most important food grain source for 40 % world's population providing about 20 % of the total dietary calories and proteins worldwide. Even though wheat is grown on more land area than any other commercial crop and its production reached almost 750 million tonnes in 2016 (Figure 1), and even 7 million tonnes more in 2017 (http://www.fao.org/worldfoodsituation/csdb/en/), the first place in world cereal production belongs to maize (Figure 1). Majority (about 92 %) of wheat production belong to the bread wheat (*T. aestivum*), utilized mainly as a flour for manufacturing a wide variety of baking products. The remaining 8 % are represented by durum wheat (*T. turgidum* ssp. *durum*), which is used for making pasta or other semolina products (https://www.ers.usda.gov/data-products/wheat-data/). Einkorn wheat (*T. monococcum*) and other hulled wheats, namely emmer (*T. dicoccum*) and spelt (*T. spelta*) are considered relic crops of minor economic importance (Dubcovsky and Dvorak, 2007).

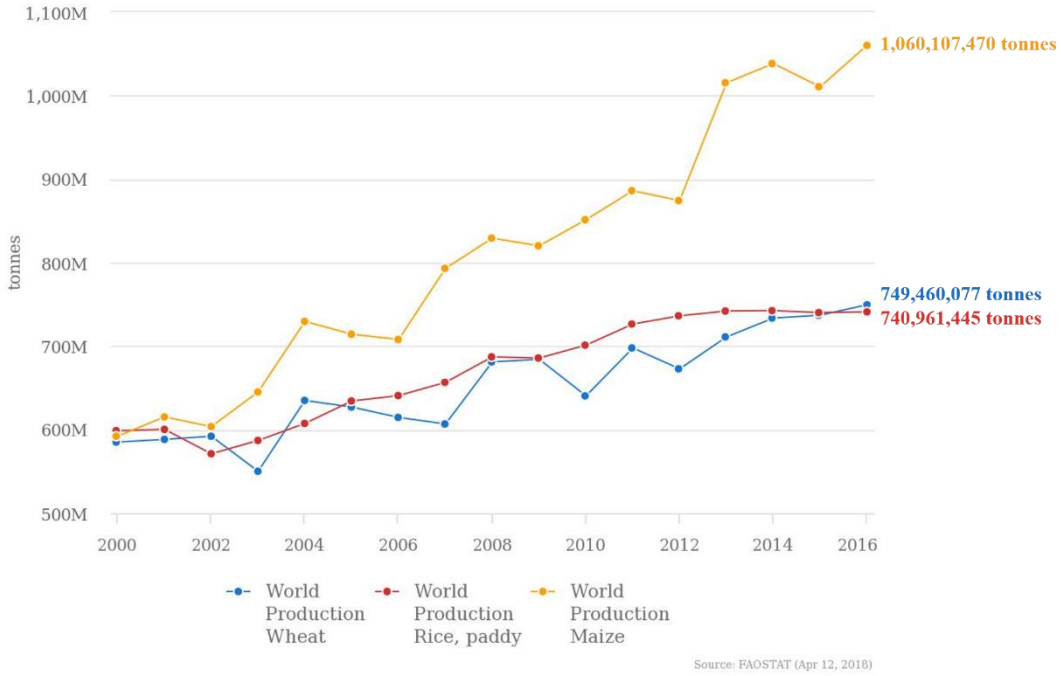


Figure 1: Comparison of wheat, maize and rice world production between 2000 and 2016 (http://www.fao.org/faostat/en/#data/FBS).

1.1.1 Evolution of the bread wheat genome

The modern bread wheat is a result of cultivation and domestication that began together with the progress and spread of agriculture in the area of Fertile Crescent (Marcussen *et al*., 2014). However, the origin of the bread wheat genome can be dated back to millions years ago (MYA) when independently evolved three ancestral diploid species, each with seven pairs of chromosomes – *Triticum urartu* (genome AA), species from Sitopsis section related to *Aegilops speltoides* (genome BB) and *Aegilops tauschii* (genome DD) (IWGSC, 2014). Bread wheat is an allohexaploid species arisen through two spontaneous hybridization events, each accompanied by polyploidization. The first hybridization occurred <0.82 MYA ago, resulting in allotetraploid *Triticum turgidum* (2n = 4x = 28; AABB) (Marcussen *et al.*, 2014), an ancestor of wild emmer wheat cultivated in the Middle East, which gave rise to *T. turgidum* ssp. *durum* grown for pasta nowadays (IWGSC, 2014). The second hybridization, occurring <0.43 MYA (Marcussen *et al.*, 2014), combined *T. turgidum* and diploid goatgrass *Ae. tauschii* (Figure 2I) to produce the huge and complex hexaploid wheat genome with individual chromosomes bigger than the whole rice genome. All three progenitors have large

genomes, about 5.5 Gb in size, that contributed to the final wheat genome size of about 16 Gb. Subsequent to the hybridization and polyploidization events, a number of structural and functional rearrangements led to genome stabilization (Feldman and Levy, 2009).

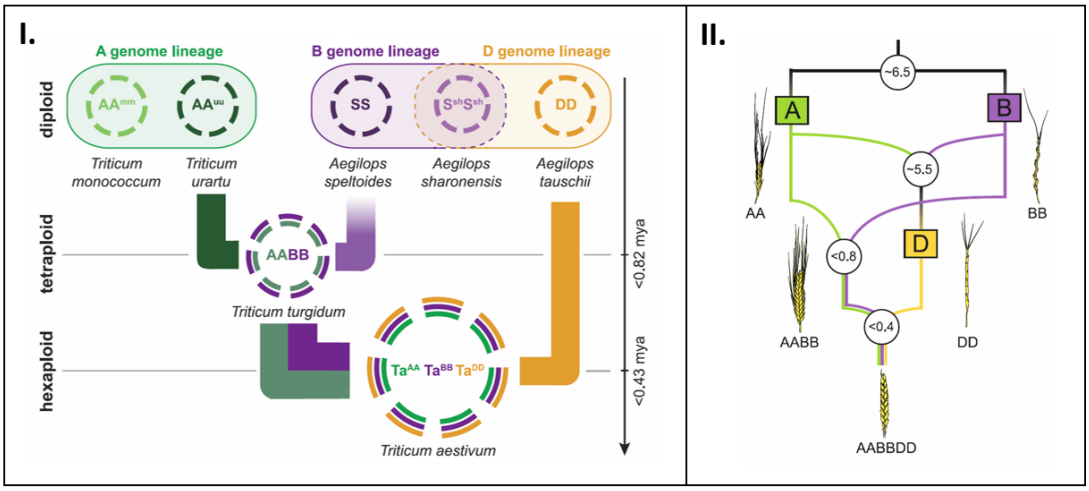


Figure 2: Two models of bread wheat genome (AABBDD) origin. I) Standard model of bread wheat genome origin (IWGSC, 2014). II) Model of wheat phylogenetic history including hypothetical D genome origin by homoploid hybrid speciation (Marcussen *et al.*, 2014).

An alternative, quite controversial hypothesis about the origin of bread wheat is described in Marcussen *et al.* (2014). In that study, authors used draft genome assemblies of bread wheat and its diploid relatives to analyse genome-wide samples of gene trees, as well as to estimate evolutionary relatedness and divergence times. First of all, they revised dating of the divergence time between the A and B lineages from initial ~4.0 MYA (Feldman *et al.*, 2005) to ~6.5 MYA, as well as dating of the hybridizations events. Moreover, they proposed three hybridization events leading to the origin of hexaploid wheat genome (Figure 2II). The first hybridization occurred about ~5.5 MYA between A and B lineages and led to the origin of the D lineage by homoploid hybrid speciation. The second hybridization between two diploid closely related ancestors, *T. urartu* (AA) and a relative (BB) of *Ae. speltoides*, occurred about 4.7 million years later and gave rise to the allotetraploid emmer wheat, *T. turgidum* (AABB). This tetraploid wheat subsequently hybridized with *Ae. tauschii*, the diploid progenitor of the D genome. This model was questioned by Li *et al.* (2015) who proposed that the history of *Ae. tauschii* might be even more

reticulated than suggested by Marcussen *et al.* (2014), involving several rounds of both recent and ancient hybridization events between the *Aegilops* and *Triticum* species.

Despite bread wheat is a structurally allopolyploid species with three sets of homeologous chromosomes belonging to the three sub-genomes, it behaves during meiosis like a diploid with pairing and crossovers between homologous chromosomes only. This diploid-like behaviour is preserved due to the action of *Ph1* gene lying on chromosome 5B (Martinez-Perez *et al.*, 2001).

Comparing the bread wheat gene sequences with gene repertoires of wheat's closest relatives showed a gene loss during the evolution of the hexaploid wheat genome and frequent gene duplications after these genomes had got together (Eversole *et al.*, 2014). Smaller gene loss and significantly lower number of pseudogenes were observed in the D sub-genome (81,905) compared to the A and B sub-genomes (99,754 and 109,097 pseudogenes, respectively; IWGSC, 2018) (Figure 3). Gene flow between the two species of A and B lineages resulted in greater sequence diversity within the A and B sub-genomes when compared with the D sub-genome (Dvořák *et al*., 2006). The D sub-genome shows a significantly lower level of polymorphism, which may be due to a differential loss of low frequency alleles during the population size bottleneck that accompanied the development of modern commercial cultivars (Chao *et al*. 2006) or to its more recent hybridization with the AABB genome.

1.1.2 Composition of the bread wheat genome

Components of the wheat genome can be divided into two main parts, each with different dynamics of evolution and importance. The first one is the conservative part that is subjected to selection pressure and mostly corresponds to the gene-space. The second part is the much larger and variable component that is under more dynamic evolution and comprises the transposable-element (TE) space as well as duplicated genes and gene fragments (Choulet *et al*., 2014b).

Knowledge of the whole genome reference sequence (IWGSC, 2018) allowed identifying 107,891 high confidence (HC) protein-coding genes with relatively equal distribution across A, B and D sub-genomes (35,345, 35,643 and 34,212, respectively). In addition, 161,537 other putative protein-coding loci were classified as low-confidence (LC) genes, representing partially supported gene models, gene fragments and orphans (IWGSC, 2018) (Figure 3). Interestingly, individual sub-genomes show

high degree of regulatory and transcriptional autonomy (Pfeifer *et al*, 2014), but no evidence was found for sub-genome dominance in gene expression (IWGSC, 2018), as observed for maize and other grasses (Schnable *et al*., 2012; Woodhouse *et al.*, 2010).

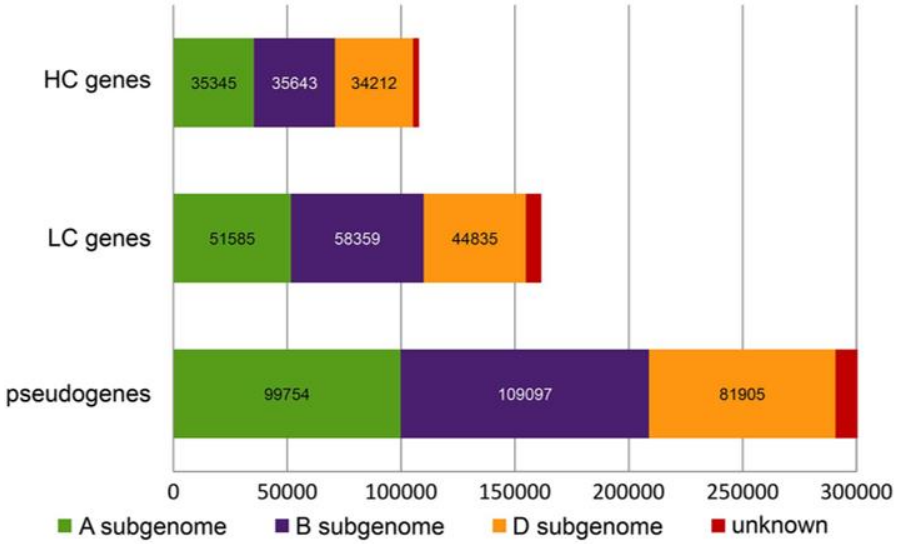


Figure 3: Gene annotation of the bread wheat genome and gene distribution across the sub-genomes (IWGSC, 2018).

Even though genes are distributed along the entire chromosome length with no gene deserts larger than a few Mb (Choulet *et al*, 2014b), they represent only a minor part (< 2 %, IWGSC, 2014) of the whole wheat genome. The majority of the genome is occupied by transposable elements, which are inventoried in several dedicated databases, including TREP (Transposable Elements Platform), listing TEs from wheat, barley and rye (Wicker *et a*l, 2002), in curated databank of repeated elements – ClariTeRep (github.com/jdaron/CLARI-TE/) or in PGSB Repeat Element database (PGSB-REdat) (pgsb.helmholtz-muenchen.de/plant/index.jsp). Wicker *et al.*, (2018), in a complex study based on the IWGSC RefSeq v1.0 assembly, identified total of 3,968,974 TEs, belonging to 505 families and representing 85 % of the IWGSC RefSeq v1.0. Though sizes of the wheat sub-genomes differ (4.93 Gb, 5.18 Gb and 3.95 Gb for A, B and D, respectively), the TE proportion is similar in all three sub-genomes, as they represented 86, 85 and 83 % of the A, B and D sub-genome sequence, respectively. The repetitive fraction is mostly dominated by TEs belonging

to class I *Gypsy* and *Copia* and class II *CACTA* superfamilies. Other superfamilies contribute to the overall genome size only marginally (Figure 4). The smaller size of the D sub-genome (about 1 Gb less than the A and B) is mainly due to a smaller amount of *Gypsy* elements. Interestingly, particular TE families show strong preferences for distinct chromosomal "niches" such as centromeric and telomeric regions. Despite majority of TEs annotated in the IWGSC RefSeq v1.0 showed relatively equal distribution across the three sub-genomes, a few were found associated with individual sub-genomes, such as *DTC_Pavel* family (*CACTA* element), which evolved centromere specificity in the D genome lineage (Wicker *et al*., 2018).

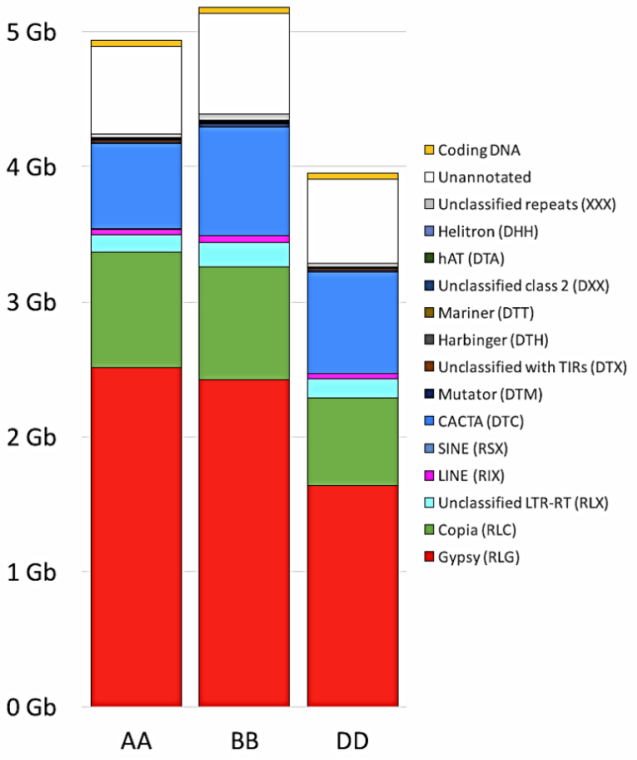


Figure 4: Composition of wheat sub-genomes with respect to particular TE superfamilies, coding DNA (yellow) and unannotated sequences (white) according to Wicker *et al*. (2018).

1.2 Sequencing of the wheat genome

The ambitious project of whole-genome sequencing of the 16-Gb bread wheat genome began as early as in 2005 with founding of the International Wheat Genome Sequencing Consortium (IWGSC). The effort towards getting the wheat reference genome took 12 years and was characterized by changes in sequencing and assembling

strategies, that reflected advances in sequencing technologies and assembling pipelines. The final outcome of this joint effort, published as IWGSC RefSeq v1.0 (IWGSC, 2018), arose from combination of multiple genomic resources and approaches described in the following chapters.

1.2.1 Reduction of genome complexity

For a long time, whole-genome sequencing of the bread wheat genome was considered a highly challenging task due to its huge size, polyploidy and a high content of repetitive DNA. The earliest efforts towards reducing the wheat genome complexity were focusing on genomes of its diploid relatives, taking advantage of smaller genome sizes and absence of polyploidy. Several BAC libraries of diploid wheat progenitors have been constructed (Akhunov *et al*., 2005; Lijavetzky *et al.*, 1999) and employed in cloning agronomically important genes. However, wheat genomes have undergone revolutionary changes following the polyploidization events including losses of DNA. This partial diploidization and other genomic changes suggest that wheat diploid ancestors, although useful resources for wheat genomics, cannot fully substitute for the genomic sequence of hexaploid wheat itself (Šimková *et al*., 2011).

Dissection of the nuclear genome into smaller well defined parts such as chromosomes or chromosomal arms, enabling ~20-40fold complexity reduction, appeared the most efficient solution for the enormous genome complexity. Microdissection of mitotic chromosomes, applied for this purpose in other cereal species (Fukui *et al*. 1992), has not become established in wheat, mainly due to laboriousness and low DNA yields obtained by this technique. Thus the best approach to dissecting the complex genome without any unintended losses of information became sorting of particular chromosomes by flow cytometry (Doležel *et al*., 1994, 2014), which allows reduction of wheat genome complexity to fractions of 224 – 580 Mb, representing 1.3 – 3.4 % of total wheat genome size (Šafář *et al*., 2010).

1.2.1.1 Flow-cytometric sorting of chromosomes

Flow cytometry was initially developed for human as a technique allowing counting blood cells and rapid measurement of individual metaphase chromosomes (Carrano *et al.*, 1979). In the course of time, it was modified for farm animals (Dixon *et al.*, 1992) and also for plant species (reviewed in Doležel, 1991). Depending

on sample type and quality, the samples are run at 1,000-2,000 particles per second and chromosomes of interest are sorted at rates of 10-40 chromosomes per second (Vrána *et al.*, 2012). Chromosomes held in liquid suspension are stained by a fluorochrome and passed into a flow chamber containing sheath fluid. The geometry of the chamber forces the chromosome suspension into a narrow stream, in which the chromosomes become aligned in a single file, and so are able to interact individually with an orthogonally directed laser beam(s). Pulses of scattered light and emitted fluorescence are detected and converted to electric pulse. If the chromosome of interest differs in fluorescence intensity from others, it is identified and sorted. The sorting is achieved by breaking the stream into droplets and by electrically charging droplets carrying chromosomes of interest. The droplets are deflected during passage through electrostatic field between deflection plates and collected in suitable containers (Figure 5) (Doležel *et al*, 2014).

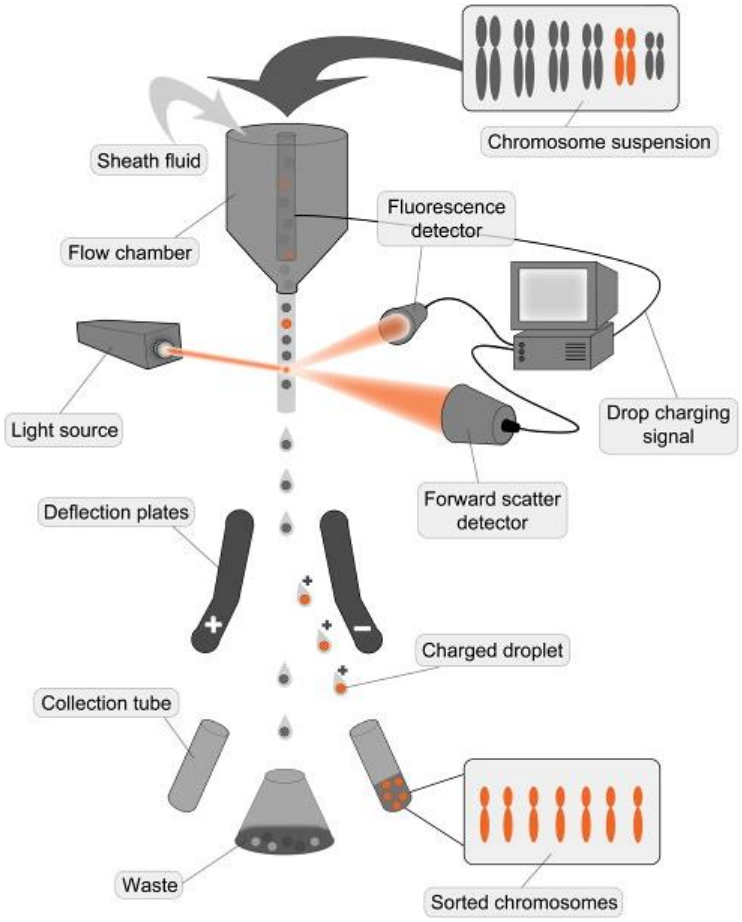


Figure 5: The mechanics of flow-sorting (Doležel *et al.*, 2014).

The outcome of the flow cytometric analysis is a flow karyotype, a histogram of relative fluorescence intensity (Figure 6). Chromosomes with the same DNA content are characterized by identical fluorescence intensity and form a separate peak in the histogram providing they differ sufficiently from other chromosomes in the genome.

Standard flow karyotype of wheat cv. Chinese Spring is composed of three peaks, each representing a mixture of chromosomes of similar size, and a minor peak composed largely of chromosome 3B, which is the only chromosome that can be discriminated and sorted from bread wheat with a standard karyotype (Vrána *et al.*, 2000, Figure 6A). This limitation can be overcome by using aneuploid ditelosomic lines that carry individual chromosome arms as telosomic chromosomes. These can be prepared from polyploid species because of their ability to tolerate aneuploidy. All telosomic chromosomes are significantly smaller in size than the standard chromosomes and thus are presented in a flow karyotype as a separated peak, which allows their discrimination and efficient sorting (Doležel *et al.*, 2014; Figure 6B). Utilization of ditelosomic lines makes it possible to sort all wheat chromosome arms except 5BL, which can be sorted from a line carrying this chromosome arm as an isochromosome (Vrána *et al*., 2012).

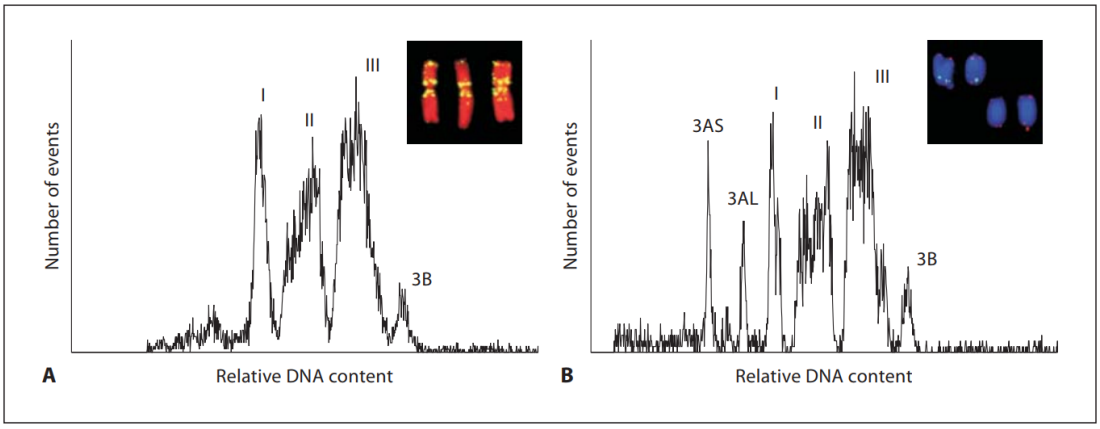


Figure 6: Chromosome analysis and sorting in wheat. A) Standard flow karyotype of hexaploid wheat (cv. Chinese Spring) showing three composite peaks (I-III) carrying groups of chromosomes and one peak containing only chromosome 3B obtained after flow cytometric analysis of DAPI-stained chromosome suspension. B**)** Flow karyotype of double ditelosomic line dDt3A of cv. Chinese Spring carrying two telocentric chromosomes 3AS and 3AL (according to Šafář *et al.*, 2010).

The set of wheat ditelosomic lines has been generated for individual chromosome arms of cv. Chinese Spring only (Sears *et* Sears, 1978), which for a long time restricted the use of the chromosome sorting to this single cultivar. Recently, this limitation was overcome by a cytogenetic technique termed FISHIS (fluorescence *in situ* hybridization in suspension) developed by Giorgi *et al*. (2013), in which mitotic chromosomes are differentially labeled by hybridizing them with fluorescently-labeled oligonucleotide probes targeting specific microsatellite sequences, typically GAA. This facilitated sorting of any chromosome from a cultivar of choice with a reasonable purity. The method was successfully used to sort chromosomes from several species of *Triticum, Aegilops* or *Agropyron* genera (Akpinar *et al*., 2015b; Molnár *et al*., 2016).

1.2.2 Sequencing strategies

There are two strategies applied in whole-genome sequencing projects, namely shotgun and clone-by-clone (CBC) sequencing (Figure 7). In the past, the CBC strategy was the major approach towards obtaining quality sequence of large genomes (IHGSC, 2001; Schnable *et al.*, 2009; Mascher *et al*., 2017), while shotgun sequencing represented an alternative approach, applied mainly for small genomes or those with limited funding. It used to provide more fragmented and less complete assemblies with limited ordering and anchoring of sequences. Both strategies have their pros and cons and can be applied individually or combined.

1.2.2.1 Whole-genome shotgun sequencing

Whole-genome shotgun (WGS) sequencing (Figure 7A) is in comparison with the CBC less laborious, cheaper and faster. If combined with the Illumina sequencing technology, it includes preparation of sequencing libraries with inserts of minimum two size categories. DNA fragments of 300-1,000 bp are sequenced from both ends producing "*pair-end*" reads that are 150-300 bp in size. Besides, it is possible to generate "*mate-pair*" reads, which are derived from fragments up to 12 kb in length. These long-distance pair-end reads are able to span problematic regions, such as longer stretches of DNA repeats, and thus are indispensable in assembling complex genomes. Alternatively, the long-distance information can be provided by long-read technologies, such as single-molecule real-time sequencing of Pacific Bioscences (PacBio) (Zimin *et al*, 2017b) or nanopore sequencing of Oxford Nanopore Technologies.

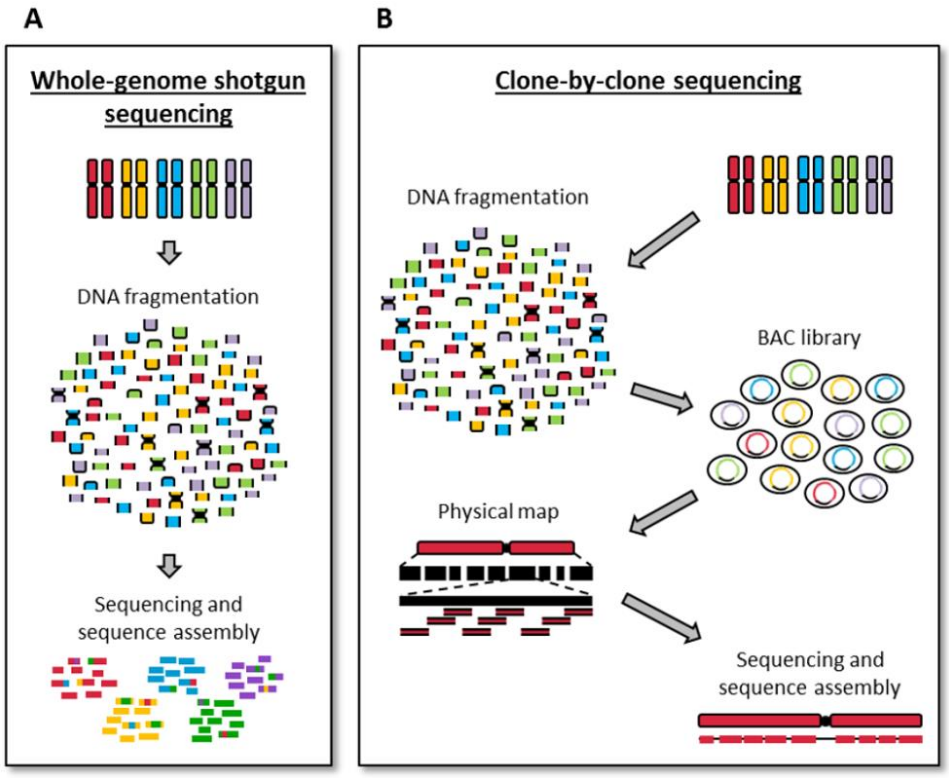


Figure 7: Two sequencing strategies. A) Whole-genome shotgun sequencing B**)** Clone-by-clone sequencing (Staňková, 2015)

1.2.2.1.1 Wheat assemblies obtained through the WGS

WGS is connected with the first efforts towards obtaining the wheat reference sequence. Brenchley *et al.* (2012) sequenced wheat cultivar Chinese Spring with 5x genome coverage using Roche/454 technology. The assembly comprised only 5.42 Gb of sequence, which is approximately one-third of the genome size, and resulted in prediction of ~95,000 genes. Later, Chapman *et al.* (2015) produced another wheat assembly generated by WGS. They sequenced wheat line "Synthetic W7984" with 30x coverage using Illumina platform. Also in this case, the resulting 9.1 Gb assembly did not cover the entire wheat genome.

In parallel with the above efforts, the IWGSC worked on generating another draft sequence of bread wheat genome on Illumina platform. The major difference between this and the previously described sequences of Brenchley *et al.* (2012) and Chapman *et al*. (2015) was the application of chromosomal approach, which employed flow cytometry to separate individual wheat chromosome or their arms (Doležel *et al*.,

2007). This step ensured significant reduction of complexity and avoided problems due to the presence of three homoelogous sub-genomes A, B and D (Choulet *et al*., 2014b). This draft chromosomal sequence assembly covered 10.2 Gb of the wheat genome and comprised 124,201 predicted genes, which were virtually ordered based on synteny with three sequenced wheat relatives (IWGSC, 2014). All three studies demonstrated that the WGS approach combined with short-read technologies enabled assembling low-copy regions of the wheat genome, but failed in capturing and correctly assembling the repetitive ones.

In 2017, Clavijo *et al*. (2017) reported a wheat genome assembly representing >78 % of the genome with scaffold N50 of 88.8 kb, which represented almost four times longer scaffolds in comparison with older wheat assemblies (Table 1). These results were achieved due to employing of mate-pair reads in scaffolding of the sequence. Besides, Clavijo *et al.* (2017) applied PacBio-sequenced full-length cDNA and Illumina RNAseq data for sequence annotation, which resulted in identification of 104,091 high-confidence genes.

Table 1: Comparison of published bread wheat assemblies

Assembly reference		Sequencing		Assembly		
		Platform	**Depth**	**Total lenght (% genome)**	**Scaffold N50**	**HC gene content**
1	Brenchley *et al.* (2012)	Roche/454	5x	5.4 Gb	< 1 kb	95,000
2	IWGSC (2014)	Illumina	30 - 241x*	10.2 Gb (48.9 %)	1.7 – 8.9 kb*	124,201
3	Chapman *et al*. (2015)	Illumina	30x	9.1 Gb (48.2 %)	24.8 kb	-
4	Clavijo *et al*. (2017)	Illumina	15 - 53x**	12.7 Gb (78 %)	88.8 kb	104,091
5	Zimin *et al*. (2017b)	Illumina PacBio	65x 36x	15.3 Gb (94 %)	232.6 kb†	-
6	IWGSC (2018)	Illumina	35 – 76x**	14.5 Gb (92 %)	7.1 Mb	107,891

*Parameters vary among chromosomes
**Parameters vary among sequencing libraries applied
† Contig-size N50

The advent of long-read sequencing technologies enabled using novel approaches towards obtaining the wheat whole-genome assembly. Zimin *et al.* (2017b) applied long-reads obtained by single-molecule real-time sequencing technology of Pacific Biosciences (PacBio) in combination with Illumina reads for sequence

assembling. The resulting assembly contained 15.3 Gb of sequence with contig-size N50 of 232.6 kb. A comparison of all hitherto generated bread wheat assemblies is shown in Table 1.

1.2.2.2 Clone-by-clone sequencing

The clone-by-clone sequencing strategy (Figure 7B) involves construction of a large-insert genomic library, cloned typically in bacterial artificial chromosome (BAC), and ordering of the BAC clones in a physical map. The next step is selection of a minimal set of overlapping BAC clones covering the whole genome, so called minimal tilling path (MTP). These are the template for sequencing. Positioning of the resulting sequences on chromosomes is done through the physical map contigs, which reduces demand on number of markers used for the anchoring.

A pilot project on the largest wheat chromosome 3B demonstrated that the CBC strategy was able to deliver a high quality reference sequence even for the complex wheat genome (Choulet *et al*., 2014b).

1.2.2.2.1 BAC libraries

Availability of a large-insert genomic library is a pre-requisite for construction of a physical map and the following CBC sequencing. First large-insert genomic libraries were cloned in yeast artificial chromosomes (YAC) with inserts up to 1,000 kb in length (Burke *et al*., 1987). YACs were used in physical mapping of human genome (Chumakov *et al*., 1992), but further studies revealed their insert instability and high level of chimerism. Thus the YAC libraries were replaced with libraries constructed in bacterial artificial chromosomes (BAC; Shizuya *et al*., 1992). The insert size of BAC clones is significantly smaller than that of YACs, typically 100 – 200 kb, but BAC libraries are easier to maintain and reproduce, show only a low level of chimerism, and are amenable to screening either by PCR or hybridization methods. Moreover, the inserts can be easily isolated by simple plasmid extraction procedures (Šafář *et al*, 2010).

BAC libraries have been constructed for a number of species from the tribe Triticeae, including *Aegilops tauschii* (Akhunov *et al*, 2005), *Triticum monococcum* (Lijavetzky *et al*., 1999), *Triticum boeticum* (Chen *et al*., 2002) or *Triticum urartu* (Akhunov *et al*., 2005). Undoubtedly, the biggest effort towards generating BAC

resources was done within the wheat genome sequencing project. In 2005, IWGSC adopted a strategy towards obtaining the reference wheat sequence that included construction of BAC libraries for each of wheat chromosomes/chromosome arms separated by flow cytometry. The first chromosome-based BAC library was constructed for the biggest wheat chromosome 3B (Šafář *et al*., 2004), followed by libraries specific for chromosomes 1D, 4D, 6D (Janda *et al*., 2004) and chromosome arm 1BS (Janda *et al*., 2006). Within twelve-year period, chromosome- or chromosome arm-specific BAC libraries were gradually constructed for all wheat chromosomes of bread wheat cv. Chinese Spring (https://olomouc.ueb.cas.cz/resources/dna-libraries; IWGSC, 2018).

1.2.2.2.2 Physical maps

Clone-based physical maps are represented by contiguous stretches of overlapping DNA fragments, inserts of the BAC clones. Their major role in sequencing projects declined with the advent of new sequencing technologies and assembling algorithms, but the physical maps and BAC clones they comprise can still serve as valuable genomic resources in gene cloning, enabling a fast access to and a detailed analysis of a region of interest (Tulpová *et al*., 2019b).

1.2.2.2.2.1 Physical map construction

The construction of a clone-based physical map requires identification of groups of clone inserts that overlap with one another. The most common method to identify the overlaps was fingerprinting, firstly utilized in case of physical map of the nematode *Caenorhabditis elegans* (Coulson *et al.*, 1986). Generally, fingerprinting is based on BAC clone digestion using restriction enzymes (RE), separation and subsequent detection of incurred fragments. In early protocols, clones were digested with a rarely cutting enzyme and radioactively labelled, then they were digested with another enzyme to produce smaller fragments suitable for separation and detection on a polyacrylamide gel. Overlapping clones from the same genome region were identified based on occurrence of several coincident fragments, and the number of shared fragments determined the extent of the overlap (Coulson *et al*., 1986). In the two-enzyme protocol, larger genomes with bigger proportion of repetitive sequences had relatively high risk of false positive overlaps (Meyers *et al.*, 2004). This stimulated development of more advanced fingerprinting techniques

involving digestion with three to five restriction enzymes (Xu *et al*., 2004). Fragments were end-labelled with fluorescently labelled ddNTPs and separated using capillary electrophoresis (Figure 8). One of the used five REs produced fragments with blunt ends, preventing their labelling and thus ensuring a reduction in number and size of detectable fragments, which made them accessible to analysis in a capillary sequencer. This approach has been called high-information-content fingerprinting (HICF) or SNaPshot technology (Luo *et al*., 2003) and it was used to construct a physical map of maize (Wei *et al*., 2007) barley (IBSC, 2012) and the D genome progenitor *Aegilops tauschii* (Luo *et al*., 2013). It was also applied in the wheat genome sequencing project to produce physical maps of individual chromosomes/chromosome arms, starting with 3B (Paux *et al*., 2008b), followed by 5A (Barabaschi *et al*., 2015) and chromosome arms 1AS, 1AL, 1BS, 1BL, 3DS, 5DS and 7DS (Breen *et al*., 2013; Lucas *et al*., 2013; Raats *et al*., 2013; Philippe *et al*., 2013; Holušová *et al*., 2017; Akpinar *et al*., 2015a; Tulpová *et al*., 2019a) (Table 2).

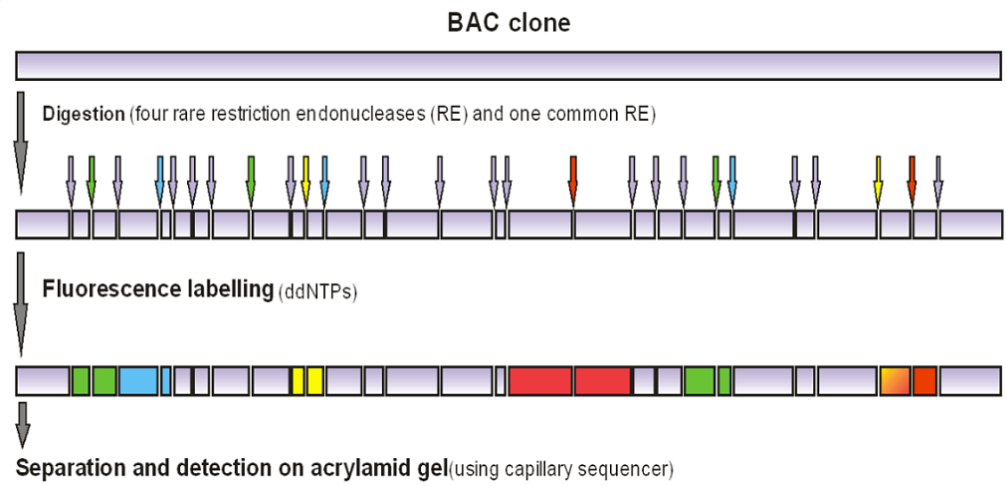


Figure 8: Scheme of the high-information-content fingerprinting. Only the coloured fragments are visualised and generate a four-colour restriction pattern

An alternative approach to obtaining fingerprints of BAC clones is a whole genome profiling (WGP) technique (van Oeveren *et al*., 2011). WGP employs next-generation sequencing (NGS) technologies to produce sequence tags along the BAC clone. BAC pooled samples are prepared from the library and cleaved by two restriction enzymes. A subset of resulting fragments are then ligated with adaptors and sequenced to produce a short sequence tag from each fragment. Finally, BAC contigs

are assembled based on overlapping sets of sequence tags. The WGP was used to produce physical maps of wheat chromosomes 6A (Poursarebani *et al.*, 2014), 6B (Kobayashi *et al.*, 2015), 2B, 2D, 4B, 5BL and 5DL (IWGSC, 2018). Besides, WGP tags were produced from MTP clones of all remaining wheat chromosomes and were utilized to support assembly of the bread wheat genome (IWGSC, 2018) (Table 2).

Table 2: Fingerprinting and assembling strategies applied in wheat chromosomal physical map projects.

Chromosome/arm	Fingerprinting method	Assembling algorithm	Reference
1AS	HICF	FPC, LTC	Breen *et al*. (2013)
1AL	HICF	FPC, LTC	Lucas *et al*. (2013)
1BS	HICF	LTC	Raats *et al*. (2013)
1BL	HICF	FPC, LTC	Philippe *et al*. (2013)
1D, 6D	HICF	LTC	IWGSC (2018)
2A	HICF	LTC	IWGSC (2018)
2B	WGP	LTC	IWGSC (2018)
2D	WGP	LTC	IWGSC (2018)
3A	HICF	FPC	IWGSC (2018)
3B	HICF	FPC	Paux *et al*. (2008b)
3DS	HICF	FPC, LTC	Holušová *et al*. (2017)
3DL	HICF	FPC	IWGSC (2018)
4A	HICF	LTC	IWGSC (2018)
4B	WGP	LTC	IWGSC (2018)
4D	HICF	LTC	IWGSC (2018)
5A	HICF	FPC, LTC	Barabaschi *et al*. (2015)
5BS	HICF	LTC	Salina *et al*. (2018)
5BL	HICF	LTC	IWGSC (2018)
5DS	HICF	LTC	Akpinar *et al*. (2015)
5DL	WGP	LTC	IWGSC (2018)
6A	WGP	FPC, LTC	Poursarebani *et al*. (2014)
6B	WGP	FPC	Kobayashi *et al*. (2015)
7A	HICF	FPC, LTC	Keeble-Gagnère *et al*. (2018)
7B	HICF	LTC	Belova *et al*. (2014)
7DS	HICF	FPC, LTC	Tulpová *et al*. (2019a)
7DL	HICF	FPC	IWGSC (2018)

To assemble a physical map from the fingerprint data, two types of statistical software can be applied: FingerPrinted Contigs (FPC; Soderlund *et al.*, 2000) and Linear Topology Contig (LTC; Frenkel *et al.*, 2010), which differ in several aspects. If utilizing FPC, the assembling process starts at very high stringency, characterized by extensive fingerprint overlaps, which is progressively decreased in further steps of the assembly. On the other hand, LTC, which utilizes graphical approach considering contig topology, starts assembling at a low stringency, resulting in branching contig structures, which is then progressively increased until a linear topology of the contigs is achieved (Figure 9). LTC has been considered more reliable in building contigs and selecting MTP (Frenkel *et al.*, 2010) and thus was used to validate wheat physical maps assembled primarily by FPC. Application of FPC and LTC for particular wheat chromosomal physical maps is shown in Table 2.

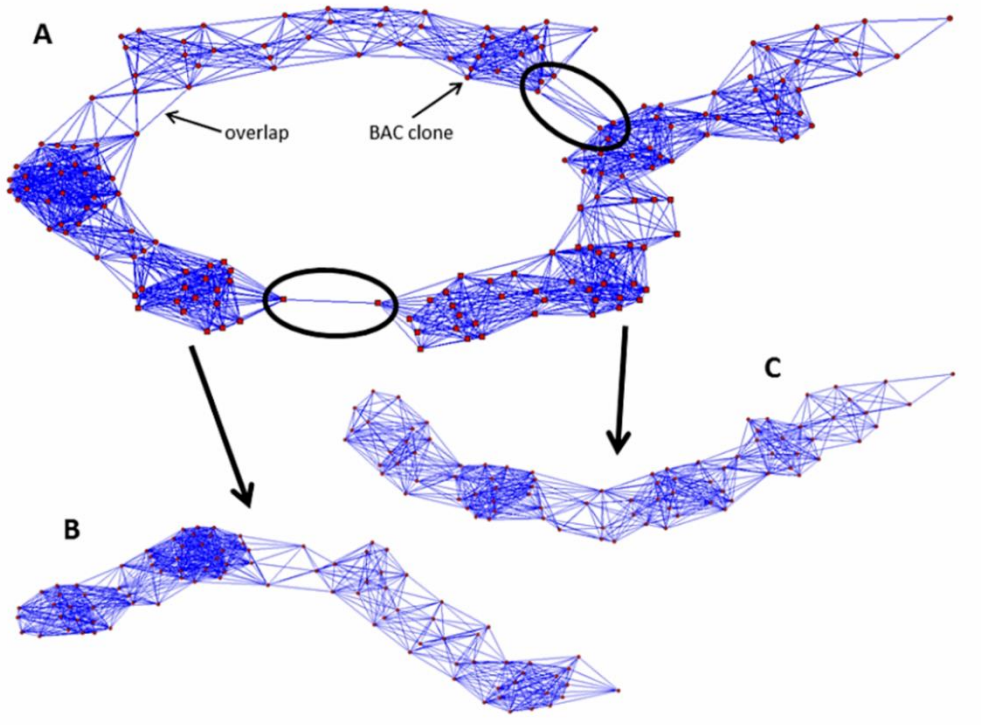


Figure 9: Identification and correction of a branching cluster visualized by LTC. Physical-map contigs are visualised as nets of significant overlaps. Red dots represent clones; blue lines represent significant clone overlaps. Questionable regions are marked by black ovals. Contig A exhibiting non-linear topological structure was split into contigs B and C by removing questionable clones causing branching (Staňková, 2015).

1.2.2.2.2.2 Physical map anchoring

To fully exploit a benefit of a physical map and to lay basis for whole-genome sequencing, the physical-map contigs need to be placed on a chromosome. The anchored physical map represents then a set of BAC contigs ordered and oriented along chromosomes. To place the contigs on a chromosome, two approaches can be used, designated as forward and reverse anchoring.

Forward anchoring utilizes markers with known genetic position, usually obtained from databases, which are subsequently localized in contigs of the physical map. The localization is done by screening of a corresponding BAC library by PCR or hybridization on high-density membranes bearing replicas of the library (Paux *et al*., 2008a; Gardiner *et al*., 2004). To accelerate the PCR screening and minimize its cost, multidimensional BAC pooling strategy was developed. Clones of BAC libraries are organized in 384-well plates. To prepare three-dimensional BAC pools, the BAC library is divided into several fractions, each comprising a specific number of plates, typically eight to ten. For each fraction, pools of BAC clones from three dimensions – plates, columns and rows – are prepared (Figure 10). A set of eight plates comprising 3,072 BAC clones can then be screened in 48 PCR reactions (Paux *et al*., 2008b). After detecting PCR products in particular pools, the BAC clone carrying the analysed marker is localized at the interception of all dimensions. Besides the simplest 3D pooling strategy, others have been proposed, including 6D, applied in sorghum (Klein *et al*., 2000) and maize (Yim *et al*., 2007), and 5D, which was considered as the most suitable for wheat and its diploid progenitors (Luo *et al*., 2009). To help with deconvolution of ambiguously identified BAC clones, perl software Elephant (*Electronic Physical Map Anchoring Tool*) was developed and used (Paux *et al*., 2008a; Holušová *et al*., 2017). The pooling strategy was also applied for high-throughput screening on array-based platforms developed originally for highly parallel SNP genotyping or expression analysis (e.g. Infinium SNP array, NimbleGen UniGene microarrays; Breen *et al*., 2013; Lucas *et al*., 2013; Luo *et al*., 2013; Philippe *et al*., 2013; Raats *et al*., 2013).

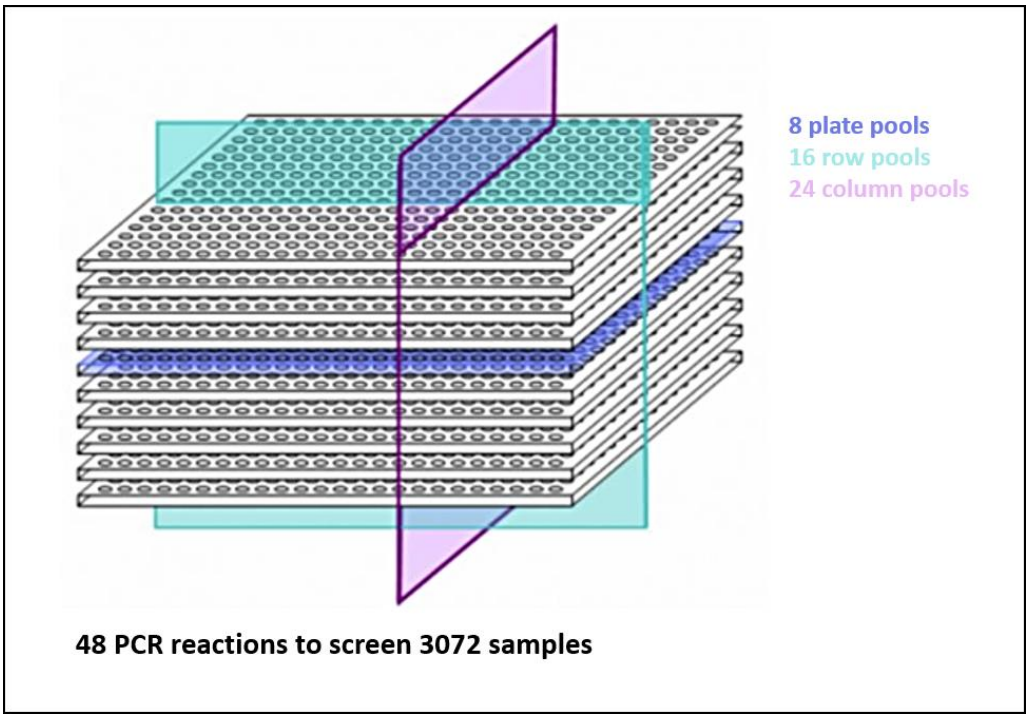


Figure 10: Three-dimensional pooling of BAC clones from eight microplates (https://cnrgv.toulouse.inra.fr/Services/Screening-services/DNA-Pool-production)

The alternative approach, reverse anchoring, utilized sequence information obtained from sequencing BAC clones or their ends (BAC-end sequences, BES). These sequences served as a resource for designing new markers that were then positioned on the chromosome by linkage mapping or using special cytogenetic stocks such as deletion lines (Raats *et al*., 2013; Philippe *et al*., 2013) or radiation hybrid (RH) panels. RH mapping, which is in wheat based on radiation-induced chromosomal deletions (Tiwari *et al*., 2016), overcomes problems due to low recombination rate in centromeric and pericentromeric regions, which is the biggest limitation of the genetic mapping (Feuillet *et al*., 2012; Balcárková *et al*., 2017; Tulpová *et al*., 2019a). The reverse anchoring was used for integration of genetic and physical maps in wheat (Paux *et al*., 2010; Tiwari *et al*., 2016) and *Ae. tauschii* (Wanjugi *et al*., 2009).

Another possibility to ordering contigs of a physical map, especially in the centromeric region, is employing information from chromosome-conformation-capture-based or optical mapping, as reported for barley whole-genome assembly by Mascher *et al*. (2017) or for wheat 7DS BAC assembly by Tulpová *et al.* (2019a). Chromosome-conformation-base mapping (Hi-C) involves chromatin crosslinking

with formaldehyde followed by digestion, biotinylation and ligation of biotinylated fragments. The resulting DNA sample thus containes ligation products consisting of fragments that were originally in close spatial proximity in the nucleus, marked with biotin at the junction. Finally, biotin-containing fragments are selected with streptavidin beads and sequenced. Hi-C allows unbiased identification of chromatin interactions across an entire genome and thus represented a rich source of long range information for assigning, ordering and orienting genomic sequences to chromosomes at megabase level (Lieberman-Aiden *et al.*, 2009; Mascher *et al*., 2017).

In a contrary Hi-C, a procedure of optical mapping does not employ massively parallel sequencing. Optical mapping is based on analysis of recognition site pattern along DNA molecules of tens to hundreds kilobases in length stretched in nano-channels (Lam *et al*., 2012). Thus the optical mapping can serve as a sapporting data set for scaffolding, scaffold ordering or orienting even regions challenging for common assemblers, such as centromere. This was demonstrated in Tulpová *et al.* (2019a) in case of (peri)centromeric region of wheat chromosome arm 7DS. Besides estimating contig order in the non-recombining region, the optical map helped revealing a structural variation close to the 7D centromere between bread wheat and its ancestor *Aegilops tauschii*.

1.2.3 IWGSC RefSeq v1.0

After 12 years of a joint effort towards getting a bread wheat reference genome, coordinated by the IWGSC, an annotated assembly of 21 chromosomes of cv. Chinese Spring has arisen. This assembly consists of 21 chromosome-based pseudomolecules, which comprise 96.8% of the assembly, and an additional dataset of unassigned scaffolds (3.2% assembly). It was built by integrating a draft *de novo* whole-genome assembly (WGA), generated from Illumina short-read data using NRGene deNovoMagic2 assembler (NRGene, Ness Ziona Israel; http://www.nrgene.com/), with additional information from multiple genomic resources (Figure 11). The assembly was superscaffolded and anchored to the chromosomes using genetic and physical mapping information, such as data from population sequencing (POPSEQ), Hi-C or chromosome-specific physical maps. Validation of the assembly was done using independent genetic and physical mapping evidence, including genotyping-by-sequencing maps, radiation hybrid (RH) maps or Bionano optical maps for group 7 homeologous chromosomes. Resulting highly contiguous IWGSC

RefSeq v1.0 assembly, covering 94 % of the wheat genome, contained 14.5 Gb sequence in scaffolds with N50 length of 7.1 Mb. Moreover, the assembly facilitated a straightforward access to 107,891 high-confidence genes, including their genomic context of regulatory sequences, and enabled a creation of a transcriptome atlas representing all stages of wheat development. Breeders can now easily get sequence-level information to precisely define the necessary changes in the genome for breeding programs (IWGSC, 2018). In addition, this assembly provides a deep insight into the distribution of genes and meiotic recombination along chromosomes, which laid the basis for their partitioning into five zones with distinct characteristics (IWGSC, 2018). Special attention was devoted to estimating precise position and size of wheat centromeres.

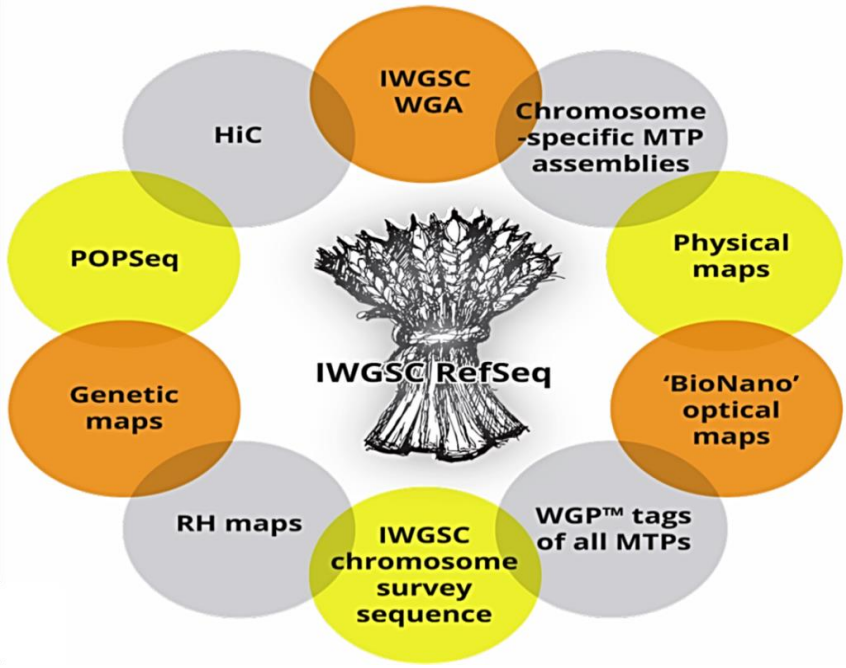


Figure 11: Integration of all resources applied in IWGSC RefSeq v1.0 building (http://www.wheatgenome.org/).

1.2.3.1 Delimiting of wheat centromeres

Centromeres constitute a crucial domain of eukaryotic chromosomes. They mediate chromosome attachment to microtubules and ensure proper segregation of the sister chromatids during mitosis and meiosis. Despite their critical importance, they are frequently incompletely covered or misassembled in reference genomes. This

is mainly due to high content of repeats and difficulty to order sequences in non-recombining centromeric regions. The genetic centromere, which is characterized by suppressed recombination, can span in wheat over hundreds megabases, covering as much as 50 % of chromosomal length (Choulet *et al*., 2014a; Keeble-Gagnère *et al*., 2018). However, the functional centromere is significantly smaller (Jiang *et al*., 2003).

On the sequence level, plant centromeres are frequently characterized by presence of GC rich satellites (Houben *et al*., 2007; Nagaki *et al*., 2004; Neumann *et al.*, 2012; Liu et al., 2015). Besides, they often contain a single class of transposable elements, in cereals mainly from *Ty3-gypsy* superfamily (Presting *et al*., 1998). These retroelements seem to be conserved in cereals since their radiation about 60 million years ago (Houben *et al*, 2007). Several of them have been characterized in grass species, namely Cereba in barley (*Hordeum vulgare*; Houben *et al*., 2007) or CRW in wild einkorn wheat (*Triticum boeticum*; Liu *et al*., 2008). According to Neumann *et al.* (2011), transcriptional activity of centromeric retrotransposons represents an active component of centromeres, plays an important role in plant centromere evolution by generating new insertions, and probably participates in normal centromere function. The core centromeric chromatin could also harbour transcriptionally active genes (Nagaki *et al.*, 2004). Despite some similarities of centromeric sequences across species, the determining feature does not seem the primary sequence but specific chromatin modifications, such as CENH3, a centromere-specific variant of histone H3, associated with centromeres of cereals (reviewed in Henikoff *et al*., 2001).

The position of the centromere is relatively easy to estimate using fluorescence *in situ* hybridization (FISH) with probes for centromere-specific repeats, in case of wheat and rye derived from a 192-bp segment of a repetitive element from wheat clone pHind258 (Ito *et al.*, 2004). Still FISH can indicate the position only roughly and does not provide information about the centromere physical size.

Another possibility to positioning the centromere is utilization of telosomic chromosomes (telosomes). These arise through misdivision of centromeres in normal chromosomes, which results in chromosome breakage. Such chromosome aberration is lethal in diploid organisms but can be tolerated in polyploids as demonstrated in several ditelosomic stocks of bread wheat (Sears *et* Sears, 1978). Sequence overlap

of two telosomes corresponding to arms of one chromosome can indicate position of the functional centromere.

Chromatin immunoprecipitation (ChIP) and derived chromatin immunoprecipitation sequencing (ChIP-seq) for CENH3 is currently the most effective and accurate method to identify the functional centromere, as demonstrated in rice (Nagaki *et al*., 2004) or maize (Wang *et al*., 2014). This approach was also applied for positioning and delimiting centromeres in the IWGSC RefSeq v1.0 assembly (IWGSC, 2018). Clear ChIP-seq peaks were evident in all wheat chromosomes and coincided with peaks of centromere-specific TEs Cereba and Quinta (Figure 12). The average centromere size estimated from the ChIP-seq data was 6.7 Mb. Comparison of this estimate with 0.75 Mb and 1.8 Mb obtained for rice (Nagaki *et al*., 2004) and maize (Wang *et al*., 2014), respectively, supported previous findings that the centromere size increases with the genome size.

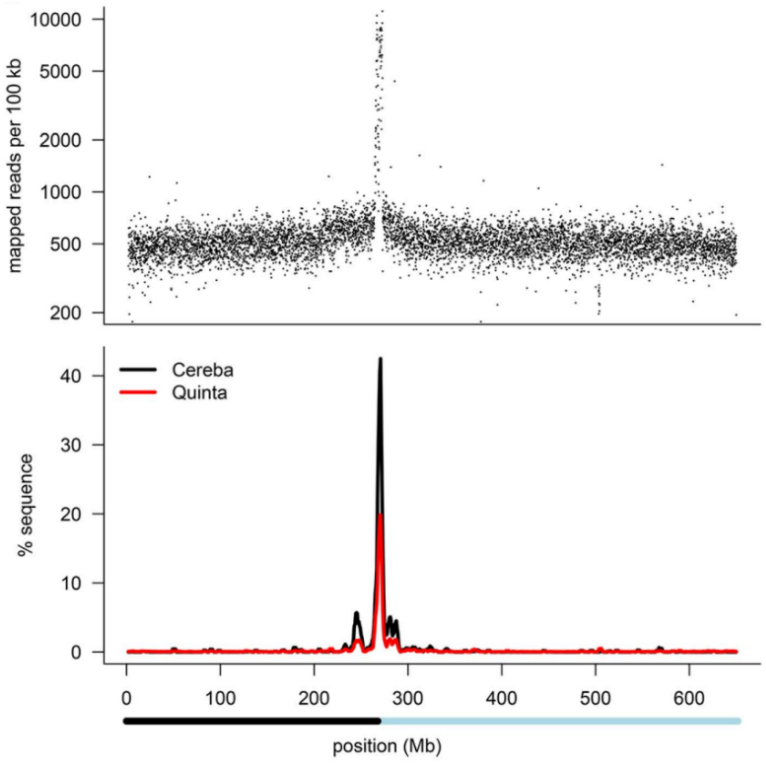


Figure 12: Positioning of the centromere in the 2D pseudomolecule. Top panel shows density of CENH3 ChIP-seq data along the wheat chromosome. Bottom panel shows distribution and proportion of the total pseudomolecule sequence composed of TEs of the Cereba and Quinta families (IWGSC, 2018).

1.3 Cloning a Russian wheat aphid resistance gene

Wheat chromosome arm 7DS, which represent the main subject of this thesis, carries several agronomically important genes, including those underlying resistance to Russian wheat aphid. Knowledge of an accurate sequence of the 7DS, together with other resources and technologies including optical mapping, facilitated and accelerated identification of the most probable candidate for a RWA resistance gene.

1.3.1 Russian wheat aphid

Russian wheat aphid (*Diuraphis noxia*, Kordjumov; RWA; Figure 15), is a small, yellow-green or gray-green elongated phloem-feeding insect of 1.4-2.3 mm in size (Botha, 2013). It belongs into a large group of about 4,500 species of hemimetabolous and hemipteran insects called *Aphididae*, the true aphids (Davis, 2012). *Aphididae*, and especially its subfamily *Aphidinae*, underwent dramatic radiation between ~5 – 26 million years ago, connected with a development of cyclical parthenogenesis and viviparity. The cyclical parthenogenesis includes seasonal alternating between one or several female parthenogenetic generations, in which unfertilized eggs develop into females, and a sexual generation, with a necessity of oocytes fertilization by male sperm. Typical life cycle begins in spring with females hatched from diapausing and frost-resistant eggs. These asexual females reproduce asexually during the summer producing a vast number of clonal offspring by viviparous parthenogenesis, resulting in large colonies that infest plants (Cuellar, 1977). Later in autumn, in a response to the shortening photoperiod, the clonal morphs produce sexual females (Davis, 2012).

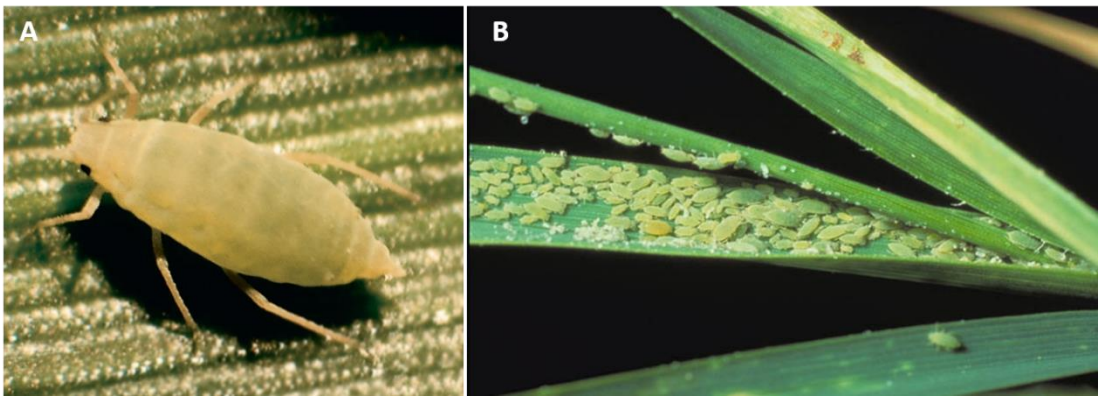


Figure 15: Russian wheat aphid (*Diuraphis noxia*). A picture of an adult female pest (A) and its colony on barley leaves (B) (https://www.invasive.org/browse/detail.cfm?imgnum=5512060)

It has been suggested that RWA had coevolved with crops in the Fertile Crescent (Botha, 2013), however its occurrence was first reported in 1978 in South Africa, followed by reports from all around the world, most recently from Australia (Yazdani *et al*., 2017). RWA devastates yields of small-grain cereals mainly in the USA where it caused economic losses of about $800 million between 1987 and 1993 (Morrison and Peairs, 1998). Moreover, process of its evolution includes adaptations to continual changes of life conditions leading to appearance of new highly virulent biotypes that are able to infest previously resistant cultivars (Botha, 2013). Thus RWA became one of the most serious invasive pests of wheat and barley but also other plants from 43 genera, including over 140 species of cultivated and wild grasses (Yazdani *et al*., 2017).

1.3.2 Types of arthropod resistance

As a defence against pathogens, an innate and adaptive immune system based on immunoglobulins evolved in human and animals. Despite of lacking specialized defence cells, plant defence strategies show similarities to those of animals, even though plants have to defend themselves by a combination of constitutive and induced defences (reviewed in Botha *et al*., 2005). Plant resistance to arthropods is defined as the sum of the constitutive, genetically inherited qualities that result in a plant of one cultivar or species being less damaged than a susceptible plant lacking these qualities. On the contrary, susceptibility is the inability of a plant to inherit qualities that express resistance to arthropods (Smith, 2005).

The pest resistance can be classified in three categories that are frequently combined – antibiosis, antixenosis and tolerance. The tolerance is an ability of plants to withstand or recover from damage caused by pest feeding. Antixenosis, or non-preference, is a type of resistance combining morphological and chemical plant factors that affect pest behaviour. These include natural physical plant barriers such as thickness of plant epidermal layer, waxy leaves, higher density of leave trichomes or production of several types of repellents and detergents that discourage the aphid from feeding and, consequently, the pest chooses an alternative host plant. Antibiosis is manifested as a negative effect of a plant on the developmental biology of the aphid, expressed mainly as lower aphid fecundity (Wang *et al*., 2004, reviewed in Botha *et al*., 2005). Mechanisms of plant defence through antixenosis and antibiosis can be overlapping.

1.3.2.1 Signalling pathways in aphid resistance

During its feeding, aphid is removing plant photoassimilates, which results in chlorosis, longitudinal streaking along the main leaf vein, head trapping, substantial reduction in biomass and, in severe cases, plant death (Burd and Burton, 1992). Resistant plants are able to reduce these devastating changes but the mechanism of their response to the aphid attack has not been clarified yet. Smith and Boyko (2007) came up with a hypothesis of two different processes involved in elicitation of plant response to aphid feeding. The first one is based on recognition of plant damage, leading to a change in plant chemistry, followed by the production of signalling molecules that activate a general stress response. The second process triggers the aphid-specific resistance by gene-for-gene recognition of elicitors that are most probably present in aphid saliva. This type of plant-pathogen interaction acts on a key-and-lock basis, meaning that the resistant host plant carries a resistance (*R*) gene specific to a pathogen gene termed the avirulence (*Avr*) gene. If the product of the *R* gene is not present or does not recognize the product of *Avr* gene, the plant is susceptible (Smith, 2005). Both processes elicit cascades of reactions mediated by several compounds, including jasmonic acid (JA), salicylic acid (SA), ethylene (ET), abscisic acid (ABA) or reactive oxygen species (ROS) (reviewed in Smith and Boyko, 2007; Figure 16).

Plants showing the antibiosis type of resistance generally use the induced gene-for-gene interaction. They employ hypersensitive response (HR) resulting in necrotic lesion on leaves (Botha, 2013) as well as other signalling cascades ending with JA, MeJA (methyl jasmonate), ET or SA (Smith and Boyko, 2007). On the contrary, aphid-tolerant plants, rather than the signalling pathway use a passive resistance, to deal with their damage on chlorophyll level and energy and nutrients removal. Thus their main aim is to compensate the reduced photosynthetic activity by upregulating components of the electron transport chain.

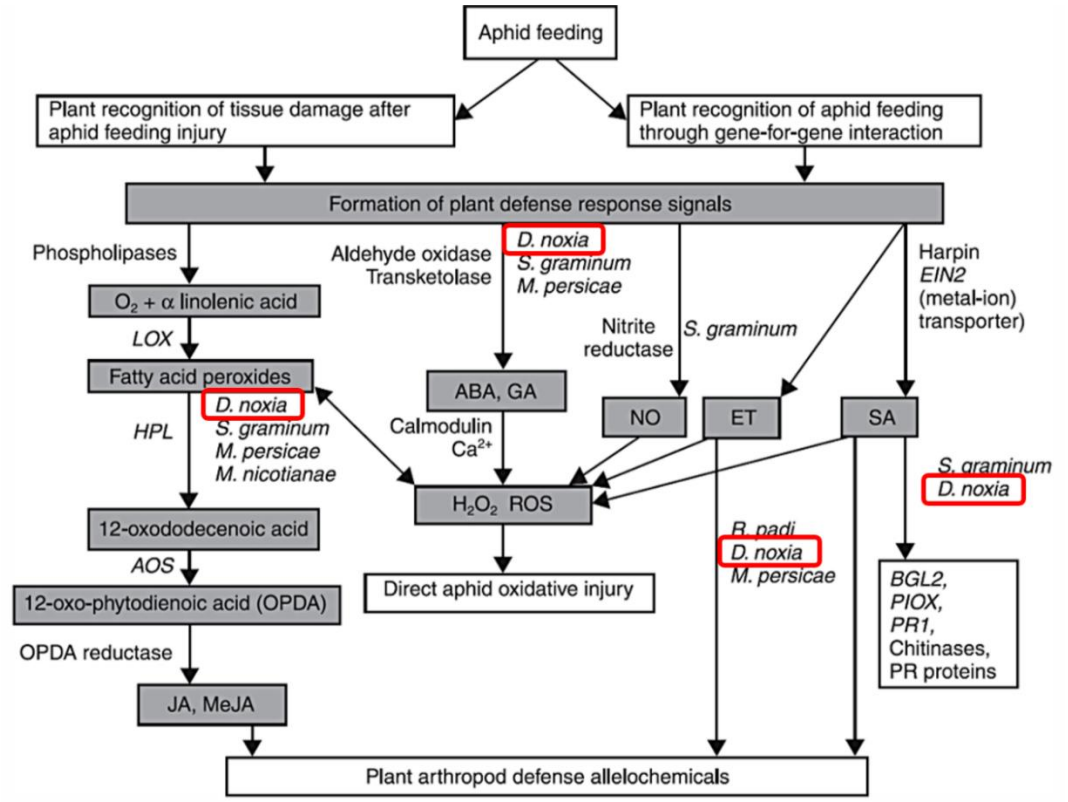


Figure 16: Scheme of signalling pathways involved in plant defence response elicited by aphid feeding (Smith and Boyko, 2007).

1.3.3 Cloning of RWA resistance genes

One of the typical symptoms of the aphid infestation is leaf rolling, which provides aphids an additional advantage, because it shelters their colonies against their natural predators as well as insecticide spraying, which reduces the effect of chemical pest treatment. Thus the most effective strategy to handle RWA attacks lies in using natural sources of aphid resistance in breeding of novel resistant cultivars. The knowledge of a gene underlying the resistance trait can speed up the breeding process by providing markers co-segregating with the trait. Moreover, it can clarify mechanisms of the plant defence, which can lead to development of new strategies for the pest treatment.

1.3.3.1 Positional cloning

Positional (or map-based) cloning represents a traditional approach towards isolation of a gene underlying a trait of interest without previous knowledge of the biochemical nature of the gene product. The process of positional cloning can be divided into four phases: 1) delimiting the gene region by genetic markers,

2) spanning the region by a contig of a physical map, 3) sequencing of co-segregating BAC clones, sequence annotation and identification of candidate gene(s), 4) functional validation of the candidate(s).

The positional cloning approach is primarily suitable for traits coded by major genes showing a clear phenotypic effect. Two individuals with a contrasting phenotype are crossed and used to derive a mapping population, usually F2 or more advanced recombinant inbred lines (RILs). For cloning quantitative trait loci (QTL), polygenes with a minor effect contributing to the trait, one of the loci needs to be selected and mendelized by preparing a back-crossed mapping population – nearly isogenic lines (NILs). Individuals of the mapping population are then assayed for available molecular markers and scored for the phenotype of interest to identify markers closely linked (less than 1 cM) with the trait and flanking the gene from both sides. For this high-resolution mapping, populations of several thousand individuals, typically RILs, may be needed (Krattinger *et al.*, 2009a).

The region delimited by flanking markers is to be spanned by BAC clones. Before availability of clone-based physical maps, this was achieved by chromosome walking (Krattinger *et al*., 2009a). Availability of the physical maps speeded up this phase significantly since screening of a BAC library with the closely linked markers identifies BAC contigs rather than individual clones. Spanning of the region with available reference sequence of the analysed species is even easier but one has to keep in mind that the region of interest may be not completely covered or may be misassembled in the reference genome (Tulpová *et al.* 2019b). This is especially relevant if the cultivar bearing the gene of interest differs significantly from the reference genome or the region bearing the gene of interest has been introgressed from a wheat relative or a landrace. If the region is completely absent from the available reference genome, a new genomic resource such as BAC library must be prepared from the gene donor (Mago *et al*., 2014; 2015). BAC clones within the delimited interval are sequenced and the obtained sequences are annotated to get a complete catalogue of genes present in the selected region.

The final step in a positional cloning project is identification and validation of candidate gene(s), which is usually performed by analysis of gene expression in the contrasting phenotypes and by methods of functional genomics, such as down-

regulation of the gene by virus induced gene silencing (VIGS; Gupta *et al*., 2013a) or the most popular gene editing using CRISPR-Cas9 system (Wang *et al*., 2015) or its modifications.

The majority of wheat genes hitherto isolated by the positional cloning approach are those underlying a resistance to different fungal pathogens, such as leaf rust – *Lr* genes (e.g. Feuillet *et al*., 2003; Krattinger *et al*., 2009b), stem rust – *Sr* genes (e.g. Mago *et al*., 2014; 2015), stripe rust – *Yr36* (Fu *et al*., 2009) or powdery mildew – *Pm3b* (Yahiaoui *et al.,* 2004). The positional cloning approach acted successfully also for a few wheat QTLs, including that for a transcription factor controlling senescence and grain nutrients (Uauy *et al*., 2006) and a grain-weight QTL that increases grain length through cell expansion in the pericarp (Brinton *et al*., 2018).

1.3.3.2 RWA resistance genes

To date, more than 15 RWA resistance genes, mostly underlying tolerance and antibiosis, have been described in wheat. Many of these originated from *Aegilops tauschii* as inferred from their map location in the D genome (Du Toit, 1987, 1989; Du Toit *et al.*, 1995; Fazel-Najafabadi *et al.*, 2015; Liu *et al.*, 2001; Ma *et al.*, 1998; Miller *et al.*, 2001; Peng *et al.*, 2007; Valdez *et al.*, 2012, Voothuluru *et al.*, 2006). At least one resistance gene, *Dn7*, originating from rye was reported (Anderson *et al.*, 2003; Marais *et al.*, 1994; Lapitan *et al.*, 2007). New highly virulent RWA biotypes have rendered previously resistant cultivars susceptible, which stirred intensive search for novel sources of resistance as well as efforts to unravel mechanisms underlying the trait. Although several components of the plant defence pathways have been proposed (Anderson *et al*., 2014, Van Eck *et al*., 2014), to date no aphid resistance gene has been cloned in wheat. Nine of the hitherto mapped RWA resistance genes, namely *Dn1*, *Dn2*, *Dn5* (Du Toit, 1987), *Dn6* (Saidi and Quick, 1996), *Dn8*, (Liu *et al.*, 2001), *Dnx* (Harvey and Martin, 1990), *Dn2401* (Voothuluru *et al.*, 2006; Fazel-Najafabadi *et al.* 2015), *Dn626580* (Valdez *et al.*, 2012), and *Dn100695* (Tonk *et al.*, 2016) are located on the short arm of wheat chromosome 7D (7DS). Except for *Dn8* located at the terminal part of the arm, the remaining genes were mapped to the interstitial part of 7DS and most of them were found linked to marker *gwm111* (Fazel-Najafabadi *et al.* 2015; Liu *et al.* 2002). It has not yet been resolved if these genes are allelic or tightly linked.

Within the framework of this thesis, we focused on the *Dn2401* gene identified in line CI2401, originating from Tajikistan. Previously, Staňková *et al.* (2015) mapped the *Dn2401* gene into 0.83-cM interval on chromosome arm 7DS and spanned it by five overlapping BAC clones from a 7DS-specific BAC library of cv. Chinese Spring. This cultivar, used as a reference genome of bread wheat, is susceptible to RWA (Peng *et al*., 2009), which raised the need to inspect the interval also in the resistant line. In the current study, we demonstrated that this goal can be facilitated by application of novel techniques, such as optical mapping.

1.4 References

AGI - The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**: 796-815.

Akhunov ED, Akhunova AR, Dvořák J (2005) BAC libraries of Triticum urartu, Aegilops speltoides and Ae.tauschii the diploid ancestors of polyploid wheat. *Theoretical ans applied genetics* **111**(8): 1617-1622.

Akpinar BA, Magni F, Yuce M, Stuart J, Lucas SJ, Šimková H, Šafář J, Vautrin S, Bergès H, Cattonaro F, Doležel J, Budak H (2015a) The physical map of wheat chromosome 5DS revealed gene duplications and small rearrangements. *BMC Genomics* **16**: 453.

Akpinar BA, Yuce M, Lucas S, Vrána J, Burešová V, Doležel J, Budak H (2015b) Molecular organization and comparative analysis of chromosome 5B of the wild wheat ancestor Triticum dicoccoides. *Scientific Reports* **5**:e10763

Anderson VA, Haley SD, Peairs FB,van Eck L., Leach JE, Lapitan NLV (2014) Virus-induced gene silencing suggests that *(1,3;1,4)-β-glucanase* is a susceptibility factor in the compatible Russian wheat aphid – wheat interaction. *Molecular Plant-Microbe Interaction* **27**: 913-922.

Avni R, Nave M, Barad O, Baruch K, Twardziok SO, Gundlach H, Hale I, Mascher M, Spannagl M, Wiebe K, Jordan KW, Golan G, Deek J, Ben-Zvi B, Ben-zvi G, Himmelbach A, MacLachlan RP, Sharpe AG, Fritz A, Ben-David R, Budak H, Fahima T, Korol A, Faris JD, Hernandez A, Mikel MA, Levy AA, Steffenson B, Maccaferri M, Tuberosa R, Cattivelli L, Faccioli P, Ceriotti A, Kashkush K, Pourkheirandish M, Komatsuda T, Eilam T, Sela H, Sharon A, Ohad N, Chamovitz DA, Mayer KFX, Stein N, Ronen G, Peleg Z, Pozniak CJ, Akhunov ED, Distelfeld A (2017) Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. *Science* **357**: 93-96.

Balcárková B, Frenkel Z, Škopová M, Abrouk M, Kumar A, Chao S, Kianian SF, Akhunov E, Korl AB, Doležel J, Valárik M (2017) A high resolution radiation hybrid map of wheat chromosome 4A. *Frontiers in Plant Science* 7: 2063.

Barabaschi D, Magni F, Volante A, Gadaleta A, Šimková H, Scalabrin S, Prazzoli MC, Bagnaresi P, Lacrima K, Michelotti V, Desiderio F, Orrù L, Mazzamurro V, Fricano A, Mastrengelox AM, Tononi P, Vitulo V, Jurman I, Frenkel Z, Cattonaro F, Morgante M, Blanco A, Doležel J, Delledonne M, Stanca AM, Cattivelli L, Valè G (2015) Physical mapping of bread wheat chromosome 5A: An integrated approach. *Plant Genome* **8**: 1-24.

Belova T, Grønvold L, Kumar A, Kianian S, He X, Lillemo M, Springer NM, Lien S, Olsen OA, Sandve SR (2014) Utilization of deletion bins to anchor and order sequences along the wheat 7B chromosome. *Theoretical and Applied Genetics* **127**: 2029-2040.

Botha AM, Li Y, Lapitan NLV (2005) Cereal host interactions with Russian wheat aphid: A review. *Journal of Plant Interactions* **1**(4): 211-222.

Botha AM (2013) A coevolutionary conundrum: the arns race between *Diuraphis noxia* (Kordjumov) a specialist pest and its host *Triticum aestivum* (L.). *Anthropod-Plant Interactions* **7**: 359-372.

Breen J, Wicker T, Shatalina M, Frenkel Z, Bertin I, Philippe R, Spielmeyer W, Šimková H, Šafář J, Cottonaro F, Scalabrin S, Magni F, Vautrin S, Bergès H, IWGSC, Paux E, Fahima T, Doležel J, Korol A, Feuillet C, Keller B (2013) A physical map of the short arm of wheat chromosome 1A. *PLoS ONE* **8**: e80272.

Brenchley R, Spannagl M, Pfeifer M, Barker GLA, D'Amore R, Allen AM, McKenzie N, Kramer M, Kerhornou A, Bolser D, Kay S, Waite D, Trick M, Bancroft I, Gu Y, Huo N, Luo MC, Sehgal S, Gill B, Kianian S, Anderson O, Kersey P, Dvorak J, McCombie WR, Hall A, Mayer KFX, Edwards KJ, Bevan MW, Hall N (2012) Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature* **491**: 705-710.

Briton J, Simmonds J, Uauy C (2018) Ubiquitin-related genes are differentially expressed in isogenic lines contrsting for pericarp cell size and grain weight in hexaploid wheat. *BMC Plant Biology* **18:**22.

Burd JD, Burton RL (1992) Characterization of plant damage caused by Russian wheat aphid (Homoptera: Aphididae). *Journal of Economic Entomology* **85**: 2017-2011.

Burke DT, Carle GF, Olson MV (1987) Clonig of large fragments of exogenous DNA into yeast by means of artificial chromosome vectors. *Science* **4803**: 806-812.

Carrano AV, Gray JW, Langlois RG, Burkhart-Schultz J, Van Dilla AM (1979) Measurement and purification of human chromosomes by flow cytometry and sorting. *Proceeding of the National Academy of Science of the United States of America* **76**(3):1382-1384.

Castro AM, Vasicek A, Ellerbrook C, Gimenéz DO, Tocho E, Tacaliti MS, Clúa A, Snape JW (2004) Mapping quantitative trait loci in wheat for resistance against greenbug and Russian wheat aphid. *Plant Breeding* **123**: 361-365.

Chao S, Zhang W, Akhunov E, Sherman J, Ma Y, Luo MC, Dubcovsky J (2009) Analysis of gene-derived SNP marker polymorphism in US wheat (*Triticum aestivum* L.) cultivars. *Molecular Breeding* **23**: 23-33.

Chapman JA, Mascher M, Buluç A, Barry K, Georganas E, Session A, Strnadova V, Jenkins J, Sehgal S, Oliker L, Schmutz J, Yelick KA, Scholtz U, Waugh R, Poland JA, Muehlbauer GJ, Stein N, Rokhsar DS (2015) A whole-genome shotgun approach for assembling and anchoring the hexaploid bread wheat genome. *Genome Biology* **16**:26.

Charles M, Belcram H, Just J, Huneau C, Viollet A, Couloux A, Segurens B, Carter M, Huteau V, Coriton O, Appels R, Samain S, Chalhoub B (2008) Dynamics and differential proliferation of transposable elements during the evolution of the B and A genomes of wheat. *Genetics* **180**(2): 1070-1086.

Choulet F, Alberti a, Theil S, Glover N, Barbe V, Daron J, Pingault L, Sourdille P, Couloux A, Paux E, Leroy P, Mangenot S, Guilhot N, Le Gouis J, Balfourier, Alaux M, Jamilloux, Poulain J, Durand C, Bellec A, Gaspin C, Safář J, Doležel J, Rogers J, Vandepoele K, Aury J-M, Mayer K, Berges H, Quesneville H, Wincker P, Feuillet C (2014a) Structural and functional partitioning of bread wheat chromosome 3B. *Science* **345** (6194): 1249721

Choulet F, Caccamo M, Wright J, Alaux M, Šimková H, Šafář J, Leroy P, Doležel J, Rogers J, Everesole K, Feuillet C (2014b) The Wheat Black Jack: Advances towards sequencing the 21 chromosomes of bred wheat. In: Tuberosa R, Graner A, Frison E (eds) Genomic of plant genetic resources. Volume 1. Managing, sequencing and minig genetic resources. pp 405-438 Springer Science + Bussiness Media, Dordrecht

Chumakov I, Rigault P, guillou S, Ougen P, Billaut A, guasconi G, Gervy P, LeGall I, Soularue P, Grinas L, Bougueleret L, Bellanné-Chantelot C, Lacroix B, Barillot E, Gesnouin P, Pook S, Vaysseix G, Frelat G, Schmitz A, Sambucy JL, Bosch A, Estivil X, Weissenbach J, Vignal A, Riethman, Cox D, Patterson D, Kathleen G, Hattoni Masahira, Sakaki Y, Ichikawa H, Ohki M, Le Paslier D, Heilig R, Antonarakis S, Cohen D (1992) Continuum of overlapping clones spanning the entire human chromosome 21q. Nature 359: 380-387.

Clavijo BJ, Venturini L, Schudoma C, Accinelli GG, Kaithakottil G, Wright J, Borrill P, Kettleborough G, Heavens D, Chapman H, Lipscombe J, Barker T, Lu FH, McKenzie N, Raats D, Ramirez-Gonzalez RH, Coince A, Peel N, Percival_Alwyn L, Duncan O, Trösch J, Yu G, Bolser DM, Namaati G, Kerhornou A, Spannagl M, Gundlach H, Haberer G, Davey RP, Fosker C, Di Palma F, Phillips AL, Millar AH, Kersey PJ, Uauy C, Krasileva KV, Swarbreck D, Bevan MW (2017) An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocation. *Genome Research* 27(5): 885-896.

Coombe L, Warren RL, Jackman SD, Yang C, Vandervalk BP, Moore RA, Pleasance S, Coope RJ, Bohlmann J, Holt RA, Jones SJM, Birol I (2016) Assembly of the complete sitka

spruce chloroplast genome using 10X Genomics GemCode sequencing data. *PLOS One*, doi: 10.1371/journal.pone.0163059

Coulson A, Sulston J, Brenner S, Karn J (1986) Towards a physical map of the genome of the nematode Caenorhabditis elegans. *Proceeding of the National Academy of Sciences of the United states of America* **83**: 7821-7825.

Cuellar O (1977) Animal parthenogenesis. *Science* **197**: 837-843.

Davis GK (2012) Cyclical parthenogenesis and viviparity in Aphids as evolutionary novelties. *Journal of Experimental Zoology* **318B**: 448-459.

Denton RD, Kudra RS, Malcom JW, Du Preez L, Malone JH (2018) the African Bullfrog (Pyxicephalus adspersus) genome unites the two ancestral ingredients for making vertebrate sex chromosomes. *bioRxiv*, doi: 10.1101/329847

Devos KM, Ma J, Pontaroli AC, Pratt LH, Bennetzen JL (2005) Analysis and mapping of randomly chosen bacterial artificial chromosome clones from hexaploid bread wheat. *PNAS* **102**(52): 19243-19248.

Dixon SC, Miller NGA, Carter NP, Tucker E (1992) Bivariate flow cytometry of farm animal chromosomes: a potetntial tool for gene mapping. *Animal Genetics* **23**: 203-210.

Doležel J (1991) Flow cytometric analysis of nuclear DNA content in higher plants. *Phytochemical Analysis*, doi: 10.1002/pca.2800020402.

Doležel J, Lucretti S, Schubert I (1994) Plant chromosome analysis and sorting by flow cytometry. *Critical Reviews in Plant Sciences* **13**(3): 275-309.

Doležel J, Kubaláková M, Paux E, Bartoš J, Feuillet C (2007) Chromosome-based genomics in cereals. *Chromosome Research* **15**: 51-66.

Doležel J, Vrána J, Cápal P, Kubaláková M, Burešová V, Šimková H (2014) Advances in plant chromosome genomics.. *Biotechnology Advances* **32**: 122-136.

Doležel J, Čížková J, Šimková H, Bartoš J (2018) One major challenge of sequencing large plant genomes is to know how big they really are. *International Journal of Molecular Science* **19**(11): 3554.

Dubcovsky J, Dvořák J (2007) Genome plasticity a key factor in the success of polyploid wheat under domestication. *Science* **316**: 1862-1866.

Du Toit F (1987) Resistance in wheat (*Triticum aestivum*) to *Diuraphis no*xia (Homoptera: *Aphididae*). *Cereal Research Communication* **15:** 175-179.

Du Toit F, Wessel WG, Marais GF (1995) The chromosome arm location of the Russian wheat aphid resistance gene, Dn5. *Cereal Research Communication* **23**: 15-17.

Dvořák J (2009) Crops and Models. In: Muehlbauer GM, Feuillet C (eds) Plant Genetics and Genomics, pp. 685-711, Springer, New York.

Eisenstein M (2015) Startups use short-read data to expand long-read sequencing market. *Nature Biotechnology* **33**(5): 433-435.

Eversole K, Feuillet C, Mayer KFX, Rogers J (2014) Slicing the wheat genome. *Science* **345**: 1257983.

Fazel-Najafabadi M, Peng J, Peairs FB, Šimková H, Kilian A, Lapitan NLV (2015) Genetic mapping of resistance to *Diuraphis noxia* (Kordjumov) biotype 2 in wheat (*Triticum aestivum* L) accession CI2401. *Euphytica* **203**: 607-614.

Feldman M, Lupton FGH, Miller TE (1995) Wheat. In: Smart J, Simmonds NW (eds) Evolution of crop plants. 2nd ed. pp 184-192 Longman Scientific and Technical, New York.

Feldman M, Levy AA (2005) Allopolyploidy – a shaping force in the evolution of wheat genomes. *Cytogenetic ans Genome Research* **109**: 250-258.

Feldman M, Levy AA (2009) Genome evolution in allopolyploid wheat – a revolutionary reprogramming followed by gradual changes. Journal of Genetics and *Genomics* **36**(9): 511-518.

Feuillet C, Travella N, Stein N, Albar L, Nublat A, Keller B (2003) Map-based isolation of the leaf rust disease resistance gene Lr10 from the hexaploid wheat (triticum aestivum L.) genome. *Proceeding of the National Academy of Science of the United States of America* **100**(25): 15253-15258.

Feuillet C, Stein N, Rossini L, Paud S, Mayer K, Schulman A, Eversole K, Appels R (2012) Integrating cereal genomics to support innovation in the Triticeae. *Functional and Itegrative Genomics* **12**(4): 573-583.

Food and Agriculture Organization of the United Nations, FAOSTAT Statistics Database, 78 (2018), http://www.fao.org/faostat/en/#data/FBS; http://www.fao.org/worldfoodsituation/csdb/en/

Frenkel Z, Paux E, Mester D, Feuillet C, Korol A (2010) LTC: a novel algorithm o improve the efficiency of contig assembly for physical mapping in complex genomes. *BMC Bioinformatics* **11**: 584.

Fu D, Uauy C, Distelfeld A, Blechl A, Epstein L, Chen X, Sela H, Fahima T, Dubcovsky J (2009) A kinase-START gene confers temperature-dependent resistance to wheat stripe rust. *Science* **323**:1357-1360.

Gardiner J, Schroeder S, Polacco ML, Sanchez-Villeda H, Fang Z, Morgante M, Landewe T, Fengler K, Useche F, Hanafey M, Tingey S, Chou H, Wing R, Soderlund C, Coe EH Jr (2004) Anchoring 9,371 maize expressed sequence tagged unigenes to the bacterial artificial chromosomes contig map by two-dimensional overgo hybridization. *Plant Physiology* **134**: 1317-1326.

Gupta B, Saha J, Sengupta A, Gupta K (2013a) Recent advances on Virus induced gene silencing (VIGS): Plant functional genomics. *Journal of Plant Biochemistery and Physiology* **1**: e116.

Guo X, Su H, Shi Q, Fu S, Wang J, Zhang X, Hu Z, Han F (2016) De novo centromere formation and centromeric sequence expansion in wheat and its wide relatives. *PLOS Genetics* **12**(4): e1005997

Harvey TL, Martin TJ (1990) Resistance to Russian wheat aphid, *Diuraphis noxia*, in wheat (*Triticum aestivum*). *Cereal Research Communication* **18**: 127-129.

Henikoff S, Ahmad K, Malik H (2001) The centromere paradox: stable inheritance with rapidly evolving DNA. *Science* **293**: 1098-1102.

Holušová K, Vrána J, Šafář J, Šimková H, Balcárková B, Frenkel Z, Darrier B, Paux E, Cattonaro F, Bergès H, Letellier T, Alaux M, Doležel J, Bartoš J (2017) Physical map of the short arm of bread wheat chromosome 3DS. *Plant Genome* **10**, doi: 10.3835/plantgenome2017.03.0021

Houben A, Schroeder-Reiter E, Nagaki K, Nasuda S, Wanner G, Murata M, Endo TR (2007) CENH3 interacts with the centromeric retrotransposon cereba and GC-rich satellites and locates to centromeric substructures in barley. *Chromosoma* **116**(3): 275-283.

IBSC – The International Barley Genome Sequencing Consortium (2012) A physical, genetic and functional sequence assembly of the barley genome. *Nature* **491**: 711 – 716.

IHGSC – The International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* **409**: 860-921.

Ito H, Nasuda S, Endo TR (2004) A direct repeat sequence associated with the centromeric retrotransposon in wheat. *Genome* **47**: 747-756.

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

IWGSC – The International Wheat Genome Sequencing Consortium (2018) Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* **361:** eaar7191.

Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT, Malla S, Marriott H, Nieto T, O'Grady J, Olsen HE, Pedersen BS, Rhie A, Richardson H, Quinlan AR, Snutch TP, Tee L, Paten B, Phillippy AM, Simpson JT, Loman NJ, Loose M (2018) Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology* **36**: 338-345.

Janda J, Bartoš J, Šafář J, Kubaláková M, Valárik M, Čihalíková J, Šimková H, Caboche M, Sourdille P, Bernard M, Chalhoub B, Doležel J (2004) Costruction of a subgenomic BAC library specific for chromosome 1D, 4D and 6D of hexaploid wheat. *Theoretical and Applied Genetics* **109**: 1337-1345.

Janda J, Šafář J, Kubaláková M, Bartoš J, Kovářová P, Suchánková P, Pateyron S, Čihalíková J, Sourdille P, Šimková H, Faivre-Rampant p, Hřibová E, Bernard M, Lukaszewski A, Doležel J, Chalhoub B (2006) Novel resources for wheat genomics: BAC library specific for the short arm of chromosome 1B. *Plant Journal* **47**: 977-986.

Jiang J, Birchler JA, Parrott WA, Dawe RK (2003) A molecular view of plant centromers. *Trends in Plant Science* **8**(12): 570-575.

Keeble-Gagnère G, Rigault P, Tibbits J, Pasam R, Hayden M, Forrest K, Frenkel Z, Korol A, Huang E, Cavanagh C, Taylor J, Abrouk M, Sharpe A, Konkin D, Sourdille P, Darrier B, Choulet F, Bernard A, Rochfort S, Dimech AM, Watson-Haigh N, Baumann U, Eckermann P, Fleury D, Juhasz A, Boisvert S, Nolin MA, Doležel J, Šimková H, Toegelová H, Šafář J, Luo MC, Camara F, Pfeifer M, Isdale D, Nystrom-Persson, IWGSC, Koo DH, Tinning M, Cui D, Ru Z, Appels (2018) Optical and physical mapping with local finishing enables megabases-scale resolution of agronomically important regions in the wheat genome. *bioRxiv*, doi: 10.1101/363465

Klein PE, Klein RR, Cartinhour SW, Ulanch PE, Dong J, Obert JA, Morishige, DT, Schlueter SD, Childs KL, Ale M, Mullet JE (2000) A high-throughtput AFLP-based method for constructing integrated genetic and physical maps: progress toward a sorghum genome map. *Genome Research* **10**: 789-807.

Kobayashi F, Wu J, Kanamori H, Tanaka T, Katagiri S, Karasawa W, Kaneko S, Watanabe S, Sakaguchi T, Hanawa Y, Fujisawa H, Kurita K, Abe C, Iehisa JC, Ohno R, Šafář J, Šimková H, Mukai Y, Hamada M, Saito M, Ishikawa G, Katayose Y, Endo TR, akumi S, Nakamura T, Sato K, Ogihara Y, Hayakawa K, Doležel J, Nasuda S, Matsumoto T, Handa H (2015) A high-resoution physical map integrating an anchored chromosome with the BAC physical map of wheat chromosome 6B. *BMC Genomics* **16**:595.

Krattinger SG, Wicker T, Keller B (2009a) Map-based cloning of genes in Triticeae (wheat and barley). In: Feuillet C, Muehlbauer GJ (eds) Genetics and genomics of the Triticeae. Pp 337-358 Springer NewYork.

Krattinger SG, Lagudah ES, Spielmeyer W, Singh RP, Huerta-Espino J, McFadden H, Bossolini E, Selter LL, Keller B (2009b) A putative ABC transporter confers durable resistance to multiple fungal pathogens in wheat. *Science* **323**:1360-1363.

Kronenberg ZN, Fiddes IT, Gordon D, Murali S, Cantsilieris S, Meyerson OS, Underwood JG, Nelson BJ, Chaisson MJP, Dougherty ML, Munson KM, Hastie AR, Diekhans M, Hormozdiari F, Lorusso N, Hoekzema K, Qiu R, Clark K, Raja A, Welch AE, Sorensen M, Baker C, Fulton RS, Armstrong J, Graves-Lindsay TA, Denli AM, Hoppe ER, Hsish P, Hill CM, Pang AWC, Lee J, Lam ET, Dutcher SK, Gage FH, Warren WC, Shendure J, Haussler D, Schneider VA, Cao H, Ventrua M, Wilson RK, Paten B, Pollen A, Eichler EE (2018) High-resolution comparative analysis of great ape genomes. *Science* **360**, doi: 10.1126/science.aar6343

Li LF, Liu B, Olsen KM, Wendel JF (2015) A re-evaluation of the homoploid hybrid origin of *Aegilops tauschii*, the donor of the wheat D-subgenome. *New Phytologist* **208**: 4-8.

Lieberman-Aiden E, van Berkum NL, williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J (2009) Comprehensive mapping of long-range interactions reveales folding principles of human genome. *Science* **326**:289.

Lijavetzky D, Muzzi G, Wicker T, Keller B, Wing R, Dubcovsky J (1999) Construction and chracterization of a bacterial artificial chromosome (BAC) library for the A genome of wheat. *Genome* **42**(6): 1176-1182.

Liu H, Liu Q, Chen Z, Zhou C, Liang Q, Ma C, Zhou J, Pan Y, Chen M, Wangjiu, Jiang W, Xiao S, Mou Z (2018a) Draft genome of Glyptosternon maculatum, an endemic fish from Tibet-plateau. *GigaScience*, doi: 10.1093/gigascience/giy104

Liu Q, Chang SY, Hartman GL, Domier LL (2018b) Assembly and annotation of a draft genome sequence for Glycine latifolia, a perennial wild relative of soybean. *Plant Journal* **95**(1): 71-85.

Liu XM, Smith CM, Gill BS, Tolmay V (2001) Microsatellite markers linked to six Russian wheat aphid resistance genes in wheat. *Theoretical and Applied Genetics* **102**: 500-511.

Liu XM, Smith CM, Gill BS (2002) Identification of microsatellite markers linked to Russian wheat aphid resistance genes *Dn4* and *Dn6*. *Theoretical and Applied Genetics* **104**: 1042-1048.

Liu Yalin, Su H, Zhang J, Liu Yang, Han F, Birchler JA (2015) Dynamic epigenetic states of maize centromeres. *Frontiers in Plant Science* **6**:904.

Liu Z, Yue W, Li D, Wang R, Kong X, Lu K, Wang G, Dong Y, Jin W, Zhang X (2008) Structure and dynamics of retrotransposons at wheat centromeres and pericentromeres. *Chromosoma* **117**: 445-456.

Lu F, Romay MC, Glaubitz JC, Bradbury PJ, Elshire RJ, Wang T, Li Y, Li Y, Semagn K, Zhang X, Hernandez AG, Mikel MA, Soifer I, Barad O, Buckler ES (2015) High-resolution genetic mapping of maize pen-genome sequence anchors. *Nature Communications* **6**: 6914.

Lucas SJ, Akpinar BA, Kantar M, Weinstein Z, Aysmoğlu F, Šafář J, Šimková H, Frenkel Z, korol A, Magni F, Cattonaro F, Vautrin S, Bellec A, Bergès H, Doležel J, Budak H (2013) Physical mapping integrated with syntenic analysis to characterize the gene space of the long arm of wheat chromosome 1A. *PLoS ONE* **8**: e59542.

Luo MC, Thomas C, You FM, Hsiao J, Ouyang S, Buell CR, Malandro M, Mcguire PE, Anderson OD, Dvorak J (2003) High-throughput fingerprinting of bacterial artificial chromosomes using the SNaPshot labeling kit and sizing of restriction fragments by capillary electrophoresis. *Genomics* **82**:378-389.

Luo MC, Xu K, Ma Y, Deal KR, Nicolet CM, Dvorak J (2009) A high-throughput strategy for screening of bacterial artificial chromosome libraries and anchoring of clones on genetic map constructed with single nucleotide polymorphisms. *BMC Genomics* **10**: 28.

Luo MC, Gu YQ, You FM, Deal KR, Ma Y, Hu Y, Huo N, Wang Y, Wang J, Chen S, Jorgensen CM, Zhang Y, McGuire PE, Paternak S, Stein JC, Ware D, Kramer M, McCombie WR, Kianian SF, Martis MM, Mayer KF, Sehgal SK, Li W, Gill BS, Bevan MW, Šimková H, Doležel J, Weining S, Lazo GR, Anderson OD, Dvorak J (2013) A 4-gigabase physical map unlocks the structure and evolution of the complex genome of *Aegilops tauschii*, the wheat D-genome progenitor. *PNAS* **110**: 7940-7945.

Mago R, Tabe L, Vautrin S, Šimková H, Kubaláková M, Upadhyaya N, Bergès H, Kong X, Breen J, Doležel J, Appels R, Ellis JG, Spielmeyer W (2014) Major haplotype divergence including multiple germin-like protein genes, at the wheat *Sr2* adult plant stem rust resistance locus. *BMC Plant Biology* **14**: 379.

Mago R, Zhang P, Vautrin S, Šimková H, Bansal U, Luo MC, Rouse M, Karaoglu H, Periyannan S, Kolmer J, Jin Y, Ayliffe MA, Bariana H, Park RF, McIntosh R, Doležel J, Bergès H, Spielmeyer W, Lagudah ES, Ellis JG, Dodds PN (2015) The wheat *Sr50* gene reveals rich diversity at a cereal disease resistance locus. *Nature Plants* **1**: 15186.

Marais GF, Horn M, Du Toit F (1994) Intergenic transfer (rye to wheat) of a gene(s) for Russian wheat aphid resistance. *Plant Breeding* **113**: 265-271.

Marcussen T, Sandve SR, Heier L, Spannagl M, Pfeifer M, IWGSC, Jakobsen KS, Wulff BBH, Steuernagel B, Mayer KFX, Olsen OA (2014) Ancient hybridizations among the ancestral genomes of bred wheat. *Science* **345**:1250092.

Martin TJ, Sears RG, Seifers DL, Harvey TL, Witt MD, Schlegel AJ, McCluskey PJ, Hatchett AJ (2001) Registration of 'Trego' wheat. *Crop Science* **41**: 929-930.

Martinez-Perez E, Shaw P, Moore G (2001) The *Ph1* locus is needed to ensure specific somatic and meiotic centromere association. *Nature* **411**: 204-207.

Mascher M, Gundlach H, Himmelbach A, Beier S, Twardziok SO, Wicker T, Radchuk V, Dockter C, Hedley PE, Russell J, Bayer M, Ramsay L, Liu H, Haberer G, Zhang XQ, Zhang Q, Barrero RA, Li L, Taudien S, Groth M, Felder M, Hastie A, Šimková H, Staňková H, Vrána J, Chan S, Muñoz-Amatriaín M, Ounit R, Wanamaker S, Bolser D, Colmsee C, Schmutzer T, Aliyeva-Schnorr L, Grasso S, Tanskanen J, Chailyan A, Sampath D, Heavens D, Clissold L, Cao S, Chapman B, Dai F, Han Y, Li H, Li X, Lin C, McCooke JK, Tan C, Wang P, Wang S, Yin S, Zhou G, Poland JA, Bellgard MI, Borisjuk L, Houben A, Doležel J, Ayling S, Lonardi S, Kersey P, Langridge P, Muehlbauer GJ, Clark MD, Caccamo M, Schulman AH, Mayer KFX, Platzer M, Close TJ, Scholz U, Hansson M, Zhang G, Braumann I, Spannagl M, Li C, Waugh R, Stein N (2017) A chromosome conformation capture ordered sequence of the barley genome. *Nature* **544**: 427-433.

Meyers BC, Scalabrin S, Morgante M (2004) Mapping and sequencing complex genomes: Let's go physical! *Genetics* **5**: 578-589.

Moll KM, Zhou P, Ramaraj T, Fajardo D, Devitt NP, Sadowsky MJ, Stupar RM, Tiffin P, Miller JR, Young ND, Silverstein KAT, Mudge J (2017) Strategies for optimizing BioNano

and Dovetail explored through a second reference quality assembly for the legume model, Medicago truncatula. *BMC Genomics* **18**: 578.

Molnár I, Vrána J, Burešová V, Cápal P, Farkas A, Darkó É, Cseh A, Kubaláková M, Molnár-Láng M, Doležel J (2016) Dissecting the U, M, S and C genomes of wild relatives of bread wheat (*Aegilops* spp.) into chromosomes and exploring their synteny with wheat. *Plant Journal* **88**(3): 452-467.

Morrison WP, Peairs FB (1998) Response model concept and economic impact. In: Quisenberry SS, Peairs FB (eds.) A response model for an introduced pest – the Russian wheat aphid. Thomas Say Publications in Entomology. Entomological Society of America, Lanham, MD: 1-11.

Nagaki K, Cheng TZ, Ouyang S, Talbert PB, Kim M, Jones KM, Henikoff S, Buell CR, Jiang J (2004) Sequencing of rice centromere uncovers active genes. *Nature Genetics* **36**: 138-145.

Navabi Z, Shiran B, Assad MT (2004) Microsatellite mapping of a Russian wheat aphid resistance gene on chromosome 7B of an Iranian tetraploid wheat line: Preliminary results. *Cereal Research Communication* **32**: 451-457.

Neumann P, Navrátilová A, Koblížková A, Kejnovský E, Hřibová E, Hobza R, Widmer A, Doležel J, Macas J (2011) Plant centromeric retrotransposons: a structural and cytogenetic perspective. *Mobile DNA* **2**: 4.

Neumann P, Navrátilová A, Schroeder-Reiter A, Koblížková A, Steinbauerová V, Chocholová E, Novák P, Wanner G, Macas J (2012) Stretching the rules: Monocentric chromosomes with multiple centromeres domains. *PLOS Genetics* **8**(6): e1002777.

Nkongolo KK, Quick JS, Limi E, Fowler DB (1991) Sources and inheritance of resistance to Russian wheat aphid in *Triticum* species amphiploids and *Triticum tauschii. Canadian Journal of Plant Science* **71**: 703-708.

Paux E, Legeai F, Guilhot N, Adam-Blondon AF, Alaux M, Salse J, Sourdille P, Leroy P, Feuillet C (2008a) Physical mapping in large genomes: accelerating anchoring of BAC contigs to genetic maps through in silico analysis. *Functional and Integrative Genomics* **8**: 29-32.

Paux E, Sourdille P, Salse J, Saitenac C, Choulet F, Leroy P, korol A, Michalak M, Kianian S, Spielmeyer W, Lagudah E, Somers D, Kilian A, Alaux M, Vautrin S, Bergès H, Eversole K, Appels R, Safar J, Simkova H, Dolezel J, Bernard M, Feuillet C (2008b) A physical map of the 1-gigabase bread wheat chromosome 3B. *Science* **322**: 101-104.

Paux E, Faure S, Choulet F, Roger D, Gauthier V, Martinant JP, Sourdille P, Balfourier F, LePaslier 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 Biotechnology Journal* **8**: 196-210.

Peng JH, Wang H, Haley SD, Peairs FB, Lapitan NLV (2007) Molecular mapping of the Russian wheat aphid resistance gene Dn2414 in wheat. *Crop Science* **47**: 2418-2429.

Peng JH, Bai Y, Haley SD, Lapitan NLV (2009) Microsatellite-based molecular diversity of bread wheat germplasm and association mapping of wheat resistance to the Russian wheat aphid. *Genetica* **135**: 95–122.

Pfeifer M, Kugler KG, Sandve SR, Zhan B, Rudi H, Hvidsten TR, International Wheat Genome Sequencing Consortium, Mayer KF, Olsen OA (2014) Genome interplay in the grain transcriptome of hexaploid bread wheat. Science 345:1250091.

Philippe R, Choulet F, Paux E, van Oeveren J, Tang J, Wittenberg AHJ, Janssen A, van Eijk MJT, Stormo K, Alberti A, Wincker P, Akhunov E, van der Vossen E, Feuillet C (2012) Whole genome profilig provides a robust framework for physical mapping and sequencing in the highly complex and repetitive wheat genome. *BMC Genomics* **13**: 47.

Philippe R, Paux E, Bertin I, Sourdille P, Choulet F, Laugier C, Šimková H, Šafář J, Bellec A, Vautrin S, Frenkel Z, Cattonaro F, Magni F, Scalabrin S, Martis MM, Mayer KF, Korol A, Bergès H, Doležel 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**: R64.

Poursarebani N, Nussbaumer T, Šimková H, Šafář J, Witsenboer H, van Oeveren J, Doležel J, Mayer KF, Stein N, Schnurbusch T (2014) Whole-genome profiling and shotgun sequencing delivers an anchored, gene-decorated, physical map assembly of bread wheat chromosome 6A. *Plant Journal* **79**: 334-347.

Presting GG, Malysheva L, Fuchs J, Schubert I (1998) A TY3/GYPSY retrotransposon-like sequence localizes to the centromeric regions of cereal chromosomes. *Plant Journal* **16**: 721-728.

Putnam NH, O'Connell BL, Stites JC, Rice BJ, Blanchette M, Calef R, Troll CJ, Fields A, Hartley PD, Sugnet CW, Haussler D, Rokhsar DS, Green RE (2016) Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Research* **26**(3): 342-350.

Qi LL, Echalier B, Lazo GR, Butler GE, Anderson OD, Akhunov ED, Dvořák J, Linkiewicz AM, Ratnasiri A, Dubcovsky J, Bermudez-Kandianis CE, Greene RA, Kantety R, La Rota CM, Munkwold JD, Sorrells SF, Sorrells ME, Dilbirligi M, Sidhu D, Erayman M, Randhawa

M, 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 tags loci and distribution of genes among the three genomes of polyploid wheat. *Genetics* **168**: 701-712.

Raats D, Frenkel Z, Krugman T, Dodek I, Sela H, Šimková H, Mahni F, Cattonaro F, Vautrin S, Bergès H, Wicker T, Keller B, Leroy P, Philippe R, Paux E, Doležel 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**: R138.

Rabinowicz PD, citek R, Budiman MA, Nunberg A, Bedell JA, Lakey N, O'Shaughnessy, Nascimento LU, McCombie WR, Martienssen RA (2005) Differential methylation of genes and repeats in land plants. *Genome Research* **15**: 1431-1440.

Rambow F, Rogiers A, Marin-Bejar O, Aibar S, Femel J, Dewaele M, Karras P, Brown D, Chang YH, Debiec-Rychter M, Adriaens C, Radaelli E, Wolter P, Bechter O, Dummer R, Levesque M, Piris A, Frederick DT, Boland G, Flaherty KT, van der Oord J, Voet T, Aerts S, Lund AW, Marine JC (2018) Toward minimal residual disease-directed therapy in melanoma. *Cell* **174**(4): 843-855.

Šafář J, Bartoš J, Janda J, Belec A, Kubaláková M, Valárik M, Pateyron S, Weiserová J, Tušková R, Čihalíková J, Vrána J, Šimková H, Faivre-Rampant P, Sourdille P, Caboche M, Bernard M, Doležel J, Chalhoub B (2004) Dissecting large and complex genomes: flow sorting and BAC cloning of individual chromosomes from bread wheat. *Plant Journal* **39**: 960-968.

Šafář J, Šimková H, Kubaláková M, Čihalíková J, Suchánková P, Bartoš J, Doležel J (2010) Development of chromosome-specific BAC resources for genomics of bread wheat. *Cytogenetic Genome Research* **129**: 211-223.

Saidi A and Quick JS (1996) Inheritance and allelic relationships among Russian wheat aphid resistance genes in winter wheat. *Crop Science* **36**:256–258.

Salina A, Nesterov MA, Frenkel Z, Kiseleva AA, Timonova EM, Magni F, Vrána J, Šafář J, Šimková H, Doležel J, Korol A, Sergeeva EM (2018) Features of the organization of bread wheat chromosome 5BS based on physical mapping. *BMC Genomics* **19**:80.

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak P, Liang CZ, Zhang JW, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlinson C, Strong C, Delehaunty K, Fronick C, Courtney B, Rock SM, Belter E, Du FY, Kim K, Abbott RM, Cotton

B, Levy A, Marchetto P, Ochoa K, Jackson SM, Gillam B, Chen WZ, Yan L, Higginbotham J, Cardenas M, Waligorski J, Applebaum E, Phelps L, Falcone J, Kanchi K, Thane T, Scimone A, Thane N, Henke J, Wang T, Ruppert J, Shah N, Rotter K, Hodges J, Ingenthron E, Cordes M, Kohlberg S, Sgro J, Delgado B, Mead K, Chinwalla A, Leonard S, Crouse K, Collura K, Kudrna D, Currie J, He RF, Angelova A, Rajasekar S, Mueller T, Lomeli R, Scara G, Ko A, Delaney K, Wissotski M, Lopez G, Campos D, Braidotti M, Ashley E, Golser W, Kim H, Lee S, Lin JK, Dujmic Z, Kim W, Talag J, Zuccolo A, Fan C, Sebastian A, Kramer M, Spiegel L, Nascimento L, Zutavern T, Miller B, Ambroise C, Muller S, Spooner W, Narechania A, Ren LY, Wei S, Kumari S, Faga B, Levy MJ, McMahan L, Van Buren P, Vaughn MW, Ying K, Yeh CT, Emrich SJ, Jia Y, Kalyanaraman A, Hsia AP, Barbazuk WB, Baucom RS, Brutnell Tp, Carpita NC, Chaparro C, Chia JM, Deragon JM, Estill JC, Fu Y, Jeddeloh JA, Han YJ, Lee H, Li PH, Lisch DR, Liu SZ, Liu ZJ, Nagel DH, McCann MC, SanMiguel P, Myers AM, Nettleton D, Nguyen J, Penning BW, Ponnala L, Schneider KL, Schwartz DC, Sharma A, Soderlund C, Springer NM, Sun Q, Wang H, Waterman M, Westerman R, Wolfgruber TK, Yang LX, Yu Y, Zhang LF, Zhou SG, Zhu Q, Bennetzen JL, Dawe RK, Jiang JM, Jiang N, Presting GG, Wessler SR, Aluru S, Martienssen RA, Clifton SW, McCombie WR, Wing RA, Wilson RK (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**: 1112–1115.

Schnable JC, Wang X, Pires J, Freeling M (2012) Escape from preferential retention following repeated whole genome duplication in plants. *Frontiers in Plant Science* **3**.

Schultz T, van Eck L, Botha AM (2015) Phi-class glutathione-S-transferase is involved in Dn1-mediated resistance. *Physiologia Plantarum* **154**: 1-12.

Sears ER, Sears LMS (1978) The telocentric chromosomes of common wheat. In: Ramanujams S (ed) Proc 5th International Wheat Genetics Symposium pp 389-407 Indian Agricultural Research Institute New Delhi.

Shiferaw B, Smale M, Braun HJ, Du veiller E, Reynolds , Muricho G (2013) Crops that feed the world 10. Past successes and future challenges to the role played by wheat in global food security. *Food Security* **5**: 291-317.

Shizuya H, birren B, Kim UJ, Mancino V, Slepak T, Tachiiri Y, Simon M (1992) Clonnig and stable maintenance of 300-kilobase-pair fragment of human DNA in Escherichia coli using F-factor-based vector. *Proceeding of the National Academy of Sciences of the United States of America* **89**: 8794-8797.

Šimková H, Šafář J, Kubaláková M, Suchánková P, Čihalíková J, Robert-Quarte H, Azhaguvel P, Weng Y, Peng J, Lapitan NLV, Ma Y, Your FM, Luo MC, Bartoš J, Doležel J

(2011) BAC libraries from wheat chromosome 7D: efficient tool for positional cloning of aphid resistance genes. *Journal of Biomedicine and Biotechnology* **2011**: 302543.

Šimková H (2017) Nové poznatky v genetice rostlin V. Genomy obilovin (téměř) dočteny. *Živa* **3/2017**: 111-114.

Smith CM (2005) Plant resistance to arthropods. Molecular and conventional approaches. Springer Dordrecht, The Netherlands.

Smith CM, Boyko EV (2007) The molecular bases of plant resistance and defense responses to aphid feeding: current status. *Entomologia Experimentalis et Applicata* **122**: 1-16.

Soderlund C, Humphray S, Dunham A, French L (2000) Contig built with fingerprints, markers and FPC v4.7. *Genome Research* **10**: 1772-1787.

Springer NM, Anderson SN, Andorf CM, Ahern KR, Bai F, Barad O, Barbazuk WB, Bass HW, Baruch K, Ben-Zvi G, Buckler ES, Bukowski R, Campbell MS, Cannon EKS, Chomet P, Dawe RK, Davenport R, Dooner HK, Du LH, Du C, Easterling KA, Gault C, Guan JC, Hunter CT, Jander G, Jiao Y, Koch KE, Kol G, Köllner TG, Kudo T, Li Q, Lu F, Mayfield-Jones D, Mei W, McCarty DR, Noshay JM, Portwood II JL, Ronen G, Settles AM, Shem-Tov D, Shi J, Soifer I, Stein JC, Stitzer MC, Suzuki M, Vera DL, Vollbrecht E, Vrebalov JT, Ware D, Wei S, Wimalanathan K, Woodhouse MR, Xiong W, Brutnell TP (2018) The maize W22 provides a foundation for functional genomics and transposon biology. *Nature Genetics*, doi: 10.1038/s41588-018-0158-0

Staňková H (2015) Construction of physical map of 7DS wheat chromosome arm and its use for positional clonig. PhD thesis. The Palacký University Olomouc. Faculty of Science.

Staňková H, Valárik M, Lapitan NLV, Berkman PJ, Batley J, Edwards D, Luo MC, Tulpová Z, Kubaláková M, Stein N, Doležel J, Šimková H (2015) Chromosomal genomics facilitates fine mapping of Russian wheat aphid resistance gene. *Theoretical and Applied Genetics* **128**: 1373-1383.

Staňková H, Hastie A, Chan S, Vrána J, Tulpová Z, Kubaláková M, Visendi P, Hayashi S, Luo MC, Batley J, Edwards D, Doležel J, Šimková H (2016) BioNano genome mapping of individual chromosomes supports physical mapping and sequence assembling in complex plant genomes. *Plant Biotechnology Journal* **14:** 1523 – 1531.

Teh BT, Lim K, Yong CH, Ng CCY, Rao SR, Rajasegaran V, Lim WK, Ong CK, Chan K, Cheng VKY, Soh PS, Swarup S, Rozen SG, Nagarajan N, Tan P (2017) The draft genome of tropical fruit durian (Durio zibethinus). *Nature Genetics* **49**: 1633-1641.

Tiwari VK, Heesacker A, Riera-Lizarazu O, Gunn H, Wang S, Wang Y, Gu YQ, Paux E, Koo DH, Kumar A, Luo MC, Lazo G, Zemetra R, Akhunov E, Friebe B, Poland J, Gill BS, Kianian S, Leonard JM (2016) A whole-genome, radiation hybrid mapping resource of hexaploid wheat. *Plant Journal* **86(**2): 195-207.

Toegelová H (2016) Optické mapování aneb jak vidět sekvenci DNA na vlastní oči. *Živa* **6/2016**: 302-304.

Tonk FA, İştipliler D, Tosun M, Turanli F, İlbi H, Çakir M (2016) Genetic mapping and inheritance of Russian wheat aphid resistance gene in accession IG 100695. *Plant Breeding* **135**: 21-25.

Tulpová Z, Luo MC, Toegelová H, Visendi P, Hayashi S, Vojta P, Paux E, Kilian A, Abrouk M, Bartoš J, Hajdúch M, Batley J, Edwards D, Doležel J, Šimková H (2019a) Integrated physical map of bread wheat chromosome arm 7DS to facilitate gene cloning and comparative studies. *New Biotechnology* 48:12-19.

Tulpová Z, Toegelová H, Lapitan NLV, Peairs F, Macas J, Novák P, Lukaszewski A, Kopecký D, Mazáčová M, Vrána J, Holušová K, Leroy P, Doležel J, Šimková H (2019b) Accessing a Russian wheat aphid resistance gene in bread wheat by long-read technologies. *Plant Genome*, doi: 10.3835/plantgenome2018.09.0065

USDA – United States Department of Agriculture, Economic Research Service (2018), http:// https://www.ers.usda.gov/data-products/wheat-data/

Valdez VA, Byrne PF, Lapitan NLV, Peairs FB, Bernardo A, Bai G, Haley SD (2012) Inheritance and genetic mapping of Russian wheat aphid resistance in Iranian wheat landrace accession PI626580. *Crop Science* **52**: 676-682.

Van Eck L, Davidson RM, Wu S, Zhao BY, Botha AM, Leach JE, Lapitan NLV (2014) The transcriptional network of WRKY53 in cereals links oxidative responses to biotic and abiotic stress inputs. *Functional Integrative Genomics* **27**: 351-362.

Van Oeveren J, de Ruiter M, Jesse T, van der Poel H, Tang J, Yalcin F, Janssen A, Volpin H, Stormo KE, Bogden R, van Eijk MJT, Prins M (2011) Sequence-based physical mapping of complex genomes by whole genome profiling. *Genome Research* **21**: 618-625.

Varshney RK, Graner A, Sorrells ME (2005) Genic microsatellite markers in plants: features and applications. *Trends in Biotechnology* **23**(1): 48-55.

Varshney RK, Nayak SN, May GD, Jackson SA (2009) Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends in Biotechnology* **27**: 522-530.

Voothuluru P, Meng J, Khajuria C, Louis J, Zhu L, Starkey S, Wilde GE, Baker CA, Smith CM (2006) Categories and inheritance of resistance to Russian wheat aphid Hmoptera: *Aphididade*) biotype 2 in a selection from wheat cereal introduction 2401. *Journal of Economic Entomology* **99**: 1854-1861.

Vrána J, Kubaláková M, Šimková H, Čihalíková J, Doležel J (2000) Flow-sorting of mitotic chromosomes in common wheat (*Triticum aestivum* L). *Genetics* **156**: 2033-2041.

Vrána J, Šimková H, Kubaláková M, Čihalíková, Doležel J (2012) Flow cytometric chromosome sorting in plants: The next generation. *Methods* **57**: 331-337.

Wang K, Wu Y, Zhang W, Dawe RK, Jiang J (2014) Maize centromeres expand and adopt a uniform size in the genetic background of oat. *Genome Researc*h **12**: e1005997

Wang M, Wang S, Xia G (2015) From genome to gene: a new epoch for wheat research? *Trends in Plant Science* **20**: 380-387.

Wang T, Quisenberry SS, Ni X, Tolmay V (2004) Enzymatic chlorophyll degradation in wheat near-isogenic lines elicited by cereal aphid (Homoptera: Aphididae) feeding. *Journal of Economic Entomology* **97**: 661-667

Wei F, Coe E, Nelson W, Bharti AK, Engler F, Butler E, Kim H, Goicoechea JL, Chen M, Lee S, Fuks G, Sanchez-Villeda H, Schroeder S, Fanf Z, McMullen M, Davis G, Bowers JE, Paterson AH, Schaeffer M, Gardiner J, Cone K, Messing J, Soderlund C, Wing RA (2007) Physical and genetic structure of the maize genome reflects its complex evolutionary history. *PLoS Genetics* **3**: e123.

Wicker T, Matthews DE, Keller B (2002) TREP: a database for *Triticeae* repetitive elements. *Trends in Plants Science* **7**(12): 561-562.

Wicker T, Gundlach H, Spannagl M, Uauy C, Borrill P, Ramírez-Gonzáles RH, De Oliveira R, IWGSC, Mayer KFX, Paux E, Choulet F (2018) Impact of transposable elements on genome structure and gene evolution in bread wheat. *Genome Biology* **19**: 103.

Wong KHY, Levy-Sakin M, Kwok PY (2018) De novo human genome assemblies reveal spectrum of alternative haplotypes in diverse populations. *Nature Communications* **9**: 3040.

Woodhouse MR, Schnable JC, Pedersen BS, Lyons E, Lisch D, Subramaniam S, Freeling M (2010) Following tetraploidy in maize, a short deletion mechanism removed genes preferentially from one of the two homeologs. *PLOS Biology*, doi: 10.1371/journal.pbio.1000409

Xu X, Bai G, Carver BF, Zhan K, Huang Y, Mornhinweg D (2015) Evaluation and resection of wheat resistance to Russian wheat aphid bitype 2. *Crop Science* **55**: 695-701.

Xu Z, Sun S, Covaleda L, Ding K, Zhang A, Wu C, Scheurin C, Zhang HB (2004) Genome physical mapping with large-insert bacterial clones by fingerprinting analysis: methodologies, source clone genome coverage, and contig map quality. *Genomics* **84**: 941-951.

Yahiaoui N, Srichumpa P, Dudler R, Keller B (2004) Genome analysis at different ploidy levels allows cloning of the powdery mildew resistance gene *Pm3b* from hexaploid wheat. *Plant Journal* **37**:528-538.

Yazdani M, Baker G, DeGraaf H, Henry K, hill K, Kimber B, Malipatil M (2017) First detection of Russian wheat aphid *Diuraphis noxia* Kordjumov (Hemiptera: Aphididae) in Australia: a major threat to cereal production. *Austral Entomology*, doi: 10.1111/aen.12292

Yim YS, moak P, Sanchez-Villeda H, Musket TA, Close P, Klein PE, Mullet JE, McMullen M, Fang Z, Schaeffer ML, Gardiner JM, Coe EH Jr., Davis GL (2007) A BAC pooling strategy combined with PCR-based screening in a large, higly repetitive genome enables integration of the maize genetic and physical maps. *BMC Genomics* **8**: 47.

Zhao G, Zou C, Li K, Wang K, Li T, Gao L, Zhang X, Wang H, Yang Z, Liu X, Jiang W, Mao L, Kong X, Jiao Y, Jia J (2017) The *Aegilops tauschii* genome reveals multiple impacts of transposons. *Nature Plants* **3**: 946-955.

Zimin AV, Puiu D, Luo MC, Zhu T, Korens S, Marçais G, Yorke JA, Dvořák J, Salzberg SL (2017a) Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome Research* **27**: 787-792.

Zimin AV, Puiu D, Hall R, Kingan S, Clavijo BJ, Salzberg SL (2017b) The first near-complete assembly of the hexaploid bread wheat genome, *Triticum aestivum*. *GigaScience* **6**: 1-7.

2 AIMS OF THE THESIS

I Anchoring and validation of the physical map of the short arm of wheat chromosome 7D

The first aim of the thesis is to anchor a 7DS physical contig map to the chromosome through integration with several types of genetic and physical genomic resources, including a physical map of the D genome progenitor, *Aegilops tauschii*. This integration will be used to compare structure of the 7DS between bread wheat and its ancestor.

II Sequencing of the chromosome arm 7DS and assembling of the reference sequence

The second aim of the work will be sequencing of a minimal set of BAC clones continuously covering the entire 7DS arm, assembling and annotating of the obtained sequence. The generated data and other genomic resources will be used to assemble and validate the reference sequence of the 7DS arm.

III Positional cloning of a Russian wheat aphid resistance gene

The third aim of the thesis will be application of the BAC clone assembly in positional cloning of a Russian wheat aphid resistance gene and in analyses of other regions of interest.

3 RESULTS

3.1. Summary

In this thesis I focused on the study of the wheat chromosome arm 7DS. The first goal of the current work was to anchor and validate the clone-based physical map of the 7DS. This effort ran in parallel with sequencing of 7DS minimal tilling path, deconvolution and assembly of BAC clone sequences and completion and validation of the 7DS reference sequence. Moreover, all these datasets were exploited in identification of a candidate for a gene underlying resistance to Russian wheat aphid, a serious cereal pest dispersed worldwide.

To anchor the 7DS physical map, several approaches were applied, including manual anchoring by PCR on 3D BAC pools or *in silico* anchoring that utilized the generated BAC sequences. Final version of the 7DS physical map (Tulpová *et al*., 2019a) integrated markers from a radiation hybrid map and three genetic maps, including one from the D-genome ancestor, *Ae. tauschii*. Besides, our approach to physical-map assembly included integration of the 7DS physical map with a whole-genome map of *Ae. tauschii* (Luo *et al.*, 2013). This together with involvement of a Bionano genome (BNG) map of the 7DS arm (Staňková *et al.*, 2016) facilitated ordering of physical-map contigs even in the non-recombining region of the genetic centromere.

Within a joint effort coordinated by the IWGSC, 4,608 MTP BAC clones from a 7DS-specific BAC library were sequenced in BAC pools of four non-overlapping clones using Illumina HiSeq platform. Resulting sequences were assembled into contigs with N50 of 72 kb. Their scaffolding was done using mate-pair data obtained from MTP-plate pools (384 clones). The final 7DS BAC assembly, composed of 9,063 scaffolds with N50 of 117 kb (Tulpová *et al*., 2019a), became a valuable data source contributing to the reference sequence of the 7DS chromosome arm (IWGSC, 2018).

In assembling the 7DS physical map and subsequent analysis of the 7DS sequence, we paid a special attention to the 7D centromere. Integration of the 7DS physical map with the map of *Ae. tauschii* enabled comparison of the two genomes and revealed a megabase-size rearrangement in the genetic centromere between wheat and its ancestor. Using two BNG maps prepared from 7DS of cv. Chinese Spring

(Staňková *et al*., 2016) and line CI2401 (Tulpová *et al*., 2019b), respectively, and a BNG map of chromosome arm 7DL (IWGSC, 2018), we were able to span the centromere in the reference sequence of the 7D chromosome (IWGSC, 2018) and prove that the assembly was correct and nearly complete in this challenging region. Position of the functional 7D centromere was predicted based on alignment of the 7DS and the 7DL BNG maps and sequences to the 7D assembly and was confirmed by mapping available (Guo *et al*., 2016) ChIP-seq reads for CENH3, a centromere-specific histone H3 variant, on the 7D sequence. This delimited the functional centromere of the 7D as a region of ~6 Mb. The ChIP-seq based approach was then applied to delimiting functional centromeres in all wheat chromosomes and enabled an estimate of average wheat centromere size of 6.7 Mb (IWGSC, 2018) (Figure 17).

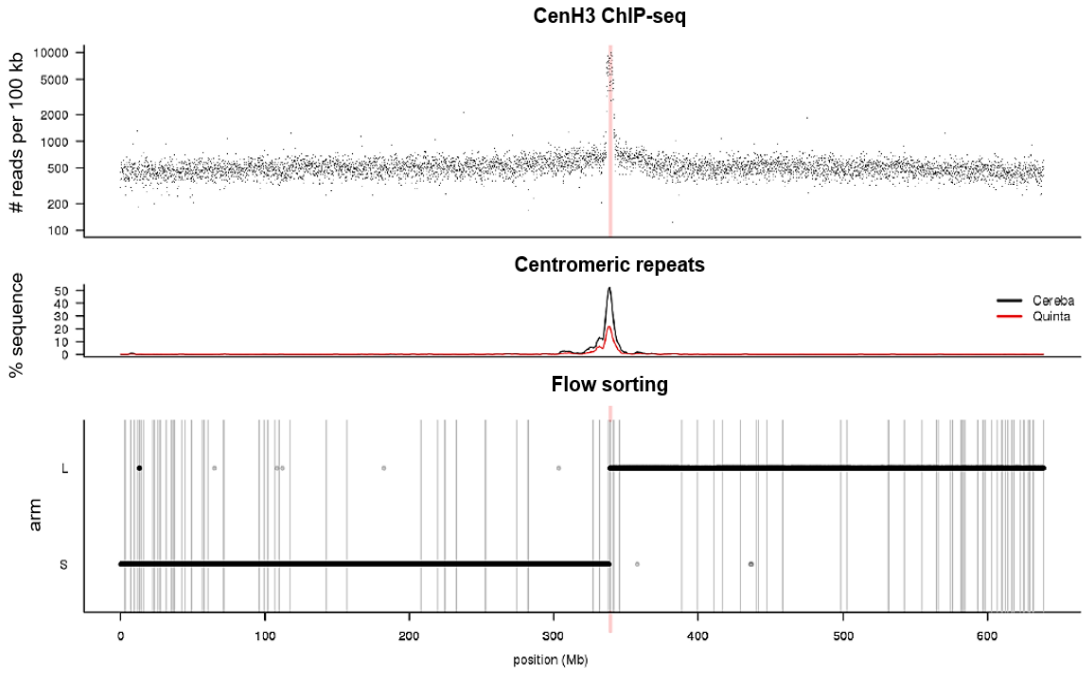


Figure 17: Positioning of the centromere in the 7D pseudomolecule. Top panel shows density of CENH3 ChIP-seq data along the wheat chromosome. Middle panel shows distribution and proportion of the total pseudomolecule sequence composed of TEs of the Cereba and Quinta families. Bottom panel shows positions of 7DS- and 7DL-specific sequences mapped on the 7D pseudomolecule. The core centromere is highlighted with the pink line (IWGSC, 2018).

The 7DS BAC assembly was also utilized in positional cloning project targeting *Dn2401* gene underlying resistance to RWA, a serious pest of small grain cereals and many grass species. In the previous study of Staňková *et al.* (2015), ~300-kb interval containing the *Dn2401* resistance gene was delimited and five BAC clones spanning this region were selected. Here we used a targeted strategy that combined traditional approaches towards gene cloning, comprising genetic mapping and Illumina sequencing of BAC clones, with novel technologies including optical mapping and long-read nanopore sequencing. Comparison of the obtained BAC hybrid assembly covering the gene region with corresponding parts in two wheat whole genome assemblies (IWGSC, 2018; Zimin *et al*, 2017b) revealed misassemblies in a close proximity of predicted candidate genes. The highly accurate BAC assembly facilitated precise annotation of the *Dn2401* region, saturation of the interval with new markers and proposing and resequencing of candidate genes. Identification of *Epoxide hydrolase 2* as the most likely *Dn2401* candidate opened an avenue to its validation by functional genomics approaches (Tulpová *et al.*, 2019b).

3.2 Original papers

3.2.1 Integrated physical map of bread wheat chromosome arm 7DS to facilitate gene cloning and comparative studies

(Appendix I)

3.2.2 Accessing a Russian wheat aphid resistance gene in bread wheat by long-read technologies

(Appendix II)

3.2.3 Shifting the limits in wheat research and breading using a fully annotated reference genome

(Appendix III)

3.2.1 Integrated physical map of bread wheat chromosome arm 7DS to facilitate gene cloning and comparative studies

Tulpová Z, Luo MC, Toegelová H, Visendi P, Hayashi S, Vojta P, Paux E, Kilian A, Abrouk M, Bartoš J, Hajdúch M, Batley J, Edwards D, Doležel J, Šimková H

New Biotechnology, 48: 12-19, 2019

doi: 10.1016/j.nbt.2018.03.0003

IF: 3.733

Abstract:

Bread wheat (*Triticum aestivum* L.) is a staple food for a significant part of the world's population. The growing demand on its production can be satisfied by improving yield and resistance to biotic and abiotic stress. Knowledge of the genome sequence would aid in discovering genes and QTLs underlying these traits and provide a basis for genomics-assisted breeding. Physical maps and BAC clones associated with them have been valuable resources from which to generate a reference genome of bread wheat and to assist map-based gene cloning. As a part of a joint effort coordinated by the International Wheat Genome Sequencing Consortium, we have constructed a BAC-based physical map of bread wheat chromosome arm 7DS consisting of 895 contigs and covering 94% of its estimated length. By anchoring BAC contigs to one radiation hybrid map and three high resolution genetic maps, we assigned 73% of the assembly to a distinct genomic position. This map integration, interconnecting a total of 1,713 markers with ordered and sequenced BAC clones from a minimal tiling path, provides a tool to speed up gene cloning in wheat. The process of physical map assembly included the integration of the 7DS physical map with a whole-genome physical map of *Aegilops tauschii* and a 7DS Bionano genome map, which together enabled efficient scaffolding of physical-map contigs, even in the non-recombining region of the genetic centromere. Moreover, this approach facilitated a comparison of bread wheat and its ancestor at BAC-contig level and revealed a reconstructed region in the 7DS pericentromere.

3.2.2 Accessing a Russian wheat aphid resistance gene in bread wheat by long-read technologies

Tulpová Z, Toegelová H, Lapitan NLV, Peairs F, Macas J, Novák P, Lukaszewski A, Kopecký D, Mazáčová M, Vrána J, Holušová K, Leroy P, Doležel J, Šimková H

Plant Genome

doi: 10.3835/plantgenome2018.09.0065

IF: 2.923

Abstract:

Russian wheat aphid (RWA) is a serious invasive pest of small grain cereals and many grass species. An efficient strategy to defy aphid attacks is to identify sources of natural resistance and transfer resistance genes into susceptible crop cultivars. Revealing the genes helps understand plant defence mechanisms and engineer plants with a durable resistance to the pest. To date, more than 15 RWA resistance genes have been identified in wheat, but none of them has been cloned. Previously, we genetically mapped RWA resistance gene *Dn2401* into an interval of 0.83 cM and spanned it with five BAC clones. Here we used a targeted strategy that combines traditional approaches towards gene cloning, comprising genetic mapping and sequencing of BAC clones, with novel technologies, including optical mapping and long-read nanopore sequencing. The latter, with reads spanning the entire length of a BAC insert, enabled us to assemble the whole region, the task not achievable with short Illumina reads only. Long-read optical mapping validated DNA sequence in the interval and revealed a difference in the locus organization between resistant and susceptible genotypes. The complete and accurate sequence of the *Dn2401* region facilitated its precise annotation, saturation of the interval with new markers and suggestion and resequencing of candidate genes. Identification of *Epoxide hydrolase 2* as the most likely *Dn2401* candidate opens an avenue to its validation by functional genomics approaches.

3.2.3 Shifting the limits in wheat research and breading using a fully annotated reference genome

International Wheat Genome Consortium

Science 361, eaar7191

doi: 10.1126/science.aar7191

IF: 41.058

Abstract:

An annotated reference sequence representing the hexaploid bread wheat genome in 21 pseudomolecules has been analyzed to identify the distribution and genomic context of coding and non-coding elements across the A, B and D sub-genomes. With an estimated coverage of 94% of the genome and containing 107,891 high-confidence gene models, this assembly enabled the discovery of tissue and developmental stage related co-expression networks using a transcriptome atlas representing all stages of wheat development. Dynamics of complex gene families involved in environmental adaptation and end-use quality were revealed at sub-genome resolution and contextualized to known agronomic single gene or quantitative trait loci. This community resource establishes the foundation for accelerating wheat research and application through improved understanding of wheat biology and genomics-assisted breeding.

3.3 Published abstracts – poster presentation

3.3.1 Integrated physical map of bread wheat chromosome arm 7DS to support evolutionary studies and genome sequencing

(Appendix IV)

3.3.2 Completing reference sequence of the wheat chromosome arm 7DS

(Appendix V)

3.3.3 Structural variation of a wheat chromosome arm revealed by optical mapping

(Appendix VI)

3.3.4 Poziční klonování genu pro rezistenci k mšici zhoubné (*Diuraphis noxia*): konstrukce vysokohustotní genetické mapy

(Appendix VII)

3.3.1 Integrated physical map of bread wheat chromosome arm 7DS to support evolutionary studies and genome sequencing

Tulpová Z, Luo MC, Toegelová H, Visendi P, Hayashi S, Vojta P, Hastie A, Kilian A, Tiwari VK, Bartoš J, Batley J, Edwards D, Doležel J, Šimková H

In: Abstracts of the "Olomouc Biotech 2017. Plant Biotechnology: Green for Good IV". Olomouc, Czech Republic 2017

Abstract:

Bread wheat (*Triticum aestivum* L.) is one of the most important crops species worldwide. Polyploid nature of wheat genome (2n = 6x = 42, AABBDD) together with huge genome size (~17 Gb) and prevalence of repetitive sequences hamper its studying. To overcome these obstacles, BAC-by-BAC sequencing strategy based on chromosome-derived physical maps has been used to support a joint effort of the International Wheat Genome Sequencing Consortium (IWGSC) towards obtaining a reference genome sequence.

Our work was focused on mapping and sequencing of the short arm of wheat chromosome 7D (7DS). Clones from 7DS-specific BAC library were fingerprinted using SNaPshot-based HICF technology and automatically assembled into contigs using FPC software. Integration of initial 7DS physical map with that of *Aegilops tauschii* (D genome ancestor) provided a clue for further merging of automatically built contigs. The assembly was verified using LTC software. Physical map has been anchored on chromosome using different approaches of forward anchoring including PCR screening of the 7DS library, or *in silico* anchoring using several genetic maps and markers or 7DS specific optical map. Hitherto, we have positioned 1864 markers in 470 contigs and anchored another 37 marker-missing contigs through an optical map generated for the 7DS. Thus we anchored 74 % of the 7DS physical map in total.

Our sequencing strategy involved pair-end sequencing of pools of four non-overlapping MTP BAC clones and mate-pair sequencing of MTP plate pools

(384 clones) using the Illumina HiSeq. Obtained reads were assembled into sequence contigs using assemblers Sassy and SSPACE.

3.3.2 Completing reference sequence of the wheat chromosome arm 7DS

Tulpová Z, Toegelová H, Luo MC, Visendi P, Hayashi S, Kilian A, Tiwari VK, Kumar A, Hastie AR, Leroy P, Rimbert H, Abrouk M, Bartoš J, Batley J, Edwards D, Doležel J, Šimková H

In: Abstracts of the International Conference "Plant and Animal Genome XXIV". P.0826. Sherago International, Inc., San Diego, 2016.

Abstract:

Bread wheat genome is characterized by high complexity (~17 Gb), polyploidy (AABBDD) and high content of repetitive sequences (>90%), which hampers genome mapping and sequencing. To overcome these obstacles, BAC-by-BAC sequencing strategy based on chromosome-derived physical maps has been adopted by the International Wheat Genome Sequencing Consortium (IWGSC) as the key strategy towards obtaining a reference genome sequence.

In the framework of this effort, we focus on mapping and sequencing of the 7DS chromosome arm. Clones from a 7DS-specific BAC library were fingerprinted and assembled into contigs using FPC and the map assembly was validated by LTC. Integration of the 7DS physical map with that of *Aegilops tauschii* (D genome progenitor) provided a clue for further merging and landing the contigs on the chromosome. Hitherto, 74% of the physical map length has been anchored to genetic and radiation hybrid maps through 1,577 markers. Our BAC sequencing strategy involved sequencing pools of four non-overlapping clones using the Illumina HiSeq. Pair-end reads thus obtained were assembled using a custom assembler Sassy. Assembling and pool deconvolution were supported by BAC-end sequences and mate-pair data generated for pools of 384 clones. In the resulting assembly, we reached 1.9 scaffolds per BAC clone on average and scaffold N50 of 114 kb. The sequences were automatically annotated by the TriAnnot pipeline and application of wheat 7DS RNAseq data. Moreover, a BioNano genome map has been constructed for the 7DS

and used to validate the physical map, anchor and orient BAC contigs, size gaps, deconvolute sequences from BAC pool data and support sequence scaffolding.

3.3.3 Structural variation of a wheat chromosome arm revealed by optical mapping

Tulpová Z, Toegelová H, Vrána J, Hastie AR, Lukaszewski AJ, Kopecký D, Lapitan NLV, Batley J, Edwards D, IWGSC, Doležel J, Šimková H

In: Abstracts of the International Conference "Plant and Animal Genome XXV". P.0837. Sherago International, Inc., San Diego, 2017.

Abstract:

Optical mapping in nanochannel arrays (BioNano mapping) provides an affordable approach to validate and improve DNA sequence assemblies and to study structural variation among individuals, cultivars or species. Coupling the optical mapping with flow sorting of individual chromosomes of bread wheat enables zooming in on particular regions of the complex (17 Gb) and hexaploid genome, making the studies more cost-efficient and reliable. In this work we purified 7DS chromosome arm (381 Mb) from ditelosomic lines of two accessions - cv. Chinese Spring (CS) and CI2401 - that carry the 7DS arm as a pair of telocentric chromosomes. Using Irys platform, we generated 68 (180x) and 78 Gb (206x) size-filtered data for CS and CI2401, respectively. The data were assembled into 371 (CS) and 468 (CI2401) genome maps with N50 of 1.3 Mb for both accessions. Comparison of the two maps showed a high degree of similarity between them. 83% of the CI2401 optical map length could be aligned to the CS map under high stringency. Alignment of the optical maps to a reference genome, generated from 'Chinese Spring' by IWGSC, revealed a significant variability in (sub)telomeric and (peri)centromeric regions. The two accessions differ in susceptibility to Diuraphis noxia and we previously located a gene underlying resistance to this pest to a 350-kb region on the 7DS. Comparison of optical maps spanning the region disclosed a 8-kb insertion in the proximity of a candidate gene in the susceptible 'Chinese Spring'. The variable regions will be scrutinized on the sequence level.

3.3.4 Poziční klonování genu pro rezistenci ke mšici zhoubné (*Diuraphis noxia*): konstrukce vysokohustotní genetické mapy

Staňková H, Valárik M, Lapitan N, Berkman P, Edwards D, Luo MC, Tulpová Z, Kubaláková M, Stein N, Doležel J, Šimková H

In: Sborník abstrakt, Bulletin České společnosti experimentální biologie rostlin, "6. Metodické dny". Seč, Česká republika, 2014

[In Czech]

Abstrakt:

Pšenice setá (*Triticum aestivum* L.) je jednou z ekonomicky nejvýznamnějších kulturních plodin, poskytující zdroj potravy pro 35 % obyvatel světa. Jedná se o allohexaploidním druh (2n = 6x = 42) s celkovou velikostí genomu téměř 17 x 10^9 bp. Genom je tvořen třemi homeologními subgenomy (A, B a D) a jeho podstatnou část (přes 80 %) tvoří repetitivní sekvence. Všechny výše zmíněné vlastnosti pšeničného genomu znesnadňují jeho analýzu, genetické i fyzické mapování, sekvenování či pozičního klonování. Třídění jednotlivých chromozómů a jejich ramen pomocí průtokové cytometrie umožňuje rozložit tento obrovský genom na malé a snadno analyzovatelné části. Na krátkém rameni chromozómu 7D (7DS) pšenice se nachází řada agronomicky významných genů, včetně genu *Dn2401* pro rezistenci ke mšici zhoubné (*Diuraphis noxia*). Mšice zhoubná je jedním z nejvýznamnějších škůdců pšenice a ječmene. Chemické i biologické postupy hubení nejsou v případě mšice zhoubné dostatečně účinné. Z tohoto důvodu se jeví jako nejvýhodnější způsob ochrany pěstování odrůd nesoucích geny pro rezistenci vůči tomuto škůdci. Konstrukce vysokohustotní genetické mapy pokrývající oblast zkoumaného genu je nezbytná pro jeho následné poziční klonování, tedy izolaci genu na základě jeho pozice na genetické či fyzické mapě. Za účelem konstrukce této mapy byla vyvinuta metoda pro cílené odvozování markerů z úzké oblasti genomu, a to v podmínkách polyploidního genomu pšenice. Tato metoda využívá syntenie mezi pšenicí a jejími příbuznými druhy (ječmen, *Brachypodium*, rýže, čirok, *Aegilops tauschii*) v kombinaci se sekvencemi jednotlivých chromozómů skupiny 7, získanými celochromozómovým

neuspořádaným (shotgun) sekvenováním. Za pomocí genových markerů vymezujících oblast genu na rameni 7DS je možno identifikovat orthologní oblasti v genomech příbuzných druhů. Geny z těchto oblastí poté slouží k nalezení odpovídajících sekvenčních kontigů pocházejících z chromozómů 7A, 7B a 7D pšenice. Na základě polymorfizmů mezi jednotlivými homeologními chromozómy je možno designovat jednolokusové, genomově specifické markery. Popsaným postupem bylo odvozeno 11 nových vysoce specifických markerů, které posloužily k zahuštění mapy v okolí genu *Dn2401*. Skríning BAC knihovny z ramene 7DS umožnil identifikaci kontigu v 7DS-specifické fyzické mapě, který kompletně překlenuje oblast genu. BAC klony z této oblasti byly osekvenovány a v současné době probíhá jejich anotace. Sekvence BAC klonů mohou rovněž posloužit k odvození dalších markerů v blízkosti genu *Dn2401*. Prezentovaná metodika vývoje markerů je obecně aplikovatelná pro jakýkoliv chromozóm pšenice.

4 CONCLUSIONS

4.1 Anchoring and validation of the physical map of the short arm of wheat chromosome7D

Anchored physical map represents a valuable genomic resource to facilitate clone-by-clone sequencing and assist map-based gene cloning. In our study, we prepared a BAC-based physical map of wheat 7DS arm, interconnecting 1,713 markers from a radiation hybrid map and three genetic maps with ordered 7DS BAC clones. These clones are available from library depositories, thus providing an opportunity to access a region of interest with advanced sequencing technologies in a focused and affordable manner. Our approach to the map assembly integrated clones from wheat 7DS with those of *Ae. tauschii*. This enabled comparative analysis of bread wheat and its ancestor, indicating regions of potential rearrangement, which could be validated using BNG map, as demonstrated for the pericentromeric region of the 7DS (Tulpová *et al*., 2019a).

4.2 Sequencing of the chromosome arm 7DS and assembling of the reference sequence

Within this work, 4,608 MTP BAC clones selected from the 7DS physical map were sequenced on Illumina HiSeq platform. Resulting assemblies became a valuable data source contributing to assembly of the7DS reference sequence (IWGSC, 2018).

4.3 Positional cloning of a Russian wheat aphid resistance gene

The BAC assemblies were also utilized in a positional cloning project focusing on *Dn2401* gene underlying resistance to Russian wheat aphid. An improved BAC hybrid assembly, combining short-read Illumina and long-read nanopore data, facilitated precise annotation of the gene region, development of new markers and proposing and resequencing of a candidate gene, which opens an avenue to its validation by functional genomics approaches (Tulpová *et al.*, 2019b). Our result can contribute to breeding novel aphid-resistant cultivars and to revealing molecular basis of the resistance trait.

5 LIST OF ABBREVIATIONS

7DS	short arm of wheat chromosome 7D
ABA	abscisic acid
BAC	bacterial artificial chromosome
BES	BAC-end sequence
BNG	BioNano genome (map)
bp	base pairs
Cas9	CRISPR associated protein 9
CBC	clone-by-clone (sequencing strategy)
CENH3	centromere-specific variant of histone H3
CRW	wheat centromeric repeat
cDNA	complementary DNA
ChIP	chromatin immunoprecipitation
ChIP-seq	chromatin immunoprecipitation sequencing
CRISPR	clustered regularly interspersed palindromic repeats
CS	Chinese Spring
ddNTP	dideoxynucleotide
DNA	deoxyribonucleic acid
ET	ethylene
FISH	fluorescence *in situ* hybridization
FISHIS	fluorescence *in situ* hybridization in suspension
FPC	FingerPrinted contigs
Gb	gigabase pairs
HC	high-confidence (gene)
Hi-C	chromosome-conformation-base mapping
HICF	high-information-content fingerprinting
HR	hypersensitive response
IWGSC	International wheat genome sequencing consortium
JA	jasmonic acid
kb	kilobase pairs

LC	low-confidence (gene)
LTC	linear topology contig
Mb	megabase pairs
MeJA	methyl jasmonate
MTP	minimal tilling path
MYA	million years ago
NGS	next-generation sequencing
NIL	nearly isogenic line
PCR	polymerase chain reaction
POPSEQ	population sequencing
QTL	quantitative trait loci
RE	restriction enzyme
RH	radiation hybrid
RIL	recombinant inbread line
ROS	reactive oxygen species
RWA	Russian wheat aphid
SA	salicylic acid
SNP	single nucleotide polymorphism
TE	transposable element
TREP	transposable elements platform
VIGS	virus induced gene silencing
WGA	whole genome assembly
WGP	whole genome profiling
WGS	whole genome sequencing
YAC	yeast artificial chromosome

6 LIST OF APPENDICES

Original Papers

Appendix I: Integrated physical map of bread wheat chromosome arm 7DS to facilitate gene cloning and comparative studies

Appendix II: Accessing a Russian wheat aphid resistance gene in bread wheat by long-read technologies

Appendix III: Shifting the limits in wheat research and breeding using a fully annotated reference genome

Published abstracts – poster presentation

Appendix IV: Integrated physical map of bread wheat chromosome arm 7DS to support evolutionary studies and genome sequencing

Appendix V: Completing reference sequence of the wheat chromosome arm 7DS

Appendix VI: Structural variation of a wheat chromosome arm revealed by optical mapping

Appendix VII: Poziční klonování genu pro rezistenci k mšici zhoubné (*Diuraphis noxia*): konstrukce vysokohustotní genetické mapy

APPENDIX I

Integrated physical map of bread wheat chromosome arm 7DS to facilitate gene cloning and comparative studies

Tulpová Z, Luo MC, Toegelová H, Visendi P, Hayashi S, Vojta P, Paux E, Kilian A, Abrouk M, Bartoš J, Hajdúch M, Batley J, Edwards D, Doležel J, Šimková H

New Biotechnology, 48: 12-19, 2019

doi: 10.1016/j.nbt.2018.03.0003

IF: 3.733

New BIOTECHNOLOGY 48 (2019) 12–19

Contents lists available at ScienceDirect

New BIOTECHNOLOGY

journal homepage: www.elsevier.com/locate/nbt

Integrated physical map of bread wheat chromosome arm 7DS to facilitate gene cloning and comparative studies

Zuzana Tulpová[a], Ming-Cheng Luo[b], Helena Toegelová[a], Paul Visendi[c], Satomi Hayashi[d], Petr Vojta[e], Etienne Paux[f], Andrzej Kilian[g], Michaël Abrouk[a], Jan Bartoš[a], Marián Hajdúch[e], Jacqueline Batley[h], David Edwards[h], Jaroslav Doležel[a], Hana Šimková[a,*]

[a] Institute of Experimental Botany, Centre of the Region Haná for Biotechnological and Agricultural Research, Šlechtitelů 31, 783 71 Olomouc, Czech Republic
[b] Department of Plant Sciences, University of California, Davis, USA
[c] Natural Resources Institute, University of Greenwich, Chatham Maritime, Kent ME4 4TB, UK
[d] Earth, Environmental and Biological Sciences, Queensland University of Technology, Brisbane, Queensland, Australia
[e] Laboratoty of Experimental Medicine, Institute of Molecular and Translational Medicine, Hněvotínská 5, 779 00 Olomouc, Czech Republic
[f] GDEC, INRA, Université Clermont Auvergne, 5 Chemin de Beaulieu, 63000 Clermont-Ferrand, France
[g] Diversity Arrays Technology Pty Ltd, University of Canberra, Bruce, Australia
[h] School of Biological Sciences and Institute of Agriculture, University of Western Australia, Crawley, WA, Australia

ARTICLE INFO

Keywords:
Triticum aestivum
BAC
BNG map
Aegilops tauschii
Centromere

ABSTRACT

Bread wheat (*Triticum aestivum* L.) is a staple food for a significant part of the world's population. The growing demand on its production can be satisfied by improving yield and resistance to biotic and abiotic stress. Knowledge of the genome sequence would aid in discovering genes and QTLs underlying these traits and provide a basis for genomics-assisted breeding. Physical maps and BAC clones associated with them have been valuable resources from which to generate a reference genome of bread wheat and to assist map-based gene cloning. As a part of a joint effort coordinated by the International Wheat Genome Sequencing Consortium, we have constructed a BAC-based physical map of bread wheat chromosome arm 7DS consisting of 895 contigs and covering 94% of its estimated length. By anchoring BAC contigs to one radiation hybrid map and three high resolution genetic maps, we assigned 73% of the assembly to a distinct genomic position. This map integration, interconnecting a total of 1713 markers with ordered and sequenced BAC clones from a minimal tiling path, provides a tool to speed up gene cloning in wheat. The process of physical map assembly included the integration of the 7DS physical map with a whole-genome physical map of *Aegilops tauschii* and a 7DS Bionano genome map, which together enabled efficient scaffolding of physical-map contigs, even in the non-recombining region of the genetic centromere. Moreover, this approach facilitated a comparison of bread wheat and its ancestor at BAC-contig level and revealed a reconstructed region in the 7DS pericentromere.

Introduction

Bread wheat (*Triticum aestivum* L.) is the staple food for 40% of the world's population. Improvements in yield and tolerance to biotic and abiotic stresses are essential to improve its production to meet the demands of the growing human population. Knowledge of the genome sequence would provide a deep insight into the genome composition and a detailed gene catalogue to facilitate genomics-assisted breeding. Bread wheat is an allohexaploid species (2n = 6x = 42, AABBDD), whose genome arose through spontaneous hybridization between tetraploid durum wheat, *Triticum turgidum* (AABB), and diploid goatgrass, *Aegilops tauschii* (DD), less than 400,000 years ago [1]. Allotetraploid *T.*

Abbreviations: 7DS, short arm of the bread wheat chromosome 7D; BAC, bacterial artificial chromosome; HICF, High Information Content Fingerprinting; FPC, FingerPrinted Contigs; LTC, Linear Topological Contig software; MTP, minimal tiling path; BNG map, Bionano genome map; CS, Chinese Spring; CsxRe genetic map, Chinese Spring x Renan genetic map; RH map, radiation hybrid map; CSS sequence, Chinese Spring survey sequence

* Corresponding author at: Institute of Experimental Botany, Centre of the Region Haná for Biotechnological and Agricultural Research, Šlechtitelů 31, CZ−783 71 Olomouc, Czech Republic.

E-mail addresses: tulpova@ueb.cas.cz (Z. Tulpová), mcluo@ucdavis.edu (M.-C. Luo), toegelova@ueb.cas.cz (H. Toegelová), P.Muhindira@greenwich.ac.uk (P. Visendi), satomi.hayashi@qut.edu.au (S. Hayashi), petr.vojta@upol.cz (P. Vojta), etienne.paux@inra.fr (E. Paux), a.kilian@diversityarrays.com (A. Kilian), abrouk@ueb.cas.cz (M. Abrouk), bartos@ueb.cas.cz (J. Bartoš), marian.hajduch@upol.cz (M. Hajdúch), Jacqueline.batley@uwa.edu.au (J. Batley), Dave.Edwards@uwa.edu.au (D. Edwards), dolezel@ueb.cas.cz (J. Doležel), simkovah@ueb.cas.cz (H. Šimková).

https://doi.org/10.1016/j.nbt.2018.03.003

Available online 08 March 2018

turgidum originated *via* hybridization between diploid *Triticum urartu* (AA) and a diploid species from Sitopsis section (BB). The presence of three homoeologous subgenomes A, B and D and a high content of repetitive sequences (~85%) contribute to the huge genome size (16 Gb/1C) [2], which together impede its mapping and sequencing.

The first coordinated efforts towards obtaining a reference wheat genome date to 2005, when the International Wheat Genome Sequencing Consortium (IWGSC) was established. At that time, a proven strategy to obtain high-quality reference sequences of large genomes was the clone-by-clone approach, i.e. sequencing clones from large-insert DNA libraries ordered in physical maps. This procedure was initially used to produce reference sequences of *Arabidopsis* and rice and more recently also maize and barley [3–5]. To overcome problems due to polyploidy and genome complexity, the IWGSC adopted a chromosome-based strategy, which relied on dissecting the genome by flow-cytometric sorting of chromosomes and/or their arms [7,8], thus significantly reducing sample complexity. The first wheat chromosome-specific BAC library was constructed from chromosome 3 B by Šafář et al. in 2004 [9] and was used to construct a BAC-based physical map of the largest wheat chromosome [10]. Chromosomal BAC libraries were then constructed from all wheat chromosome arms of cv. Chinese Spring [11] (http://olomouc.ueb.cas.cz/dna-libraries/cereals) and laid the basis for constructing chromosomal physical maps for the whole wheat genome [10,12–20] (www.wheatgenome.org). They proved a favorable resource for map-based gene cloning [21] as well as generating a whole-chromosome sequence [22]. The availability of a new assembling algorithm DeNovoMAGIC, which is powerful enough to produce quality assemblies from whole-genome shotgun Illumina data even in large polyploid genomes, such as tetraploid wild emmer wheat [23], led to a change in the IWGSC plan and integration of the whole-genome assembly (WGA) approach as a key component of the sequencing strategy. WGA together with chromosomal maps, available BAC sequences and other resources resulted in a superior bread wheat reference sequence (www.wheatgenome.org), comparable to that obtained for wild emmer wheat.

Even with the progress in genome sequencing and assembly technologies, wheat chromosomal physical maps and the BAC clones they comprise remain a valuable resource, enabling a fast access to and a detailed analysis of a region of interest. This is because even the latest genome assemblies do not completely cover the genome. In the case of emmer wheat, the reference sequence represents 87.5% of the estimated genome size and the missing part is considered a combination of both unresolved repetitive sequences and difficult-to-sequence regions [23]. Moreover a part of sequence scaffolds (4.1% of emmer wheat assembly) remains non-assigned to a specific genome position. A highly complete and accurate sequence of the region of interest is a prerequisite for successful gene cloning. The availability of BAC clones with known genomic context enables focused and affordable re-sequencing of the critical region with more advanced technologies, such as PacBio and Oxford Nanopore, facilitating the identification of the desired gene or gene cluster [24,25].

In the current work, we targeted the short arm of bread wheat chromosome 7D (7DS) with the size of 381 Mb [11], which is known to harbor numerous agronomically important genes and QTLs [26–29]. An early version of a physical map constructed from a 7DS-specific BAC library [30], was used to delimit a region of a gene underlying an aphid resistance [29]. BAC clone Illumina sequences generated in the current study supplemented with OxfordNanopore sequences of a selected BAC clone are being used to identify candidate genes for the pest resistance (Tulpová, unpublished). The presented version of the 7DS physical map integrates markers from a radiation hybrid (RH) map and three genetic maps, including one from the D-genome ancestor *Ae. tauschii*. Our approach to physical map assembly included the previously published integration of a 7DS physical map with a whole-genome physical map of *Ae. tauschii* [31]. Co-assembly with *Ae. tauschii*, supported by a recently generated Bionano genome (BNG) map of the 7DS arm [32], facilitated partial ordering of physical map contigs even in the non-recombining region of the genetic centromere, which generally poses the major challenge in whole-genome assemblies. Moreover, this approach enabled comparison of genomes of bread wheat and its ancestor at a BAC-contig level and indicated a megabase-size region in the genetic centromere showing structural variation between *Ae. tauschii* and bread wheat 7DS.

Materials and methods

Physical map assembly

A total of 49,152 clones from the 7DS-specific BAC library TaaCsp7DShA [30] were previously fingerprinted using SNaPShot-based HICF technology [33], the resulting fingerprints were processed as described in Šimková et al. [30] and 39,765 useful fingerprints were used for automatic contig assembly in FPC [34]. The initial assembly was performed by incremental contig building with a cut-off value of 1×10^{-75}, tolerance of 0.4 bp and gel length 3600 followed by six iterations of single-to-end and end-to-end (Match: 1, FromEnd: 50) merging with decreasing cut-off up to the terminal value 1×10^{-45}. The DQer function was used after each merge to break up all contigs, that contained more than 10% questionable (Q) clones (Step: 3). In parallel, fingerprints of clones from the TaaCsp7DShA library were FPC-assembled together with fingerprints of all clones from *Ae. tauschii* BAC libraries previously used to produce the *Ae. tauschii* physical map [31]. This co-assembly resulted in *Ae. tauschii* supercontigs with embedded 7DS clones (Fig. 1), which were associated with contig signatures from the initial 7DS assembly. 7DS contigs showing apparent overlap in the co-assembly were subjected to end-merging in FPC with decreased stringency (Sulston-score value rising up to 1×10^{-15}).

As the next step, the 7DS assembly was validated using LTC software [35]. 3D view in LTC allowed visualization of missassembled or chimeric contigs, which were disjoined in FPC. Subsequently, contigs of two clones, likely originating from low-quality fingerprints or other wheat chromosomes that contaminated the 7DS fraction in the process of BAC library preparation, were removed (kill function). MTP clones were selected from this (pre-final) 7DS assembly using FPC with the following parameters: min FPC overlap = 25, from end = 0, min shared bands = 12. Final contig assembly merging was done upon the availability of MTP sequence data (see below). Deconvoluted sequence contigs of MTP BAC clones or whole BAC pools were compared using BLASTn (http://blast.ncbi.nlm.nih.gov/) with default parameters. Subjects of comparison were i) clones from different contigs sharing a marker, ii) clones proposed as overlapping by *Ae. tauschii* co-assembly, that could not be joined at 1×10^{-15}, and iii) clones, whose overlap was indicated by 7DS BNG map [32]. Physical-map contigs were merged if fulfilling at least three out of four criteria: marker sharing, sequence overlap, overlap in *Ae. tauschii* co-assembly, overlap supported by BNG map.

MTP sequencing and BNG map alignments

A total of 4608 MTP BAC clones were pair-end sequenced using a pooling strategy in which 96 pools, each consisting of four non-overlapping clones, were indexed and sequenced on a single lane of the Illumina HiSeq2000 platform [32]. DNA sequences were de-multiplexed and assembled using SASSY as described in detail by Visendi et al. [36]. DNA sequence deconvolution was supported by BAC-end sequences (BES), obtained by Sanger sequencing of all MTP clones from both ends, and by utilizing overlaps between BAC clones in physical map contigs. Mate-pair data was obtained by Illumina sequencing MTP-plate pools (384 clones per pool), and mate-pair reads were then applied to build scaffolds by SSPACE as described in Visendi et al. [36]. Alignments of sequence scaffolds to the 7DS BNG map were performed in IrysView 2.1.1. *Cmaps* were generated from *fasta* files of individual

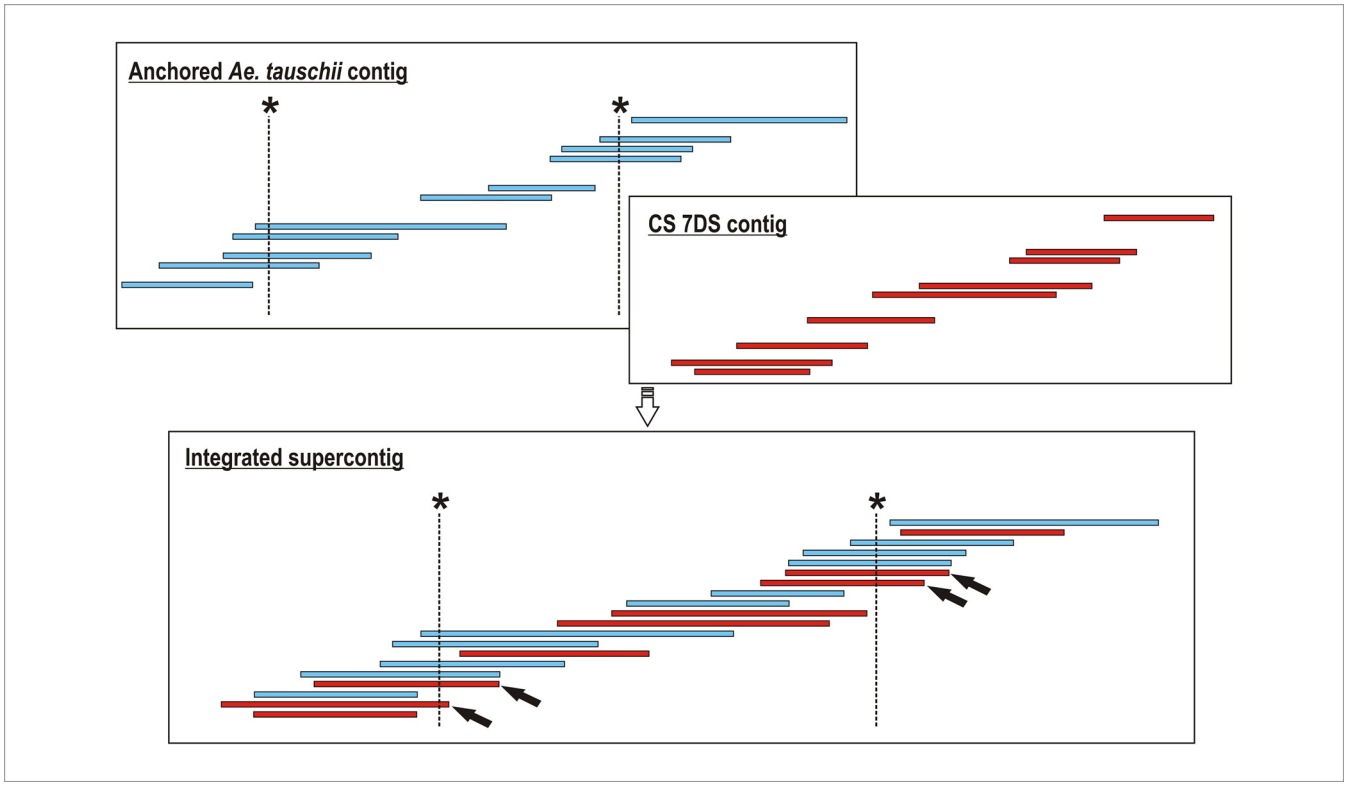


Fig. 1. Contig anchoring based on integration of 7DS and *Ae. tauschii* physical maps. A supercontig resulting from a co-assembly of BAC clones from *Ae. tauschii* (blue bars) and CS 7DS (red bars) comprises *Ae. tauschii* markers (asterisks) anchored to *Ae. tauschii* clones that can be tentatively assigned to 7DS clones (arrows) based on their overlap in FPC visualization. The assignment was validated by BLASTn of marker sequences to 7DS MTP BAC sequences.

MTP BAC clones or BAC pools. Query-to-anchor comparison was performed with default parameters and variable *P*-value threshold ranging from $1e^{-6}$ to $1e^{-10}$.

Physical map anchoring and inter-map comparison

The physical map of 7DS was anchored by combining three approaches. First, the TaaCsp7DShA library was screened with 35 publically available markers (Suppl. Table A) from GrainGenes database (https://wheat.pw.usda.gov/GG2/) using PCR on three-dimensional BAC pools following the procedure described in Šimková et al. [30]. The second approach utilized the 7DS-*Ae. tauschii* map integration described above. *Ae. tauschii* contigs were previously anchored to a linkage map comprising 1159 7D-specific SNP markers [31], part of which could be assigned to particular 7DS BAC clones based on *Ae. tauschii*-7DS clone overlap, as visualized by FPC (Fig. 1). The third approach – *in silico* anchoring – was used to anchor physical map contigs to *Ae. tauschii* linkage map, Chinese Spring x Renan F_6 RIL map (2413 7D SNP markers [37]), Chinese Spring consensus map v.3 (2323 7D DArTseq markers, Suppl. Table B) and Chinese Spring radiation hybrid map (485 7DS SNP markers from iSelect 90 K SNP array, [38,39]). Sequences of all markers were analyzed for homology with 7DS MTP sequences using BLASTn. Only unique hits with > 98% homology along the whole marker length were considered. For markers from the CsxRe map, CSS sequence contigs [40] comprising the markers were applied. To anchor *Ae. tauschii* markers, overlap with marker-bearing *Ae. tauschii* clones in the FPC visualization was required.

To perform inter-map comparison, anchored markers from individual maps were divided into groups of syntenic markers, which were compiled on the basis of co-localization in particular 7DS contigs. Information on genetic position of each marker and its syntenic relationship (Suppl. File A) were processed by Strudel software [41].

Results

Physical map assembly

A total of 39,765 high-quality fingerprints previously generated by SNaPShot HICF technology from a 7DS-specific BAC library [30] were automatically assembled into contigs using FPC, which resulted in assembly of 29,850 clones ordered in 1767 physical map contigs with average size of 264 kb, totaling approximately 468 Mb. This corresponded to 123% of the estimated 7DS chromosome arm length. The largest contig had a size of approximately 2382 kb. The assembly was manually end-merged based on the integration with the *Ae. tauschii* physical map and verified by 3D visualization in LTC. This reduced the number of contigs from 1767 to 1481 and increased their average size to 300 kb, totaling 445 Mb assembly length (117% of 7DS arm length). In the next step, contigs containing two clones were killed, which resulted in 931 physical map contigs comprising 28,339 clones with average contig size of 388 kb, N50 of 528 kb and L50 of 205. This pre-final assembly had estimated assembly length of 362 Mb, which corresponded to 95% of the 7DS arm length (Table 1).

The pre-final assembly was subject to MTP selection using FPC software. This resulted in an MTP of 4608 clones, which were sequenced. This provided an assembly totaling 9063 scaffolds with N50 of 117 kb, average scaffold size 63 kb and 2.5% Ns. On average, we obtained 1.9 scaffolds per BAC clone. The availability of MTP sequences enabled identification of new contig overlaps based on homology revealed by BLASTn search. In this way, 61 cases of potential overlaps were identified on edges of 121 contigs, out of which 71 could be end-merged under a lower-stringency condition in FPC. The remaining 50 contigs did not meet at least two out of three requirements (shared markers, fingerprint overlap with cutoff value $\leq 1 \times 10^{-15}$, clone overlap supported by BNG map) and were not merged. BNG map alignment revealed several misassemblies and proposed one contig

Table 1
Development of 7DS physical map assembly.

Assembly	Initial ($1e^{-45}$)	After integration with *Ae. tauschii* and LTC validation	After killing 2-clone contigs	After manual contig merging
No. contigs	1767	1481	931	895
Average contig size	264 kb	300 kb	388 kb	401 kb
Total assembly length (% arm length)	468 Mb (123%)	445 Mb (117%)	362 Mb (95%)	359 Mb (94%)
Contig N50	308 kb	424 kb	528 kb	555 kb
Contig L50	418	294	205	197

split. The final physical map assembly consisted of 895 contigs with average size of 401 kb, N50 of 555 kb and L50 of 197, representing approximately 359 Mb (94%) of the estimated chromosome-arm length (Table 1).

Physical map anchoring and inter-map comparison

Several approaches were combined to order 7DS physical-map contigs on the chromosome. First, 3-dimensional (3D) BAC pools were screened using PCR with 35 genetically mapped markers from GrainGenes database. Of these (Suppl. Table A), 21 were unambiguously assigned a distinct genetic-map position. The remainder were not anchored because they provided PCR products from two or more loci on 7DS. The second approach made use of integration of the 7DS physical map with the map of *Ae. tauschii*, which had been anchored to a high-density genetic SNP map [31]. The *Ae. tauschii* map of chromosome 7D spans 228.104 cM in total and comprises 1159 markers mapped to 488 distinct positions. The integration of 7DS and *Ae. tauschii* BAC clones in one physical map made it possible to delimit the short-arm part of the genetic map, which spans from 0 to 114.563 cM, and to identify 635 SNP markers in 7DS, which were mapped to 285 distinct positions. Out of the 635 markers, 582 could be tentatively assigned to 250 contigs of the 7DS physical map based on the graphical view in FPC (Fig. 1). Verification of this anchoring approach involved PCR screening of the 7DS-specific BAC library with bread wheat STS markers derived from their *Ae. tauschii* orthologs [29] and *in silico* anchoring based on MTP sequence data. The expected positions were confirmed for all STS markers and 460 out of 582 *Ae. tauschii* SNP markers. A failure to confirm additional *Ae. tauschii* markers was mainly due to an inability to deconvolute BAC sequence data or to find a significant sequence match for those markers. *In silico* anchoring positioned 621 SNP markers from the CsxRe genetic map, 147 SNP markers from the CS RH map and 464 DArTseq markers from the CS consensus map into contigs of the 7DS physical map. These efforts enabled anchoring 460 contigs of the 7DS physical map with 1713 markers (Table 2). The anchoring enabled delimiting the 7DS-specific part in the applied maps as follows: 0–117.597 cM (out of total 226.128 cM for 7D), 0–190.086 cM (out of total 371.634 cM) and 0–516.3 cR (out of total 758.7 cR) for CsxRe, CS consensus and CS RH maps, respectively.

In a previous study, the 7DS BNG map was demonstrated to be a useful tool for validating physical map assembly, scaffolding physical-map contigs, and integrating multiple maps [32]. Furthermore, the BNG map joined here an additional 45 marker-free 7DS contigs to the marker-anchored ones, thus increasing the number of anchored contigs to a total of 506 (Table 2). Combining all approaches, we were able to anchor 263.6 Mb of the 7DS physical map, corresponding to 73.4% of the assembly length.

Comparison of the three genetic maps integrated through the 7DS physical map showed a high collinearity along the 7DS arm (Fig. 2**A**) except for the subtelomeric region, where the CS consensus map showed a collinearity disruption with respect to *Ae. tauschii* and CsxRe maps (Fig. 2**B**). This disruption is roughly delimited by map interval 0–40.87 cM in the CS consensus map, corresponding to 0–31.66 cM and 0–12.994 cM in *Ae. tauschii* and CsxRe maps, respectively (Fig. 2B). The RH map showed overall collinearity but did not have sufficient resolution in the terminal region to resolve the inversion.

Table 2
BAC contig anchoring summary.

Anchoring method		No. anchored markers	No. anchored contigs
Pool screening[a]		21	21
Integration-based prediction	- SNP – *Ae. tauschii* map[b]	582	274
In silico anchoring	- SNP – *Ae. tauschii* map[b]	460	251
	- SNP – CsxRe map[c]	621	254
	- DArTseq – CS consensus map[d]	464	249
	- SNP – CS RH map[e]	147	98
BNG map[f]		NA	45
Overall		1713*	506

Map references.
* only sequence-validated (*in silico* anchored) *Ae. tauschii* markers considered.
[a] GrainGenes (Suppl. Table A).
[b] [31].
[c] [37].
[d] Suppl. Table B.
[e] [38].
[f] [32].

Ordering the centromeric region

Ordering physical map and sequence contigs in pericentromeric regions poses a major challenge in genomic projects due to suppressed recombination. Here we applied an integrated multiple-map approach to order physical map contigs as well as genetic markers around the genetic centromere of the 7D chromosome. It was delimited by a cluster of 88 markers co-segregating at 114.563 cM in the genetic map of *Ae. tauschii*. 85 markers from the cluster were positioned in 35 contigs from *Ae. tauschii* physical-map spanning over 77,090.5 kb in *Ae. tauschii* 7D. Using the physical map co-assembly, we identified 95 7DS contigs embedded in *Ae. tauschii* supercontigs associated with the centromeric marker cluster (Suppl. Table C). An additional five 7DS contigs were assigned to the cluster through BNG map-based scaffolding. A total of 100 7DS physical-map contigs spanned over 59,581 kb in Chinese Spring 7DS, corresponding to 16.6% of the total 7DS assembly length. No 7DS contigs were anchored beyond the 114.563 cM position, confirming affiliation of the cluster to the centromere. Seven out of 35 centromeric *Ae. tauschii* contigs did not comprise 7DS BAC clones and were proposed to belong to the 7DL chromosome arm (Suppl. Table C). The 7DL location was confirmed for all of them by finding matches between markers comprised in these contigs and the 7DL survey sequence [40]. Thus, a total of 15 *Ae. tauschii* markers from the centromeric cluster were assigned to the 7D long arm. The BNG map-based scaffolding also proposed joining an additional three marker-free *Ae. tauschii* contigs to the genetic centromere (Suppl. Table D), expanding it to 82,522.5 kb (including the 7DL part).

Applying all resources, we attempted to resolve the order and orientation of physical map contigs in this region. Besides the unambiguously assigned *Ae. tauschii* markers from the centromeric cluster, the 7DS centromeric contigs comprised 18 DArTseq markers mapping

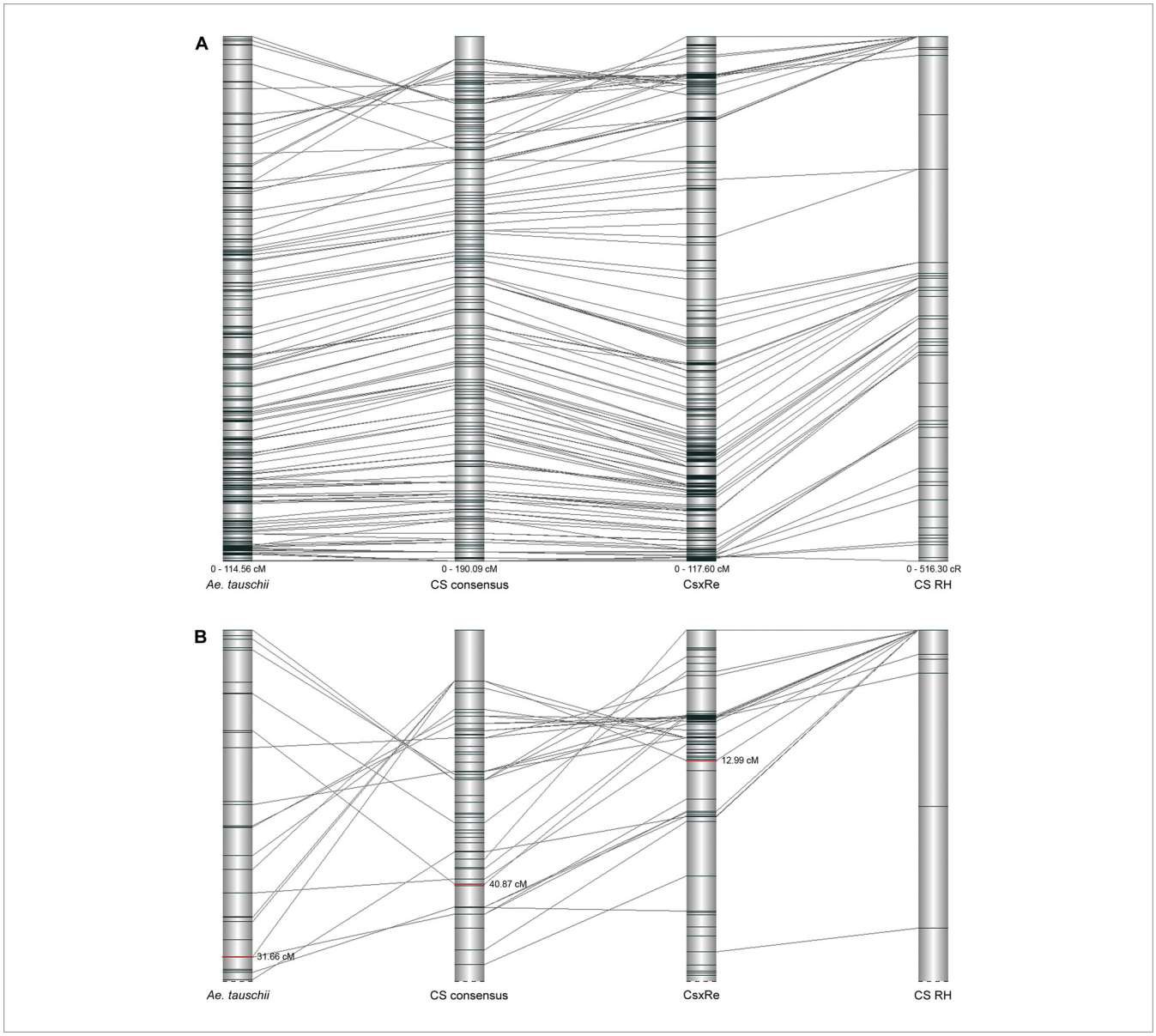


Fig. 2. Comparison of 7DS genetic and RH maps through 7DS physical map. Pairwise comparison of genetic/RH maps was performed after integrating particular marker datasets in contigs of the physical map. (**A**) Comparison of *Ae. tauschii* SNP map, CS consensus DArTseq map, CsxRe SNP map and CS RH map visualized by Strudel software. (**B**) Detail of collinearity disruption in CS consensus DArTseq map.

Table 3
Physical map of the genetic centromere (100 7DS physical-map contigs associated with *Ae. tauschii* centromeric marker cluster).

	No. 7DS contigs anchored	Total anchored length	No. scaffolds/ map positions	No. contig joins
Ae. tauschii co-assembly	95	55,481 kb	28	75
BNG map	66	49,755 kb	54	22
CsxRe map[a]	22	19,222.4 kb	5	NA
CS consensus map[a]	12	10,916 kb	4	NA
Combined[b]	**100**	**59,581 kb**	**16**	**86**

[a] Excluding markers mapping outside the most proximal map positions.
[b] *Ae. tauschii* co-assembly plus BNG map.

to five positions of the CS consensus map and 28 markers assigned to six positions of the CsxRe map (Suppl. Table C). Seventeen of the DArTseq markers mapped to 185.522–190.086 cM, which is the most proximal interval on the CS consensus map of 7DS. DArTseq marker 7D_3222174 assigned to 7DS ctg833 was allocated to 175.571 cM, suggesting a more distal position of the contig. Excluding this marker, we could anchor twelve of 7DS centromeric contigs to four positions of the CS consensus map (Table 3) and propose their tentative order. Out of 28 SNP markers from the CsxRe map, 26 mapped to interval 117.187-117.588, neighbouring the last 7DS position identified in the CsxRe map (117.597). While these 26 markers mapped to a total of five positions, they did not allow unambiguous ordering of 22 contigs containing them (Suppl. Table C). The remaining two SNP markers allocated in 7DS ctg3991 mapped to position 5.79 cM, which indicated a subtelomeric rather than centromeric location of the contig. None of the 7DS contigs associated with the centromeric marker cluster contained a SNP marker from the CS RH map and thus this resource could not be used for ordering contigs in the centromere.

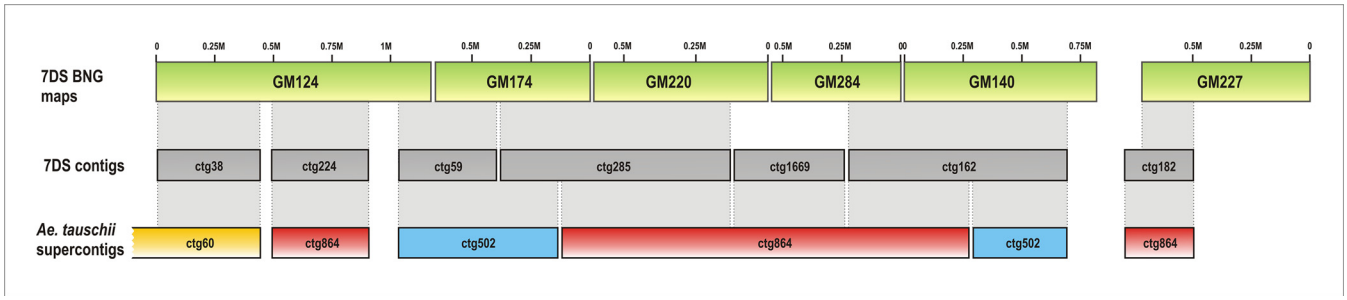


Fig. 3. Rearrangement in the pericentromeric region of Chinese Spring 7DS. 7DS contigs (grey boxes) assigned to *Ae. tauschii* supercontigs 60, 864 and 502 (yellow, red and blue boxes) by co-assembly were aligned through sequences of their BAC clones to 7DS BNG maps (green boxes), which indicated their different arrangement in 7DS of bread wheat compared to its ancestor. 7DS ctg1669 could not be aligned to any BNG map due to the lack of significant sequence match.

The co-assembly with *Ae. tauschii* aligned 95 7DS centromeric contigs with the *Ae. tauschii* genome, thus proposing 75 joins between separated 7DS contigs and combining the contigs in 28 physical-map scaffolds, corresponding to individual *Ae. tauschii* supercontigs (Table 3, Suppl. Table C). In parallel, projection of MTP sequences from centromeric contigs on 7DS BNG map aligned 66 contigs to particular genome maps, which proposed 22 contig joins. As a result, the centromeric contigs were ordered in 54 BNG map-based scaffolds. The 7DS BNG map confirmed the majority of scaffolds proposed by *Ae. tauschii* co-assembly and suggested 13 joins of *Ae. tauschii* contigs or their parts. Another two joins were indicated by 7DS contigs, which were split in two *Ae. tauschii* supercontigs each (Suppl. Table D). Combining both scaffolding approaches, we were able to organize the 7DS region corresponding to the *Ae. tauschii* centromeric marker cluster, spanning over 59,581 kb and comprising 100 7DS physical-map contigs, into 16 superscaffolds. A region delimited by *Ae. tauschii* supercontigs 864 and 502 appears to be rearranged in bread wheat with respect to *Ae. tauschii* since parts of the two *Ae. tauschii* contigs are interspersed in the CS 7DS, as proved by the BNG map (Fig. 3, Suppl. Table D).

Data accessibility

The 7DS physical-map assembly can be accessed through the GBrowse interface at https://urgi.versailles.inra.fr/gb2/gbrowse/wheat_phys_7DS_v2. The 7DS – *Ae. tauschii* co-assembly and 7DS BNG map are available on request. The BNG map is also deposited at the URGI repository https://urgi.versailles.inra.fr/download/iwgsc/IWGSC_RefSeq_Annotations/v1.0/iwgsc_refseqv1.0_optical_maps_group7.zip. The 7DS MTP scaffolds have been deposited at DDBJ/ENA/GenBank under the accession No. PKRY00000000 (BioSample SAMN07709812). The version described in this paper is version PKRY01000000 and can be assigned to particular clones or BAC pools using Suppl. File B.

Discussion

Physical map assembly and anchoring

The construction of a 7DS physical map involved the most widely used methods, such as contig fingerprinting by SNaPShot-based HICF technology and contig building by FPC. Alternative approaches, used in other wheat physical-map projects, included whole-genome profiling (WGP) to generate fingerprints of BAC clones [16,17] and application of LTC software for contig assembly. WGP was not compatible with our approach, which employed co-assembly with previously fingerprinted *Ae. tauschii* clones, since the co-assembly was conditioned by identical processing of clones from CS 7DS and *Ae. tauschii*. This also determined the use of FPC as the primary assembly software. LTC was shown to outperform FPC in terms of contig number and length (N50 value) in several wheat chromosomal projects, but the LTC assemblies generally covered smaller portions of the estimated chromosome arm lengths [12,13,15,19]. In the present work, LTC was only used for contig validation. Despite relying on FPC, the parameters of our 7DS assembly were comparable or superior to those obtained for other chromosomes. In our assembly, we considered all contigs > 2 clones, which were subject to MTP selection, whereas other physical maps excluded contigs with less than five or six clones, mainly due to difficulty in anchoring them. BNG map applied in our project resolved this problem as it reliably anchors contigs with only two sequenced clones [32]. Thanks to the inclusion of the short contigs, we achieved one of the highest chromosome-arm coverages (94%) among all wheat physical maps published to date. In comparison with other physical maps assembled by FPC, our full assembly's N50 value (555 kb) is higher than that obtained for 3DS (445 kb) [20], 1AL (460 kb) [13] or 5A (296 kb and 252 kb for 5AS and 5AL) [19] but lower than the N50 value for 1BS (1033 kb) [14] or 5DS (1141 kb) [18]. Excluding contigs < 6 clones, the N50 of the 7DS assembly increased up to 616 kb, which brings the value closer to that obtained for 3 B (783 kb) [10].

In order to position contigs of the 7DS physical map on the 7DS, we combined several approaches and genomic resources, including three genetic maps, one RH map, the 7DS BNG map and co-assembly with the *Ae. tauschii* physical map. Using 1713 markers with distinct genomic positions, we ordered 506 contigs, totaling 74% of the assembly, along the chromosome arm. The most efficient means of allocation was *in silico* anchoring, based on searching homology between marker and MTP-clone sequences, which positioned 460 contigs on the chromosome. Another 45 marker-free contigs were joined to those anchored by BNG map, which poses a highly reliable and recombination-independent tool to scaffold assemblies. The resulting high rate of assembly anchoring was achieved without employing high-throughput platforms such as NimbleGen UniGene microarray, used in other wheat projects [12–15]. Several projects [12,14,19,20] also exploited synteny-based approaches and anchored physical map assemblies to so called genome zippers [42], which predict a virtual gene order on the basis of synteny with three grass genomes (rice, brachypodium and sorghum). The limitation of this approach can be a relatively high proportion of genes (or non-recognized pseudogenes) in non-syntenic positions [12] and also low quality or different origin of genetic map used to build the zipper [18].

Ordering the centromeric region

Ordering physical-map contigs and sequences in non-recombining regions around genetic centromeres has been the most challenging part of genome-sequencing projects. In our study, the genetic centromere was delimited as a cluster of contigs associated with markers of the *Ae. tauschii* map co-segregating at the most proximal position of the 7DS arm. This corresponded to a cumulative length of 70,882.5 kb and 59,581 kb for 7DS-assigned *Ae. tauschii* and Chinese Spring contigs, respectively. These values are with a high probability underestimates of the real size of the non-recombining region since we considered only contigs anchored to *Ae. tauschii* markers and those added to them by

BNG map-based scaffolding. Contigs anchored to the most proximal positions of the other two linkage maps used in our study are candidates for expanding the region of the 7DS genetic centromere. The 59,581 kb region, representing 16.6% of the 7DS assembly length, is only a part of the difficult-to-order low-recombining region, which was reported to span over more than 40% of the chromosome length in wheat chromosomes [10,15].

As expected, integrating information from additional two genetic maps did not aid in resolving the order of contigs around the centromere (Table 3, Suppl. Table C). The wheat radiation hybrid map, not relying on genetic recombination and showing a high resolution in the pericentromeric region of 7D chromosome [38], appeared a promising genomic resource. Unfortunately, we failed to allocate SNP markers from the RH map to the 100 7DS centromeric contigs. This may reflect the fact that the iSelect 90 K SNP array [39] used to genotype the CS RH panel mostly comprises markers from genic/low-copy regions that are relatively sparsely represented in the centromeric part of RH maps [43]. Repeat junction markers (RJM) or ISBP markers, derived from junctions of transposable elements and dispersed throughout the genome, were shown to be a favorable complement to the gene-based markers in RH mapping [44]. In our physical map project, we tested 380 RJM markers intended for mapping on the Chinese Spring 7D RH panel (V. Tiwari, personal communication) and managed to align six of them to the centromeric contigs (data not shown), thus indicating a modest potential of the anticipated RH map for the centromeric region.

Alternative methods for ordering contigs/BAC clones in low-recombining regions include cytogenetic mapping [45], which is laborious and cannot be used in a high-throughput manner, or synteny-based anchoring to genes virtually ordered in genome zipper [42]. The synteny-based approach did not prove to be beneficial for centromeric regions because of low representation of genes [12], which are mostly in non-syntenic positions around the centromeres [44]. The most efficient approach to assembling non-recombining regions was recently reported by Mascher et al. [6] who applied an innovative method of chromosome conformation capture-based ordering to complete BAC-based assembly of barley. This approach is applicable only at final stages of sequencing projects, as its resolution in large genomes is relatively limited (1 Mb in the barley project) and thus is conditional on availability of large sequence scaffolds.

BNG map-based ordering applied in our study also relies on the availability of BAC clone sequences. However, as we demonstrated earlier [32], assemblies of 90 kb or even less are sufficient to align to the BNG map if used in the context of a physical map. Anchoring 100 centromeric contigs to BNG maps proposed 22 joins and split the region into 54 physical-map scaffolds. The potential of the BNG map could not be fully exploited in the pericentromere due to a high content of DNA repeats, which hampered sequence assembly and deconvolution and thus reduced the number of sequences that could be aligned to the BNG map. Overall, the best clue to ordering contigs in the genetic centromere was the co-assembly with the *Ae. tauschii* physical map, which indicated 75 contig joins and distributed the 7DS centromeric contigs into 28 scaffolds. Nevertheless, the *Ae. tauschii*-based ordering must be taken with caution as the fingerprint co-assembly was done with less stringent parameters and might have resulted in erroneous clone joins. Furthermore, structural variation between bread wheat and its ancestor should be considered. Nevertheless, the co-assembly with *Ae. tauschii* provided a valuable clue for ordering contigs of the CS 7DS physical map and this beneficial effect was reciprocal, since the co-assembly, combined with alignment of the 7DS BNG map, added three marker-free *Ae. tauschii* contigs to this region and provided a clue for scaffolding a large part of *Ae. tauschii* centromeric contigs. The CS 7DS – *Ae. tauschii* co-assembly was previously used to validate the *Ae. tauschii* physical map assembly [31] and here we show further extension of this approach, that can be used to support *Ae. tauschii* genome assembly in the centromeric region of 7DS. To summarize, by combining both physical-mapping approaches, we could organize the region of 59,581 kb, associated with the 7DS genetic centromere, into 16 validated physical map scaffolds, thus outperforming outcomes of other wheat physical map projects.

Comparison of bread wheat and Ae. tauschii

Ae.tauschii is a donor of the youngest component of the bread wheat genome, which implies relatively low diversity between the D-genome of bread wheat and its ancestor [1]. Several rearrangements on the genic level and a small duplication were indicated by physical mapping on chromosome 5DS [18]. Structural variability between wheat and its ancestor was also proposed in chromosome 2D [46]. Chromosome arm 7DS seems highly collinear in wheat cv. Chinese Spring and *Ae. tauschii*, as indicated by inter-map comparison (Fig. 2A). The disruption of collinearity at the distal end of 7DS, apparent from the comparison of *Ae. tauschii* and CS consensus map (Fig. 2B), was not confirmed by the CsxRe map (Suppl. Figure A) and is likely due to an error in building the map consensus or to a structural variability in one of consensus-map components. The analysis of the distribution of CS 7DS contigs in the co-assembled supercontigs indicated several regions of potential variability between bread wheat and *Ae. tauschii* at a BAC-contig level. Validation of supercontigs from the centromeric cluster by the BNG map confirmed local rearrangements in a region spanning over more than 5 Mb (Fig. 3, Suppl. Table D), which could not be detected by the genetic maps.

Conclusions

This work resulted in a GBrowse-accessible 7DS physical map, which will facilitate gene cloning and comparative genome analyses in wheat. The physical map interconnects a total of 1713 markers from one radiation-hybrid map and three genetic maps with ordered 7DS BAC clones. All clones are available from library depositories at IEB (http://olomouc.ueb.cas.cz/dna-libraries/cereals) or CNRGV (http://cnrgv.toulouse.inra.fr/en/library/wheat), thus providing an opportunity to access a region of interest with advanced sequencing technologies in a focused and affordable manner. Mate-pair Illumina assemblies of MTP BAC clones generated in this study are available from NCBI and we anticipate their utilization in the envisaged version 2 of the bread wheat reference sequence. The co-assembly of 7DS with *Ae. tauschii* indicated regions of potential structural variation between bread wheat and its ancestor, which can be validated using BNG map, as demonstrated for the pericentromeric region of 7DS.

Author contributions

Z.T. performed physical-map anchoring, final stage of map assembly and centromere analyses. She drafted the manuscript. M.C.L. carried out fingerprinting of 7DS-specific BAC library and generated *Ae. tauschii* – 7DS co-assembly. H.T. participated in the construction of physical map and MTP sequencing. P.Visendi and D.E. assembled MTP sequences. S.H., J. Batley, P.Vojta and M.H. were engaged in MTP sequencing. E.P. and A.K. provided CS consensus and CsxRe maps, respectively. M.A. was involved in map integration and comparison. J. Bartoš supervised 7DS physical map assembly. J.D. participated in designing the study and revised the manuscript. H.Š. designed and coordinated the study and completed the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank to Dr. Vijay Tiwari from the University of Maryland for providing prepublication access to Chinese Spring RH map and marker sequences. We acknowledge excellent technical assistance of Radka Tušková, Marie Seifertová and Helena Tvardíková. This work was supported by the Czech Science Foundation (award No. P501/12/

2554), and by the Ministry of Education, Youth and Sports of the Czech Republic (grant LO1204 from the National Program of Sustainability I). The research leading to these results has received funding from the French Government managed by the Research National Agency (ANR) under the Investment for the Future programme (BreedWheat project ANR-10-BTBR-03), from FranceAgriMer, French Funds to support Plant Breeding (FSOV) and from INRA.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.nbt.2018.03.003.

References

[1] Marcussen T, Sandve SR, Heier L, Spannagl M, Pfeifer M, Rogers J, et al. Ancient hybridizations among the ancestral genomes of bread wheat. Science 2014;345:1250092.
[2] Arumuganathan K, Earle ED. Nuclear DNA content of some important plant species. Plant Mol Biol Rep 1991;9(3):208–18.
[3] AGI. Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 2000;408:796–815.
[4] IRGSP – International Rice Genome Sequencing Project. The map-based sequence of the rice genome. Nature 2005;436:793–800.
[5] Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, et al. The B73 maize genome: complexity, diversity, and dynamics. Science 2009;326(5956):1112–5.
[6] Mascher M, Gundlach H, Himmelbach A, Beier S, Twardziok SO, Wicker T, et al. A chromosome conformation capture ordered sequence of the barley genome. Nature 2017;544:427–33.
[7] Doležel J, Kubaláková M, Paux E, Bartoš J, Feuillet C. Chromosome-based genomics in the cereals. Chromosome Res 2007;15:51–66.
[8] Feuillet C, Eversole K. Physical mapping of the wheat genome: a coordinated effort to lay the foundation for genome sequencing and develop tools for breeders. Isr J Plant Sci 2007;55:307–13.
[9] Šafář J, Bartoš J, Janda J, Bellec A, Kubaláková M, Valárik M, et al. Dissecting large and complex genomes: flow sorting and BAC cloning of individual chromosomes from bread wheat. Plant J 2004;39:960–8.
[10] Paux E, Sourdille P, Salse J, Saintenac C, Choulet F, Leroy P, et al. A physical map of the 1-gigabase bread wheat chromosome 3B. Science 2008;322:101–4.
[11] Šafář J, Šimková H, Kubaláková M, Čihalíková J, Suchánková P, Bartoš J, et al. Development of chromosome-specific BAC resources for genomics of bread wheat. Cytogenet Genome Res 2010;129:211–23.
[12] Breen J, Wicker T, Shatalina M, Frenkel Z, Bertin I, Philippe R, et al. A physical map of the short arm of wheat chromosome 1A. PLoS One 2013;8(11):e80272.
[13] Lucas SJ, Akpinar BA, Kantar M, Weinstein Z, Aydınoğlu F, Šafář J, 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.
[14] Raats D, Frenkel Z, Krugman T, Dodek I, Sela H, Šimková H, et al. The physical map of wheat chromosome 1BS provides insights into its gene space organization and evolution. Genome Biol 2013;14:R138.
[15] Philippe R, Paux E, Bertin I, Sourdille P, Choulet F, Laugier Ch, et al. A high density physical map of chromosome 1BL supports evolutionary studies, map-based clonig and sequencing in wheat. Genome Biol 2013;14:R64.
[16] Poursarebani N, Nussbaumer T, Šimková H, Šafář J, Witsenboer H, van Oeveren J, et al. Whole-genome profiling and shotgun sequencing delivers an anchored, gene-decorated, physical map assembly of bread wheat chromosome 6A. Plant J 2014;79:334–47.
[17] Kobayashi F, Wu J, Kanamori H, Tanaka T, Katagiri S, Karasawa W, et al. A high-resolution physical map integrating an anchored chromosome with the BAC physical maps of wheat chromosome 6B. BMC Genom 2015;16:595.
[18] Akpinar BA, Magni F, Yuce M, Lucas SJ, Šimková H, Šafář J, et al. The physical map of wheat chromosome 5DS revealed gene duplications and small rearrangements. BMC Genom 2015;16:453.
[19] Barabaschi D, Magni F, Volante A, Gadaleta A, Šimková H, Scalabrin S, et al. Physical mapping of bread wheat chromosome 5A: an integrated approach. Plant Genome 2015;8:1–24.
[20] Holušová K, Vrána J, Šafář J, Šimková H, Balcárková B, Frenkel Z, et al. Physical map of the short arm of bread wheat chromosome 3D. Plant Genome 2017;10:1–11.
[21] Krattinger S, Wicker T, Keller B. Map-based cloning of genes in *Triticeae* (wheat and barley). In: Feuillet C, Muehlbauer GJ, editors. Genetics and genomics in the Triticeae. New York: Springer; 2009. p. 337–58.
[22] Choulet F, Alberti A, Theil S, Glover N, Barbe V, Daron J, et al. Structural and functional partitioning of bread wheat chromosome 3B. Science 2014;345(6194):1249721.
[23] Avni R, Nave M, Barad O, Naruch K, Twardziok SO, Gundlach H, et al. Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. Science 2017;357(6346):93–7.
[24] Schneider LM, Adamski NM, Christensen CE, Stuart DB, Vatrin S, Hansson M, et al. The *Cer-cqu* gene cluster determines three key players in a β-diketone synthase polyketide pathway synthesizing aliphatics in epicutilar waxes. J Exp Bot 2016;67(9):2715–30.
[25] Roselli S, Olry A, Vautrin S, Coriton O, Ritchie D, Galati G, et al. A bacterial artificial chromosome (BAC) genomic approach reveals partial clustering of the furanocoumarin pathway genes in parsnip. Plant J 2017;89(6):1119–32.
[26] Röder MS, Huang XQ, Börner A. Fine mapping of the region on wheat chromosome 7D controlling grain weight. Funct Integr Genomics 2008;8:79–86.
[27] Hirao K, Nishijima R, Sakaguchi K, Takumi S. Fine mapping of Hch1, the causal D-genome gene for hybrid chlorosis in interspecific crosses between tetraploid wheat and *Aegilops tauschii*. Genes Genet Syst 2015;90:283–91.
[28] Reddy INBL, Chandrasekhar K, Zewdu Y, Dinoor A, Keller B, Ben-David R. Identification and genetic mapping of PmAF7DS a powdery mildew resistance gene in bread wheat (*Triticum aestivum* L.). Theor Appl Genet 2016;129(6):1127–37.
[29] Staňková H, Valárik M, Lapitan N, Berkman PJ, Batley J, Edwards D, et al. Chromosomal genomics facilitates fine mapping of a Russian wheat aphid resistance gene. Theor Appl Genet 2015;128(7):1373–83.
[30] Šimková H, Šafář J, Kubaláková M, Suchánková P, Čihalíková J, Robert-Quatre H, et al. BAC libraries from wheat chromosome 7D: efficient tool for positional cloning of aphid resistance genes. J Biomed Biotechnol 2011:302543.
[31] Luo MC, Gu YQ, You FM, Deal KR, Ma Y, Hu Y, et al. A 4-gigabase physical map unlocks the structure and evolution of the complex genome of *Aegilops tauschii*, the wheat D-genome progenitor. PNAS 2013;110:7940–5.
[32] Staňková H, Hastie A, Chan S, Vrána J, Tulpová Z, Kubaláková M, et al. BioNano genome mapping of individual chromosomes supports physical mapping and sequence assembling in complex plant genomes. Plant Biotechnol J 2016;14:1523–31.
[33] Luo MC, Thomas C, You FM, Hsiao J, Ouyang S, Buel CR, et al. High-throughput fingerprinting of bacterial artificial chromosomes using the SNaPshot labelling kit and sizing of restriction fragments by capillary electrophoresis. Genomics 2003;82:378–89.
[34] Soderlund C, Humphray S, Dunham A, French L. Contig built with fingerprints, markers and FPC V4.7. Genome Res 2000;10:1772–87.
[35] 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 Bioinf 2010;11:584–601.
[36] Visendi P, Berkman PJ, Hayashi S, Golicz AA, Bayer PE, Ruperao P, et al. An efficient approach to BAC based assembly of complex genomes. Plant Methods 2016;12:2.
[37] Rimbert H, Darrier B, Navarro J, Kitt J, Choulet F, Leveugle M, et al. High throughput SNP discovery and genotyping in hexaploid wheat. PLoS One 2017;13(1):e0186329.
[38] Tiwari VK, Heesacker A, Riera-Lizarazu O, Gunn H, Wang SC, Wang Y, et al. A whole-genome radiation hybrid mapping resource of hexaploid wheat. Plant J 2016;86(2):195–207.
[39] Wang S, Wong D, Forrest K, Allen A, Chao S, Huang BE, et al. Characterization of polyploid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array. Plant Biotechnol 2014;12(6):787–96.
[40] IWGSC. A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. Science 2014;345:1251788.
[41] Bayer I, Milne I, Stephen G, Shaw P, Cardle L, Wright F, et al. Comparative visualization of genetic and physical maps with Strudel. Bioinformatics 2011;27(9):1307–8.
[42] Mayer KF, Taudien S, Martis M, Šimková H, Suchánková P, Gundlach H, et al. Gene content and virtual gene order of barely chromosome 1H. Plant Physiol 2009;151:496–505.
[43] Balcárková B, Frenkel Z, Škopová M, Abrouk M, Kumar A, Chao SM, et al. A high resolution radiation hybrid map of wheat chromosome 4A. Plant Sci 2017;7:2063.
[44] Kumar A, Simons K, Iqbal MJ, Michalak de Jimenéz M, Bassi FM, Ghavami F, et al. Physical mapping resources for large plant genomes: radiation hybrids for wheat D-genome progenitor *Aegilops tauschii*. BMC Genom 2012;13:597.
[45] Aliyeva-Schnorr L, Beier S, Karafiatova M, Schmutzer T, Scholz U, Doležel J, et al. Cytogenetic mapping with centromeric bacterial artificial chromosomes contigs shows that this recombination-poor region comprises than half of barley chromosome 3H. Plant J 2015;84(2):285–394.
[46] Thind AK, Wicker T, Šimková H, Fossati D, Moullet O, Brabant C, et al. Rapid cloning of genes in hexaploid wheat using cultivar-specific long-range chromosome assembly. Nat Biotechnol 2017;35:793–6.

Glossary

BAC-based physical map: Constructed from overlapping clones of large-insert libraries, maintained in bacterial-artificial-chromosome (BAC) vector. The overlaps are identified by pairwise comparison of fingerprints generated from particular BAC clones.

Bionano genome (BNG) map: Also called optical map. Physical map of short sequence motifs recognition sites of a nicking enzyme – along hundreds to thousands kilobases-long stretches of DNA. It is created by imaging labelled DNA molecules on nanochannel arrays.

APPENDIX II

Accessing a Russian wheat aphid resistance gene in bread wheat using long-read technologies

Tulpová Z, Toegelová H, Lapitan NLV, Peairs F, Macas J, Novák P, Lukaszewski A, Kopecký D, Mazáčová M, Vrána J, Holušová K, Leroy P, Doležel J, Šimková H

Plant Genome, 2019

doi: 10.3835/plantgenome2018.09.0065

IF: 2.923

Publisher: AGRONOMY; Journal: TPG:The Plant Genome; Copyright: Will notify...
Volume: Will notify...; Issue: Will notify...; Manuscript: tpg-2018-09-0065; DOI: ; PII: <txtPII>
TOC Head: ; Section Head: ; Article Type: Original Research

Accessing a Russian wheat aphid resistance gene

Core ideas

An accurate sequence of the *Dn2401* region was generated from Illumina and nanopore reads.

Structural variation in the region was analyzed by optical mapping and resequencing.

New markers located *Dn2401* within a 133.2-kb interval with six high-confidence genes.

Epoxide hydrolase 2 was identified as the most likely resistance gene candidate.

Accessing a Russian Wheat Aphid Resistance Gene in Bread Wheat by Long-Read Technologies

Zuzana Tulpová, Helena Toegelová, Nora L. V. Lapitan, Frank B. Peairs, Jiří Macas, Petr Novák, Adam J. Lukaszewski, David Kopecký, Mira Mazáčová, Jan Vrána, Kateřina Holušová, Philippe Leroy, Jaroslav Doležel, and Hana Šimková*

Z. Tulpová, H. Toegelová, D. Kopecký, M. Mazáčová, J. Vrána, K. Holušová, J. Doležel, H. Šimková, Inst. of Experimental Botany, Centre of the Region Haná for Biotechnological and Agricultural Research, Šlechtitelů 31, 78371 Olomouc, Czech Republic; N.L.V. Lapitan, Bureau for Food Security, United States Agency for International Development, Washington, DC 20004; F.B. Peairs, Dep. of Bioagricultural Sciences and Pest Management, Colorado State Univ., Fort Collins, CO 80523; J. Macas, P. Novák, Biology Centre, Czech Academy of Sciences, Institute of Plant Molecular Biology, Branišovská 31, CZ-37005 České Budějovice, Czech Republic; A.J. Lukaszewski, Dep. of Botany and Plant Sciences, Univ. of California, Riverside, CA 92521; P. Leroy, Genetics, Diversity and Ecophysiology of Cereals, INRA, Univ. Clermont Auvergne, 5 chemin de Beaulieu, 63039 Clermont-Ferrand, France. Received 10 Sept. 2018. Accepted 10 Dec. 2018. *Corresponding author (simkovah@ueb.cas.cz).

ABSTRACT

Russian wheat aphid (RWA) (*Diuraphis noxia* Kurdjumov) is a serious invasive pest of small-grain cereals and many grass species. An efficient strategy to defy aphid attacks is to identify sources of natural resistance and transfer resistance genes into susceptible crop cultivars. Revealing the genes helps understand plant defense mechanisms and engineer plants with durable resistance to the pest. To date, more than 15 RWA resistance genes have been identified in wheat (*Triticum aestivum* L.) but none of them has been cloned. Previously, we genetically mapped the RWA resistance gene *Dn2401* into an interval of 0.83 cM on the short arm of chromosome 7D and spanned it with five bacterial artificial chromosome (BAC) clones. Here, we used a targeted strategy combining traditional approaches toward gene cloning (genetic mapping and sequencing of BAC clones) with novel technologies, including optical mapping and long-read nanopore sequencing. The latter, with reads spanning the entire length of a BAC insert, enabled us to assemble the whole region, a task that was not achievable with short reads. Long-read optical mapping validated the DNA sequence in the interval and revealed a difference in the locus organization between resistant and susceptible genotypes. Complete and accurate sequencing of the *Dn2401* region facilitated the identification of new markers and precise annotation of the interval, revealing six high-confidence genes. Identification of *Epoxide hydrolase 2* as the most likely *Dn2401* candidate opens an avenue for its validation through functional genomics approaches.

Abbreviations: 2OG-Fe(II) oxygenase, 2-oxoglutarate and Fe(II)-dependent oxygenase; 7DS, short arm of wheat chromosomes 7D; BAC, bacterial artificial chromosome; BNG map, Bionano genome map; CS, cultivar Chinese Spring; EH, epoxide hydrolase; *EH2*, *epoxide hydrolase 2*; IWGSC, International Wheat Genome Sequencing Consortium; PCR, polymerase chain reaction; RWA, Russian wheat aphid; SNP, single nucleotide polymorphism; UTR, untranslated region

Publisher: AGRONOMY; Journal: TPG:The Plant Genome; Copyright: Will notify...
Volume: Will notify...; Issue: Will notify...; Manuscript: tpg-2018-09-0065; DOI: ; PII: <txtPII>
TOC Head: ; Section Head: ; Article Type: Original Research

Russian wheat aphid (RWA) was first reported in 1978 as a pest of small-grain cereals in South Africa (Walters et al., 1980). Since then, the aphid has spread around the world, most recently to Australia, and has become a serious invasive pest not only of wheat and barley (*Hordeum vulgare* L.) but also of many other plants from 43 genera, including over 140 species of cultivated and wild grasses (Yazdani et al., 2017). During feeding, aphids remove plant photoassimilates, which results in chlorosis, longitudinal streaking around the main leaf vein, head trapping, a substantial reduction in biomass, and, in severe cases, plant death (Burd and Burton, 1992). Another undesirable effect accompanying aphid feeding is leaf rolling, which that serves as a shelter against natural predators of the aphid and also against insecticide spraying. Thus the most efficient strategy to defy RWA attacks lies in identifying sources of natural resistance and introducing them into susceptible cultivars.

Plant defenses against insect damage include three mechanisms that are frequently combined: antibiosis (mechanical and chemical defensive factors impacting insect biology, including fertility), antixenosis (nonpreference of a particular plant as a host), and tolerance (the ability of a plant to withstand insect damage) (Botha et al., 2005). To date, more than 15 RWA resistance genes, mostly underlying tolerance and antibiosis, have been described in wheat. Many of these originated from *Aegilops tauschii* Coss. as inferred from their map location in the D genome (Du Toit, 1987, 1989; Du Toit et al., 1995; Fazel-Najafabadi et al., 2015; Liu et al., 2001; Ma et al., 1998; Miller et al., 2001; Peng et al., 2007; Valdez et al., 2012; Voothuluru et al., 2006). At least one resistance gene, *Dn7*, originating from rye (*Secale cereale* L.) has been reported (Anderson et al., 2003; Marais et al., 1994; Lapitan et al., 2007). New highly virulent RWA biotypes have rendered previously resistant cultivars susceptible, which has stimulated an intensive search for novel sources of resistance as well as efforts to unravel the mechanisms underlying the trait. Although several components of the plant defense pathways have been proposed (Anderson et al., 2014; Van Eck et al., 2014), to date, no aphid resistance gene has been cloned in wheat. Nine of the hitherto mapped RWA resistance genes, namely *Dn1*, *Dn2*, *Dn5* (Du Toit, 1987), *Dn6* (Saidi and Quick, 1996), *Dn8*, (Liu et al., 2001), *Dnx* (Harvey and Martin, 1990), *Dn2401* (Voothuluru et al., 2006; Fazel-Najafabadi et al. 2015), *Dn626580* (Valdez et al., 2012), and *Dn100695* (Aykut Tonk et al., 2016) are located on the short arm of wheat chromosome 7D (7DS). Except for *Dn8*, which is located at the terminal part of the arm, the remaining genes were mapped to the interstitial part of 7DS and most of them were found linked to marker *gwm111* (Liu et al. 2002; Fazel-Najafabadi et al. 2015). It has not yet been resolved if these genes are allelic or tightly linked.

Here, we focused on the *Dn2401* gene identified in line ‘CI2401’, which originated from Tajikistan. Previously, Staňková et al. (2015) mapped the *Dn2401* gene into a 0.83-cM interval on chromosome arm 7DS, delimited by the markers *Xowm705* and *Xowm711*. The interval was spanned by five overlapping BAC clones from a 7DS-specific BAC library of the cultivar Chinese Spring (Šimková et al., 2011). Chinese Spring, a reference genome of bread wheat, is susceptible to RWA (Peng et al., 2009). In the current work, we sequenced, assembled, and annotated the five CS BAC clones, proposing a list of genes located in the interval. The study of Staňková et al., (2015) pointed to a striking decrease (a minimum of eightfold) in the physical/genetic map distance ratio in the *Dn2401* interval compared with the neighboring regions. We speculated that this decrease is likely to be caused by a higher recombination rate in the region. However, an alternative explanation might be a structural variation between CS, in which we measured the physical distance, and CI2401 and ‘Glupro’, which were used to construct the genetic map. In the case of the larger deletion in CS, the list of candidate genes

Publisher: AGRONOMY; Journal: TPG:The Plant Genome; Copyright: Will notify...
Volume: Will notify...; Issue: Will notify...; Manuscript: tpg-2018-09-0065; DOI: ; PII: <txtPII>
TOC Head: ; Section Head: ; Article Type: Original Research

deduced from the CS reference sequence might not be complete. Considering the size of the region (~300 kb) and a highly repetitive nature of the wheat genome, it was unrealistic to analyze the entire interval in CI2401 by resequencing. Thus an easier approach to a fast and affordable comparative analysis of the region was chosen: a long-read optical mapping on the nanochannel platform of Bionano Genomics (San Diego, CA) (Lam et al. 2012), which generates maps of short sequence motif hundreds to thousands of kb long that can be used to support or validate sequence scaffolding and to perform large-scale comparative analyses. Thanks to a previously constructed Bionano genome (BNG) map of the 7DS arm from CS (Staňková et al., 2016), a newly prepared BNG map of 7DS from CI2401 facilitated a straightforward comparison of the entire region between the two accessions.

An accurate and complete sequence of the interval delimited by the flanking markers is a prerequisite for successful gene cloning. We approached the region through low complexity sequencing of individual BAC clones on the short-read Illumina platform, which did not produce a continuous sequence of the entire interval. Recently, two advanced whole-genome assemblies of bread wheat were released: Triticum 3.1 (Zimin et al., 2017), combining short Illumina)San Diego, CA) and long Pacific Biosciences (Menlo Park, CA) reads with the fully annotated International Wheat Genome Sequencing Consortium (IWGSC) RefSeq v1.0 (International Wheat Genome Sequencing Consortium, 2018), which was based on short reads only and coupled the whole-genome assembly with assemblies of physical map-ordered BAC clones. Triticum 3.1 and IWGSC RefSeq v1.0 cover 97 and 92% of the estimated wheat genome size of 15.76 Gb (International Wheat Genome Sequencing Consortium, 2018), respectively. Such reference genome assemblies promise to be an excellent tool for accelerating gene cloning. Unfortunately, not even these assemblies allowed us to close the gap in the BAC assembly of the region of interest. To finally resolve the *Dn2401* region, we adopted an alternative long-read technology: nanopore sequencing on the MinION platform of Oxford Nanopore Technologies, which produces reads reaching tens or even hundreds of kb (Deschamps et al., 2018), thus having the potential to span long transposable elements– the main obstacle in assembling complex genomes. Coupling nanopore and short reads of the BAC clones allowed us to complete the sequence of the *Dn2401* region, which enabled precise annotation of the region, identification of new markers, and identification and resequencing of candidate genes.

MATERIALS AND METHODS

Plant Material

Genetic mapping was performed in an F_2 population derived from a cross between a RWA-resistant winter wheat line CI2401 and the susceptible cultivar Glupro. The current population (designated CI2401–Glupro F_2) consisted of 333 plants and combined 184 individuals from a previous study (CI2401–Glupro F_2–1)(Staňková et al., 2015) with a new set of 149 F_2 plants (CI2401–Glupro F_2–2) being included in this work.

The short arms of homeologous group 7 chromosomes of CS were flow-sorted from the respective double ditelosomic lines. Seed samples of these lines were kindly provided by Prof. Bikram Gill (Kansas State Univ., Manhattan, KS). To enable flow-sorting of the 7DS arm from line CI2401, a 7DS ditelosomic line was generated by misdivision of univalent $7D_{CI2401}$ in a hybrid between CS nullisomic 7D (CS N7D) × CI2401. Among 280 progeny screened, three

Publisher: AGRONOMY; Journal: TPG:The Plant Genome; Copyright: Will notify...
Volume: Will notify...; Issue: Will notify...; Manuscript: tpg-2018-09-0065; DOI: ; PII: <txtPII>
TOC Head: ; Section Head: ; Article Type: Original Research

telocentric $7DS_{CI2401}$ chromosomes were identified and the respective plants were self-pollinated. Ditelosomic 7DS were selected from their progeny and propagated.

BAC Clone Sequencing and Assembly

Five overlapping BAC clones from CS 7DS arm-specific BAC library (Šimková et al., 2011) spanning the *Dn2401* region (*TaaCsp7DS088E19*, *TaaCsp7DS010E01*, *TaaCsp7DS112N11*, *TaaCsp7DS044E09*, and *TaaCsp7DS086H04*) were purified with a NucleoSpin 96 Flash kit (Macherey-Nagel, Düren, Germany) and individually pair-end sequenced on the Illumina MiSeq platform. The 2 × 250 bp reads obtained were assembled with RAY (Boisvert et al., 2010) separately for each clone. The resulting 90 contigs, which were mostly doubled because of BAC overlaps, were manually merged in Geneious version 7.1.2 (http://www.geneious.com, accessed 14 Feb. 2019)(Kearse et al., 2012), with support from BLAST and the 7DS BNG map of CS (Staňková et al., 2016). A larger gap in the clone *TaaCsp7DS088E19* was closed with scaffolds assembled by SASSY from the 2 × 150 bp reads obtained from pooled 7DS BAC clones (Tulpová et al., 2019). The SASSY scaffolds did not close the 14.4 kb gap identified in *TaaCsp7DS086H04*.

Additional data for the BAC clone TaaCsp7DS086H04 were obtained by long-read nanopore sequencing. To maximize the read length, BAC DNA was extracted manually by the alkaline lysis method followed by phenol-chloroform extraction and ethanol precipitation. As a final step, the DNA was purified by a 5-min incubation with 1:1 AMPure XP beads (Beckman Coulter, Miami, FL) and eluted into 30 μL of 10 mM tris(hydroxymethyl)-aminomethane (pH 8.5). The sequencing library was prepared using the Rapid Sequencing Kit (SQK-RAD003, Oxford Nanopore Technologies, Oxford, UK) and the sample was run on the MinION platform (Oxford Nanopore Technologies). Size-selected reads (352 in total) ranging from 10 to 131 kb and representing 55-fold coverage of the BAC sequence were combined with the 2 × 250 bp Illumina reads for hybrid assembly with MaSuRCA (Zimin et al., 2013). The resulting assembly was manually verified by comparison with the longest nanopore reads spanning the entire (or a majority of) the insert length to confirm correct assembly of structurally complex repeated regions.

Sequence Analysis

Annotation of the *Dn2401* region was performed with the TriAnnot (online version 4.3.1) pipeline (Leroy et al., 2012) and predictions were manually curated with GenomeView 2250 software (genomeview.org, accessed 14 Feb. 2019)(Abeel et al., 2012). This manual annotation was compared with the IWGSC RefSeq Annotation version 1.0 (International Wheat Genome Sequencing Consortium, 2018), once it was released (Table 1). To resequence candidate genes from CI2401 and Glupro, primers were designed to cover all of the coding sequences as well as the upstream and downstream regulatory regions. Primers, polymerase chain reaction (PCR) conditions and amplicon sizes are provided in the Supplemental Table S3. For templates with a high GC content, 5% dimethyl sulfoxide was added into a PCR reaction. Amplification products were cleaned up and Sanger-sequenced on an ABI 3730xl DNA analyzer (Applied Biosystems, Waltham, MA). Polymorphisms between parental lines were identified by sequence alignment in Geneious version 7.1.2.

Publisher: AGRONOMY; Journal: TPG:The Plant Genome; Copyright: Will notify...
Volume: Will notify...; Issue: Will notify...; Manuscript: tpg-2018-09-0065; DOI: ; PII: <txtPII>
TOC Head: ; Section Head: ; Article Type: Original Research

Bionano Genome Map and Analysis of Structural Variation

To investigate the overall sequence organization of the *Dn2401* region in CI2401 and to detect a possible structural variation between CI2401 and the reference genome of CS, we constructed a new BNG map of the chromosome arm 7DS originating from CI2401 and compared it with the previously assembled 7DS BNG map of CS (Staňková et al., 2016). The 7DS chromosome arm was flow-sorted from a 7DS ditelosomic line of CI2401 and the BNG map was constructed as described in Staňková et al. (2016) with minor modifications. HMW DNA molecules labeled at Nt.BspQI sites (GCTCTTC) with Alexa546-dUTP fluorochromes were analyzed on the Irys platform (Bionano Genomics). In total, 79 Gb of raw data greater than 150 kb, corresponding to 207× of the 7DS arm (Šafář et al., 2010), were collected from a single IrysChip (Supplemental Table S1). De novo assembly of the optical map was done through a pairwise comparison of all single molecules and graph building in IrysView software (Bionano Genomics). A *p*-value threshold of 1×10^{-10} was used during the pairwise assembly, 1×10^{-11} for the extension and refinement steps, and 1×10^{-15} for the final refinement. Sequence-to-map and map-to-map alignments were done with IrysView 2.1.1 software. To align the sequences, *cmaps* were generated from fasta files of the BAC sequence assemblies and 7D pseudomolecules of wheat whole-genome assemblies IWGSC RefSeq v1.0 (International Wheat Genome Sequencing Consortium, 2018) or Triticum 3.1 (Zimin et al., 2017). Query-to-anchor comparisons were performed with default parameters and a *p*-value threshold of 1×10^{-10}.

Narrowing Down the *Dn2401* Interval

To identify new markers within the interval delimited by the flanking markers *Xowm705* and *Xowm711*, we adopted a strategy proposed in our previous study (Staňková et al., 2015) with some modifications. Briefly, to design D-genome-specific primers, assembled sequences of five BAC clones spanning the *Dn2401* region in 7DS arm of CS were compared with homeologous sequence contigs from the short arms of CS chromosome 7A and 7B (Berkman et al., 2013). The specificity of the proposed primers was tested and PCR was optimized with DNA from the short arms of chromosomes 7A, 7B, and 7DS, which were flow-sorted from respective ditelosomic lines of CS. New polymorphisms between CI2401 and Glupro were identified by amplicon sequencing, with DNA of flow-sorted chromosome arm 7DS (CI2401) and chromosome 7D (Glupro) as templates. The 7D chromosome could be distinguished from other chromosomes of Glupro after FISHIS labeling of GAA repeats (Giorgi et al., 2013). Sequences of the PCR products from both parents were compared with Geneious version 7.1.2 and their specificity to the *Dn2401* region was verified by alignment to sequences of the assembled BAC clones and the CS RefSeq v1.0 (International Wheat Genome Sequencing Consortium, 2018).

Genetic mapping of the newly identified single nucleotide polymorphism (SNP) markers *Xowm713*, *714*, and *716* was carried out by Sanger-sequencing the respective amplicons from plants of the extended CI2401–Glupro F_2 mapping population. The presence–absence variation of the marker *Xowm715* was assayed by PCR and electrophoresis. Primers and PCR conditions for all markers are given in Table 2. The 149 newly added individuals of the mapping population (CI2401–Glupro F2–2) were also assayed for the original flanking markers *Xowm705* and *Xowm711*. Scoring of the RWA response was done at the Colorado State University Insectary in the same fashion as described for the first part of the mapping population (Staňková et al., 2015). Genotype and phenotype data of all 333 individuals from the extended CI2401–Glupro F_2

Publisher: AGRONOMY; Journal: TPG:The Plant Genome; Copyright: Will notify...
Volume: Will notify...; Issue: Will notify...; Manuscript: tpg-2018-09-0065; DOI: ; PII: <txtPII>
TOC Head: ; Section Head: ; Article Type: Original Research

mapping population were processed with JoinMap version 4.0 software (Van Ooijen and Voorrips, 2001).

Data Availability

The BAC hybrid assembly was submitted to NCBI (GenBank accession number MH806875). The optical map of CS 7DS is available at https://urgi.versailles.inra.fr/download/iwgsc/IWGSC_RefSeq_Annotations/v1.0/iwgsc_refseqv1.0_optical_maps_group7.zip (accessed 14 Feb. 2019). The optical map of CI2401 7DS is available on request.

RESULTS

Sequencing of Candidate BAC Clones

Five candidate BAC clones, spanning the *Dn2401* interval as delimited by Staňková et al., (2015), were Illumina sequenced and assembled, resulting in two scaffolds of 70.8 and 367 kb (Supplemental Fig. S1). To evaluate the completeness and correctness of the assembly, the scaffolds were aligned to the previously constructed BNG map of the CS 7DS arm (Staňková et al., 2016), which showed a 14.4-kb gap in the region corresponding to the BAC clone *TaaCsp7DS086H04*. We attempted to close that gap with two recently published whole-genome CS sequence assemblies (Zimin et al., 2017; International Wheat Genome Sequencing Consortium, 2018) but the BNG map indicated misassemblies in the region of interest for both of them (Fig. 1A). To complete the sequence of the region, long-read nanopore data were generated for the BAC clone *TaaCsp7DS086H04* and were combined with the BAC short-read data (Supplemental File S1), which resulted in a hybrid BAC assembly of 452,442 bp that spanned the entire interval between *Xowm705* and *Xowm711* by a continuous sequence of 304,109 bp (no Ns) and was in full agreement with the BNG map of CS 7DS (Fig. 1B, Supplemental Fig. S1).

Candidate Gene Identification and Comparison

Gene modeling of the assembled sequence of CS BAC clones spanning the *Dn2401* interval led to the prediction of 13 coding loci (Table 1, Supplemental Fig. S2), including the gene *HVA22*, whose coding sequence lies 182 bp distal to the interval but, because of its close proximity, the downstream regulatory region of the gene might potentially fall within the interval.

Narrowing down the *Dn2401* interval by mapping the new marker *Xowm713* excluded six proximally located genes (Table 1). Three of the remaining genes, *HVA22*, *Epoxide hydrolase 2* (*EH2*), and a gene coding for a 2-oxoglutarate and Fe(II)-dependent oxygenase [2OG-Fe(II) oxygenase] superfamily protein, were considered potential candidates for the RWA resistance gene *Dn2401*. *HVA22*, initially identified in barley (Shen et al., 1993), is a gene induced by abscisic acid, which is known to mediate developmental and physiological processes such as seed development and stress responses, including plant responses to aphid feeding (Smith and Boyko, 2007). *EH2* was found previously to be one of the genes showing increased expression after aphid feeding in a RWA-resistant line, PI220127 (Boyko et al., 2006). 2-Oxoglutarate and Fe(II)-dependent oxygenase appears to be involved in plant response to biotic stress because the

Publisher: AGRONOMY; Journal: TPG:The Plant Genome; Copyright: Will notify...
Volume: Will notify...; Issue: Will notify...; Manuscript: tpg-2018-09-0065; DOI: ; PII: <txtPII>
TOC Head: ; Section Head: ; Article Type: Original Research

underlying gene was found to be overexpressed after inoculating wheat with fungal pathogens (Zhang et al., 2014)(www.wheat-expression.com, accessed 14 Feb. 2019).

To compare the *HVA22* locus between the parents of the mapping population (the RWA-resistant line CI2401 and the RWA-susceptible Glupro), we sequenced in both of them a continuous region starting 3.2 kb upstream and ending 765 bp downstream of the *HVA22* coding sequence. The gene consisted of five exons that did not differ between the parents, and four introns, the second of which bore one SNP [A/G]. The A variant of the SNP is shared by the resistant line CI2401 and the susceptible CS. In the promoter region of both CI2401 and Glupro, we observed abscisic acid response elements ABRE2 and ABRE3 as well as three TATAA domains localized 1105, 717, and 163 bp upstream of the gene. In the 3' untranslated region (UTR), in both accessions, we found two AUAAA domains and in the downstream, region we precisely positioned a 43-bp deletion that was specific for Glupro, which was earlier identified as the marker *Xowm711* (Staňková et al., 2015). The sequence spanning this indel was also inspected in CS and it was confirmed that the susceptible CS had the same haplotype as the resistant CI2401 in the *HVA22* locus. Hence this polymorphism was not associated with the resistant phenotype.

To assess the *EH2* locus, we compared sequences starting 2871 bp upstream and ending 963 bp downstream of the *EH2* coding sequence that consisted of three exons and two introns. The total length of the coding sequence was 2275 bp. In the 5' flanking region of the gene, we found a *Ty1/Copia* retroelement located 1550 bp upstream of the coding sequence (Supplemental Fig. S3, Supplemental File S1). Alignment of the CI2401 and Glupro sequences revealed no polymorphism in the entire *EH2* region except for a single SNP [C/T] located in the 3'UTR, 290 bp downstream from the coding sequence. The T variant was shared by the susceptible cultivars CS and Glupro.

We also resequenced and compared a continuous DNA segment of 8470 bp covering the entire *2OG-Fe(II) oxygenase* locus as well as a neighboring gene coding for a bacterial trigger factor protein, including their up- and downstream regulatory regions (Supplemental Fig. S4). The sequence comparison revealed a single polymorphism corresponding to the marker *Xowm714*, located in the second exon of the gene for the bacterial trigger factor protein. This enabled a haplotype analysis of the locus, revealing that the A variant of the SNP was shared by the resistant line CI2401 and the susceptible CS. The data obtained by resequencing the two loci did not support them as the true *Dn2401* candidates. In the search for new markers in the interval, we also analyzed the entire gene for aspartic peptidase, including the adjacent regions and the majority of the gene for dual specificity protein phosphatase (Supplemental Table S3) without detecting any polymorphisms.

Bionano Genome Mapping and Structural Variations in the *Dn2401* Region

With the aim of inspecting the overall structure in the *Dn2401* interval in the CI2401 line and to compare it with the reference genome of CS, we investigated the region by optical mapping. By using the Bionano Genomics Irys platform and the approach applied previously for constructing the 7DS BNG map of CS, we generated a BNG map of the 7DS chromosome arm of CI2401 that consisted of 468 BNG map contigs with an average length of 765 kb and N50 of 1.36 Mb. This novel 7DS-CI2401 BNG map had a total length of 358 Mb and covered 94% of the estimated arm length (Supplemental Table S1).

Publisher: AGRONOMY; Journal: TPG:The Plant Genome; Copyright: Will notify...
Volume: Will notify...; Issue: Will notify...; Manuscript: tpg-2018-09-0065; DOI: ; PII: <txtPII>
TOC Head: ; Section Head: ; Article Type: Original Research

Alignments of the two 7DS BNG maps to the BAC assembly of the region led to the identification of two BNG map contigs spanning the *Dn2401* region in each of the accessions: contigs 279 and 280 for CS and contigs 165 and 199 for CI2401 (Fig. 1B). The reason for the BNG map split was the presence of a "fragile site" that arose from the occurrence of proximally located Nt.BspQI nicking sites on opposite DNA strands. This introduced a gap of ~15 kb in both BNG maps. Alignment of the two maps indicated an 8.2-kb indel at the distal end of the BNG map contigs 280 and 199 of CS and CI2401, respectively (Fig. 1B). Detailed analysis of the variable region in CS positioned the CS-specific insert ~12.7 kb upstream of the *EH2* gene and identified it as a recombined copy of a long terminal repeat retroelement from the *Ty3/gypsy* superfamily (Fig. 2), which was probably present as the native single-copy element in CI2401. The absence of the duplicated retroelement in CI2401 was validated by PCR with primers designed for diagnostic SNPs in long terminal repeats of the retroelements (Supplemental Table S2), which confirmed the presence of this duplicated recombined element in Glupro. Apart from this rearrangement, the alignment of the optical maps did not reveal other structural variations, indicating that the list of genes obtained from the CS sequence was most likely to be complete.

Narrowing Down the *Dn2401* Interval

In a previous study on the F_2 population of 184 individuals, we identified *Xowm705* and *Xowm711* as markers flanking the *Dn2401* gene and delimiting an interval of 0.83 cM (Staňková et al., 2015). To further increase the resolution of the genetic map, we expanded the mapping population by an additional 149 F_2 individuals and identified three new recombinants for the interval delimited by *Xowm705* and *Xowm711*. The new data extended the interval to 0.90 cM and localized *Dn2401* as 0.60 cM distal to *Xowm705* and 0.30 cM proximal to *Xowm711* (Fig. 3).

To further narrow down the *Dn2401* region, 46 new pairs of primers were designed and, after optimization, 44 primer pairs (96%) provided a single 7DS-specific PCR product. Sequencing of 41 amplicons from both parents did not detect any polymorphism, whereas one amplicon provided a mixed sequence, probably caused by a duplication within the 7DS. Only two amplicons yielded new SNPs, which were mapped in the complete CI2401–Glupro F_2 population. The SNPs, named *Xowm713* and *Xowm714* (Table 2), were mapped into the *Dn2401* interval. *Xowm713* became a new proximal flanking marker, reducing the interval to 0.45 cM and 137.5 kb, whereas *Xowm714* was found to cosegregate with *Dn2401* (Fig. 3). The structural variation revealed in the retroelement upstream of the *EH2* gene with *STS2* primers (Fig. 2) could be mapped as a presence–absence polymorphism in the mapping population. This new marker, labeled *Xowm715*, was found to cosegregate with *Xowm714* and *Dn2401*. Finally, we mapped the SNP identified in the 3'UTR of the *EH2* gene between *Xowm711* and *Xowm715*. This novel distal flanking marker, named *Xowm716* (Table 2), delimited a reduced *Dn2401* interval of 0.3 cM and 133.2 kb (Fig. 3).

DISCUSSION

Sequencing and Analysis of the *Dn2401* Region

The *Dn2401* region proved to be highly challenging for several sequencing technologies and approaches applied. The biggest obstacle to complete and correct assembly was the duplicated recombined retroelement (Fig. 2), which was found to be misassembled in the whole-genome

Publisher: AGRONOMY; Journal: TPG:The Plant Genome; Copyright: Will notify...
Volume: Will notify...; Issue: Will notify...; Manuscript: tpg-2018-09-0065; DOI: ; PII: <txtPII>
TOC Head: ; Section Head: ; Article Type: Original Research

assembly generated from short-read Illumina data (International Wheat Genome Sequencing Consortium, 2018) and could not be resolved even after reducing sample complexity to a single BAC clone, despite applying the several assemblers including SASSY and RAY, which, in combination, succeeded in assembling the remaining part of the interval. We expected that the duplication might be resolved in the hybrid assembly of Zimin et al. (2017), which combined short-read Illumina data with long-read data obtained through Pacific Biosciences sequencing. Surprisingly, optical mapping indicated even more discrepancies in the Triticum 3.1 assembly of the *Dn2401* region. The entire *Dn2401* interval could be resolved only after adding long nanopore reads, which enabled a reliable scaffolding of precise Illumina reads.

Bacterial artificial chromosome libraries, used previously for generating reference genomes, are available for many important crops, including maize (*Zea mays* L.) (Schnable et al. 2009), barley (IBGSC, 2012), and wheat (International Wheat Genome Sequencing Consortium, 2018). Chromosome- or chromosome-arm-specific BAC libraries have been constructed for all chromosomes of CS bread wheat (Šafář et al. 2010; International Wheat Genome Sequencing Consortium, 2018) and can be screened easily through pools of BAC clones (http://cnrgv.toulouse.inra.fr/en/library/wheat, accessed 14 Feb. 2019). Moreover, short-read assemblies of BAC clones have been generated for several chromosomes of CS and are publicly accessible (International Wheat Genome Sequencing Consortium, 2018; Tulpová et al. 2019). The availability of the BAC resources makes the precision resequencing of a region of interest (as used in the present work) widely applicable. In previous wheat cloning studies, the BAC-based approach was coupled with the Pacific Biosciences technique rather than nanopore sequencing (Schweiger et al., 2016; Roselli et al., 2017). This alternative is handicapped by shorter reads [e.g., an N50 of ~10 kb was reported in Zimin et al., (2017)] that hardly span the entire length of a BAC insert, typically exceeding 100 kb. Schweiger et al. (2016) sequenced 10 wheat BAC clones spanning their region of interest with Roche 454 and Pacific Biosciences technologies, but they were unable to produce a continuous assembly because of two gaps of several kb around transposable elements. The MinION sequencer is highly affordable (https://nanoporetech.com/products/minion, accessed 14 Feb. 2019) and relatively undemanding in terms of DNA input (less than 1µg per sequencing library), which makes it a favorable platform for small-scale in-house sequencing of specific BAC clones.

Optical BNG maps proved to be a useful and cost-efficient tool for overall sequence validation and the identification of structural variations in the region of interest. The minimum size of indels that can be resolved by optical mapping ranges from 1 kbp (Zhu et al., 2018) to as little as 500 bp (https://bionanogenomics.com/technology/dls-technology/, accessed 14 Feb. 2019) for the Bionano Genomics Irys and Saphyr platforms, respectively. To generate BNG maps, we took advantage of complexity reduction by flow-sorting the 7DS arm from wheat ditelosomic lines carrying 7DS arms as telocentric chromosomes. Preparation of a particular telosomic line from a cultivar of interest is technically simple but requires an amount of cytology that is difficult to predict, as the misdivision frequencies of individual chromosomes vary and cannot be gauged in advance. In several applications of the flow-sorted chromosomes, the use of telosomics has been circumvented by applying FISHIS (Giorgi et al. 2013), which labels DNA at trinucleotide microsatellites, typically GAA. This enables the flow-sorting of specific wheat chromosomes from the majority of wheat accessions with high purity (Vrána et al., 2016). In a pilot experiment, FISHIS labeling turned out to be problematic for BNG mapping of the Nt.BspQI nickase (unpublished data), perhaps because of an overlap of GAA with the Nt.BspQI recognition site. It is yet to be determined if the FISHIS procedure interferes with the Direct

Publisher: AGRONOMY; Journal: TPG:The Plant Genome; Copyright: Will notify...
Volume: Will notify...; Issue: Will notify...; Manuscript: tpg-2018-09-0065; DOI: ; PII: <txtPII>
TOC Head: ; Section Head: ; Article Type: Original Research

Label and Stain chemistry, introduced recently by Bionano Genomics as an alternative to nickase-based labeling (Deschamps et al., 2018). Whole-genome BNG maps are available for reference genomes of all major cereal crops, including rice (*Oryza sativa* L.) (Chen et al., 2017), maize (Jiao et al., 2017), barley (Mascher et al., 2017), and wheat (Luo et al. 2017); all of them have been generated on the Irys platform of Bionano Genomics. The new Bionano platform Saphyr, which has improved throughput and image resolution, mitigates the need for the complexity reduction and allows analyses of Gbp-sized genomes of interest in their entirety in a relatively short time and at a reasonable cost.

The *Dn2401* Candidate

Annotation of the *Dn2401* region delimited by the markers *Xowm711* and *Xowm713* indicated the presence of seven HC genes (Table 1; Supplemental Fig. S2); of these, *HVA22* (*TraesCS7D01G224100*), *EH2* (*TraesCS7D01G224200*), and *2OG-Fe(II) oxygenase* (*TraesCS7D01G224500*) were considered as potential resistance gene candidates on the basis of their preliminary functional analysis. Mapping of a new marker, *Xowm716*, into the region shifted *HVA22* further away from the *Dn2401* interval. Furthermore, we revealed that the resistant CI2401 and the susceptible CS shared the same haplotype in the *HVA22* region, which excluded *HVA22* as the *Dn2401* candidate. In contrast, the haplotype in the *EH2* region is common to the susceptible cultivars and is different from that of the resistant line CI2401. *EH2* was among the genes showing increased expression after aphid feeding in the RWA-resistant line PI220127, which bears the *Dnx* resistance gene in chromosome arm 7DS (Boyko et al., 2006). Both *Dn2401* and *Dnx* belong to a group of RWA resistance genes linked with the marker *gwm111* (Liu et al. 2001; Fazel-Najafabadi et al., 2015), which might be allelic (Liu et al. 2002). This makes *EH2* a robust candidate for the *Dn2401* gene.

Epoxide hydrolases (EHs) are enzymes present in all living organisms; they transform epoxide-containing lipids through the addition of water. Plants contain multiple soluble EH isoforms, which can be both constitutive and infection-induced (Newman et al., 2005). The transcription of the inducible enzymes can be increased by exogenous exposure to hormones, including the host-defense regulator methyl jasmonate (Stapleton et al., 1994). The substrate specificity and regulatory behavior of soluble plant EHs argue for a primary function of this enzyme in host defense and growth. The defensive functions can be related to both passive (cutin biosynthesis) and active (antifungal chemical synthesis) roles (Newman et al. 2005). The effect of the cuticle lipid (wax and cutin) composition has been considered in plant defense against both pathogens and pests. Studies on wheat's response to feeding by Hessian fly (*Mayetiola destructor*) larvae suggested cuticle composition and integrity are important components of host resistance (Kosma et al. 2010; Khajuria et al. 2013). The involvement of EHs in cutin biosynthesis matches well with the tolerance type of RWA resistance identified in CI2401 by Dong et al. (1997) and Voothuluru et al. (2006) and assayed in the present work.

We were also intrigued by a role of EHs in insects, whose juvenile hormone EH irreversibly degrades juvenile hormones, which is essential for insect metamorphosis. The gene for juvenile hormone EH was thus proposed as a promising target for hemipteran pest management (Tusun et al., 2017). In aphids, with their reproductive polyphenism (i.e., alternating reproductive modes from parthenogenesis to sexual reproduction in response to short photoperiods), the juvenile hormone degradation pathway appears to be involved in regulating the transition from asexual to sexual reproduction (Ishikawa et al., 2012). Interestingly, the EHs of cress and potato were

Publisher: AGRONOMY; Journal: TPG:The Plant Genome; Copyright: Will notify...
Volume: Will notify...; Issue: Will notify...; Manuscript: tpg-2018-09-0065; DOI: ; PII: <txtPII>
TOC Head: ; Section Head: ; Article Type: Original Research

shown to efficiently hydrolyze insect juvenile hormones (Morisseau et al., 2000), which lets us speculate about a possible pest–host interaction that could explain antibiosis (reduced populations) observed in the CI2401 line (Voothuluru et al., 2006).

During the analysis of the *EH2* locus, we resequenced a total of 6109 bp (Supplemental Fig. S3) from the resistant CI2401 and the susceptible Glupro and compared them with the reference genome of the susceptible CS. The only polymorphism discovered was a single SNP located in the 3'UTR, 290 bp downstream of the gene. A larger structural variation (a 8.2-kb indel) between the resistant line CI2401 and the susceptible cultivars CS and Glupro was identified 12.7 kb upstream of the gene. Both polymorphisms were found to cosegregate with the trait. It is yet to be determined by functional analysis if the SNP in the 3'UTR or the indel in the upstream region affect the expression of *EH2* and thus contribute to the resistance of CI2401.

The gene coding for the 2OG-Fe(II) oxygenase (*TraesCS7D01G224500*) was selected as a potential candidate because of its reported overexpression in response to a biotic stress, specifically, wheat infection by powdery mildew (*Blumeria graminis* f. sp. *tritici*) and stripe rust (*Puccinia striiformis* f. sp. *tritici*) (Zhang et al., 2014). On the other hand, no pathogen-dependent expression patternw were observed after inoculation of a resistant wheat line with *Fusarium* head blight (Schweiger et al., 2016), which suggested that 2OG-Fe(II) oxygenase was not a universal component of plant defense pathways. This, together with the absence of polymorphisms in the coding and regulatory regions of the gene, reduce the chance of *2OG-Fe(II) oxygenase* being the true *Dn2401* candidate. As to the remaining genes in the interval, we have not found any published evidence of their changed expression in response to aphid feeding (Boyko et al., 2006; Smith and Boyko, 2007). The comprehensive wheat transcriptomics database built on the RNA-seq data (http://www.wheat-expression.com, accessed 14 Feb. 2019)(Ramírez-González et al., 2018) does not currently include any pest-response related studies. However, it comprises data from multiple projects on host–pathogen interactions, which did not indicate any significant pathogen-related response for the genes *TraesCS7D01G224300* (aspartic peptidase), *TraesCS7D01G224400* (a bacterial trigger factor protein), *TraesCS7D01G224600* (nuclear pore complex protein), and *TraesCS7D01G224700* (dual specificity protein phosphatase) located in the *Dn2401* interval. Previous transcriptomic studies on wheat's response to RWA infestation were based on subtractive hybridization, cDNA-AFLP, RT-qPCR on a selected gene set, and microarray hybridization, which may provide limited information. We expect that RT-qPCR analysis of all genes from the critical interval will provide more comprehensive data to ultimately resolve the true *Dn2401* candidate.

Author Contributions

ZT, HT, KH, JM, and PN were involved in sequencing and assembling the *Dn2401* region. The annotation was done by PL, JM, and ZT. The 7DS ditelosomic line from CI2401 was prepared by AJL and DK and the optical map was generated by HT. Flow-sorting of particular chromosome arms was done by JV. Analysis of structural variation was performed by ZT. Marker development and genetic mapping were done by ZT and MM. Phenotyping was conducted by NLVL and FBP. The manuscript was drafted by ZT and HŠ, who conceived and coordinated the project. JD participated in project organization and manuscript writing.

Publisher: AGRONOMY; Journal: TPG:The Plant Genome; Copyright: Will notify...
Volume: Will notify...; Issue: Will notify...; Manuscript: tpg-2018-09-0065; DOI: ; PII: <txtPII>
TOC Head: ; Section Head: ; Article Type: Original Research

Supplemental Information

Supplemental Table S1: CI2401 7DS BNG map preparation and statistics
Supplemental Table S2: Primers and PCR conditions applied for interrogating local structural variation in the *Dn2401* region
Supplemental Table S3: Primers and PCR conditions applied for resequencing of candidate genes from parental lines
Supplemental Fig. S1: Comparison of a short-read and a hybrid BAC assembly of the *Dn2401* region.
Supplemental Fig. S2: Distribution of coding sequences in the wider *Dn2401* region in 'Chinese Spring'.
Supplemental Fig. S3: Analysis of the sequence in the EH2 region.
Supplemetal Fig. S4: Resequencing of loci for 2OG-Fe(II) oxygenase and bacterial trigger factor protein.

Conflict of Interest Disclosure

The authors declare that there is no conflict of interest.

ACKNOWLEDGMENTS

We thank Prof. B.S. Gill for providing seeds of the wheat ditelosomic lines. We acknowledge the excellent assistance of Romana Šperková, Eva Jahnová, Radka Tušková, and Marie Seifertová in chromosome sorting, preparation of chromosomal DNA, and BAC isolation. We also acknowledge the help of Hana Vanžurová with construction of the genetic map. The authors thank the IWGSC for prepublication access to IWGSC RefSeq v1.0. This work was supported by the Czech Science Foundation (award No. P501/12/2554); the Czech Ministry of Education, Youth and Sports (National Program of Sustainability I, grant LO1204); and the Czech Academy of Sciences (RVO:60077344).

REFERENCES

<jrn>Abeel, T., T. Van Parys, Y. Saeys, J. Galagan, and Y. Van de Peer. 2012. GenomeView: A next-generation genome browser. Nucleic Acids Res. 40. e12. doi:10.1093/nar/gkr995</jrn>

<jrn>Anderson, G., D. Papa, J. Peng, M. Tahir, and N.L.V. Lapitan. 2003. Genetic mapping of *Dn7*, a rye gene conferring resistance to the Russian wheat aphid in wheat. Theor. Appl. Genet. 107:1297–1303. doi:10.1007/s00122-003-1358-1</jrn>

<jrn>Anderson, V.A., S.D. Haley, F.B. Peairs, L. van Eck, J.E. Leach, and N.L.V. Lapitan. 2014. Virus-induced gene silencing suggests that *(1,3;1,4)-β-glucanase* is a susceptibility factor in the compatible Russian wheat aphid– wheat interaction. Mol. Plant Microbe Interact. 27:913–922. doi:10.1094/MPMI-05-13-0141-R</jrn>

<jrn>Aykut Tonk, F.A., D. İştipliler, M. Tosun, F. Turanli, H. İlbi, and M. Çakir. 2016. Genetic mapping and inheritance of Russian wheat aphid resistance gene in accession IG 100695. Plant Breed. 135:21–25. doi:10.1111/pbr.12339</jrn>

<jrn>Berkman, P.J., P. Visendi, H.C. Lee, J. Stiller, S. Manoli, M.T. Lorenc, et al. 2013. Dispersion and domestication shaped the genome of bread wheat. Plant Biotechnol. J. 11:564–571. doi:10.1111/pbi.12044</jrn>

<jrn>Boisvert, S., F. Laviolette, and J. Corbeil. 2010. Ray: Simultaneous assembly of reads from a mix of high-throughput sequencing technologies. J. Comput. Biol. 17:1519–1533. doi:10.1089/cmb.2009.0238</jrn>

<jrn>Botha, A.M., Y. Li, and N.L.V. Lapitan. 2005. Cereal host interactions with Russian wheat aphid: A review. J. Plant Interact. 1:211–222. doi:10.1080/17429140601073035</jrn>

<jrn>Boyko, E.V., C.M. Smith, V.K. Thara, J.M. Bruno, Y. Deng, S.R. Starkey, et al. 2006. Molecular basis of plant gene expression during aphid invasion: Wheat *Pto*- and *Pti*-like sequences are involved in interactions between wheat and Russian wheat aphid (Homoptera: Aphididae). J. Econ. Entomol. 99:1430–1445. doi:10.1093/jee/99.4.1430</jrn>

<jrn>Burd, J.D., and R.L. Burton. 1992. Characterization of plant damage caused by Russian wheat aphid (Homoptera: Aphididae). J. Econ. Entomol. 85:2017–2022. doi:10.1093/jee/85.5.2017</jrn>

<jrn>Chen, P., X. Jing, B. Liao, Y. Zhu, J. Xu, R. Liu, et al. 2017. BioNano genome map resource for *Oryza sativa* ssp. *japonica* and *indica* and its application in rice genome sequence correction and gap filling. Mol. Plant 10:895–898. doi:10.1016/j.molp.2017.02.003</jrn>

<jrn>Deschamps, S., Y. Zhang, V. Llaca, L. Ye, G. May, and H. Lin. 2018. A chromosome-scale assembly of the sorghum genome using nanopore sequencing and optical mapping. Nat. Commun. 9: 4844.</jrn>

<jrn>Dong, H., J.S. Quick, and Y. Zhang. 1997. Inheritance and allelism of Russian wheat aphid resistence in several wheat lines. Plant Breed. 116:449–453. doi:10.1111/j.1439-0523.1997.tb01029.x</jrn>

<jrn>Du Toit, F. 1987. Resistance in wheat (*Triticum aestivum*) to *Diuraphis noxia* (Homoptera: Aphididae). Cereal Res. Commun. 15:175–179.</jrn>

<jrn>Du Toit, F. 1989. Inheritance of resistance in two *Triticum aestivum* lines to Russian wheat aphid (Homoptera: Aphididae). J. Econ. Entomol. 82:1251–1253. doi:10.1093/jee/82.4.1251</jrn>

<jrn>Du Toit, F., W.G. Wessel, and G.F. Marais. 1995. The chromosome arm location of the Russian wheat aphid resistance gene, *Dn5*. Cereal Res. Commun. 23:15–17.</jrn>

<jrn>Fazel-Najafabadi, M., J. Peng, F.B. Peairs, H. Šimková, A. Kilian, and N.L.V. Lapitan. 2015. Genetic mapping of resistance to *Diuraphis noxia* (Kordjumov) biotype 2 in wheat (*Triticum aestivum* L.) accession CI2401. Euphytica 203:607–614. doi:10.1007/s10681-014-1284-0</jrn>

<jrn>Giorgi, D., A. Farina, V. Grosso, A. Gennaro, C. Ceoloni, and S. Lucretti. 2013. FISHIS: Fluorescence in situ hybridization in suspension and chromosome flow sorting made easy. PLoS One. doi:10.1371/journal.pone.0057994</jrn>

<jrn>Harvey, T.L., and T.J. Martin. 1990. Resistance to Russian wheat aphid, *Diuraphis noxia*, in wheat (*Triticum aestivum*). Cereal Res. Commun. 18:127–129.</jrn>

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

<jrn>Ishikawa, A., K. Ogawa, H. Gotoh, T.K. Walsh, D. Tagu, J.A. Brisson, et al. 2012. Juvenile hormone titre and related gene expression during the change of reproductive modes in the pea aphid. Insect Mol. Biol. 21:49–60. doi:10.1111/j.1365-2583.2011.01111.x</jrn>

<jrn>International Wheat Genome Sequencing Consortium. 2018. Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science 361: eaar7191.</jrn>

<jrn>Jiao, Y., P. Peluso, J. Shi, T. Liang, M.C. Stitzer, B. Wang, et al. 2017. Improved maize reference genome with single-molecule technologies. Nature 546:524–527.</jrn>

<jrn>Kearse, M., R. Moir, A. Wilson, S. Stones-Havas, M. Cheung, S. Sturrock, et al. 2012. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics. 28:1647–1649. doi:10.1093/bioinformatics/bts199</jrn>

<jrn>Khajuria, C., H.Y. Wang, X.M. Liu, S. Wheeler, J.C. Reese, M. El Bouhssini, et al. 2013. Mobilization of lipids and fortification of cell wall and cuticle are important in host defense against Hessian fly. BMC Genomics 14:423. doi:10.1186/1471-2164-14-423</jrn>

<jrn>Kosma, D.K., J.A. Nemacheck, M.A. Jenks, and C.E. Wiliams. 2010. Changes in properties of wheat leaf cuticle during interaction with Hessian fly. Plant J. 63:31–43.</jrn>

<jrn>Lam, E.T., A. Hastie, C. Lin, D. Ehrlich, S.K. Das, M.D. Austin, et al. 2012. Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. Nat. Biotechnol. 30:771–776. doi:10.1038/nbt.2303</jrn>

<jrn>Lapitan, N.L.V., J. Peng, and V. Sharma. 2007. A high-density map and PCR markers for Russian wheat aphid resistance gene *Dn7* on chromosome 1RS/1BL. Crop Sci. 47:811–818. doi:10.2135/cropsci2006.08.0529</jrn>

<jrn>Leroy, P., N. Guilhot, H. Sakai, A. Bernard, F. Choulet, S. Theil, et al. 2012. TriAnnot: A versatile and high performance pipeline for the automated annotation of plant genomes. Front. Plant Sci. 3:5. doi:10.3389/fpls.2012.00005</jrn>

<jrn>Liu, X.M., C.M. Smith, B.S. Gill, and V. Tolmay. 2001. Microsatellite markers linked to six Russian wheat aphid resistance genes in wheat. Theor. Appl. Genet. 102:504–510. doi:10.1007/s001220051674</jrn>

<jrn>Liu, X.M., C.M. Smith, and B.S. Gill. 2002. Identification of microsatellite markers linked to Russian wheat aphid resistance genes *Dn4* and *Dn6*. Theor. Appl. Genet. 104:1042–1048. doi:10.1007/s00122-001-0831-y</jrn>

<jrn>Luo, M.C., Y.Q. Gu, D. Puiu, H. Wang, S.O. Twardziok, K.R. Deal, et al. 2017. Genome sequence of the progenitor of the wheat D genome *Aegilops tauschii*. Nature 551:498–502.</jrn>

<jrn>Ma, Z.Q., A. Saidi, J.S. Quick, and N.L.V. Lapitan. 1998. Genetic mapping of Russian wheat aphid resistance genes *Dn2* and *Dn4* in wheat. Genome 41:303–306. doi:10.1139/g98-013</jrn>

<jrn>Marais, G.F., M. Horn, and F. Du Toit. 1994. Intergeneric transfer (rye to wheat) of gene(s) for Russian wheat aphid resistance. Plant Breed. 113:265–271. doi:10.1111/j.1439-0523.1994.tb00735.x</jrn>

<jrn>Mascher, M., H. Gundlach, A. Himmelbach, S. Beier, S.O. Twardziok, T. Wicker, et al. 2017. A chromosome conformation capture ordered sequence of the barley genome. Nature 544:427–433. doi:10.1038/nature22043</jrn>

<jrn>Miller, C.A., A. Altinkut, and N.L.V. Lapitan. 2001. A microsatellite marker for tagging a wheat gene conferring resistance to the Russian wheat aphid. Crop Sci. 41:1584–1589. doi:10.2135/cropsci2001.4151584x</jrn>

<jrn>Morisseau, C., J.K. Beetham, F. Pinot, S. Debernard, J.W. Newman, and B.D. Hammock. 2000. Cress and potato soluble epoxide hydrolases: Purification, biochemical characterization, and comparison to mammalian enzymes. Arch. Biochem. Biophys. 378:321–332. doi:10.1006/abbi.2000.1810</jrn>

<jrn>Newman, J.W., C. Morisseau, and B.D. Hammock. 2005. Epoxide hydrolases: Their roles and interactions with lipid metabolism. Prog. Lipid Res. 44:1–51. doi:10.1016/j.plipres.2004.10.001</jrn>

<jrn>Peng, J., H. Wang, S.D. Haley, F.B. Peairs, and N.L.V. Lapitan. 2007. Molecular mapping of the Russian wheat aphid resistance gene *Dn2414* in wheat. Crop Sci. 47:2418–2429. doi:10.2135/cropsci2007.03.0137</jrn>

<jrn>Peng, J.H., Y. Bai, S.D. Haley, and N.L.V. Lapitan. 2009. Microsatellite-based molecular diversity of bread wheat germplasm and association mapping of wheat resistance to the Russian wheat aphid. Genetica (The Hague) 135:95–122.</jrn>

<jrn>Ramírez-González, R. H., Borrill, P., D. Lang, S. A. Harrington, J. Brinton, L. Venturini, et al. 2018. The transcriptional landscape of polyploid wheat. Science 361: eaar6089. doi:10.1126/science.aar6089</jrn>

<jrn>Roselli, S., A. Olry, S. Vautrin, O. Coriton, D. Ritchie, G. Galati, et al. 2017. A bacterial artificial chromosome (BAC) genomic approach reveals partial clustering of the furanocoumarin pathway genes in parsnip. Plant J. 89:1119–1132. doi:10.1111/tpj.13450</jrn>

<jrn>Šafář, J., H. Šimková, M. Kubaláková, J. Číhalíková, P. Suchánková, J. Bartoš, et al. 2010. Development of chromosome-specific BAC resources for genomics of bread wheat. Cytogenet. Genome Res. 129:211–223. doi:10.1159/000313072</jrn>

<jrn>Saidi, A., and J.S. Quick. 1996. Inheritance and allelic relationships among Russian wheat aphid resistance genes in winter wheat. Crop Sci. 36:256–258. doi:10.2135/cropsci1996.0011183X003600020007x</jrn>

<jrn>Schnable, P.S., D. Ware, R.S. Fulton, J.C. Stein, F. Wei, P. Pasternak, et al. 2009. The B73 maize genome: Complexity, diversity, and dynamics. Science 326:1112–1115. doi:10.1126/science.1178534</jrn>

<jrn>Schweiger, W., B. Steiner, S. Vautrin, T. Nussbaumer, G. Siegwart, M. Zamini, et al. 2016. Suppressed recombination and unique candidate genes in the divergent haplotype encoding *Fhb1*, a major *Fusarium* head blight resistance locus in wheat. Theor. Appl. Genet. 129:1607–1623. doi:10.1007/s00122-016-2727-x</jrn>

<jrn>Shen, Q., J.U. Scot, and T.H.D. Ho. 1993. Hormone response complex in a novel abscisic acid and cycloheximide-inducible barley gene. J. Biol. Chem. 268:23652–23660.</jrn>

<jrn>Šimková, H., J. Šafář, M. Kubaláková, P. Suchánková, J. Číhalíková, H. Robert-Quatre, et al. 2011. BAC libraries from wheat chromosome 7D: Efficient tool for positional cloning of aphid resistance genes. J. Biomed. Biotechnol. 2011:302543. doi:10.1155/2011/302543</jrn>

<jrn>Smith, C.M., and E.V. Boyko. 2007. The molecular bases of plant resistance and defense responses to aphid feeding: Current status. Entomol. Exp. Appl. 122:1–16. doi:10.1111/j.1570-7458.2006.00503.x</jrn>

<jrn>Staňková, H., M. Valárik, N.L.V. Lapitan, P.J. Berkman, J. Batley, D. Edwards, et al. 2015. Chromosomal genomics facilitates fine mapping of a Russian wheat aphid resistance gene. Theor. Appl. Genet. 128:1373–1383. doi:10.1007/s00122-015-2512-2</jrn>

<jrn>Staňková, H., R. Hastie, S. Chan, J. Vrána, Z. Tulpová, M. Kubaláková, et al. 2016. BioNano genome mapping of individual chromosomes supports physical mapping and sequence assembly in complex plant genomes. Plant Biotechnol. J. 14:1523–1531. doi:10.1111/pbi.12513</jrn>

<jrn>Stapleton, A., J.K. Beetham, F. Pinot, J.E. Garbarino, D.R. Rockhold, M. Friedman, et al. 1994. Cloning and expression of soluble epoxide hydrolase from potato. Plant J. 6:251–258. doi:10.1046/j.1365-313X.1994.6020251.x</jrn>

<jrn>Tulpová, Z., M.C. Luo, H. Toegelová, P. Visendi, S. Hayashi, P. Vojta, et al. 2019. Integrated physical map of bread wheat chromosome arm 7DS to facilitate gene cloning and comparative studies. N. Biotechnol. 48:12–19. doi:10.1016/j.nbt.2018.03.003</jrn>

<jrn>Tusun, A., M. Li, X.Z. Liang, T. Yang, B. Yang, and G.R. Wang. 2017. Juvenile hormone epoxide hydrolase: A promising target for hemipteran pest management. Sci. Rep. 7:789. doi:10.1038/s41598-017-00907-0 [erratum: 8: 6246]</jrn>

<jrn>Valdez, V.A., P.F. Byrne, N.L.V. Lapitan, F.B. Peairs, A. Bernardo, G. Bai, et al. 2012. Inheritance and genetic mapping of Russian wheat aphid resistance in Iranian wheat landrace accession PI626580. Crop Sci. 52:676–682. doi:10.2135/cropsci2011.06.0331</jrn>

<jrn>Van Eck, L., R.M. Davidson, S. Wu, B.Y. Zhao, A.M. Botha, J.E. Leach, et al. 2014. The transcriptional network of *WRKY53* in cereals links oxidative responses to biotic and abiotic stress inputs. Funct. Integr. Genomics 14:351–362. doi:10.1007/s10142-014-0374-3</jrn>

<jrn>Van Ooijen, J.W., and R.E. Voorrips. 2001. JoinMap 3.0, software for the calculation of genetic linkage maps. Plant J. 3:739–744.</jrn>

<jrn>Voothuluru, P., J. Meng, C. Khajuria, J. Louis, L. Zhu, S. Starkey, et al. 2006. Categories and inheritance of resistance to Russian wheat aphid (Homoptera: Aphididade) biotype 2 in a selection from wheat cereal introduction 2401. J. Econ. Entomol. 99:1854–1861. doi:10.1093/jee/99.5.1854</jrn>

<unknown>Vrána, J., P. Cápal, H. Šimková, M. Karafiátová, J. Čížková, and J. Doležel. 2016. Flow analysis and sorting of plant chromosomes. Curr. Protoc. Cytom. 78:5.3.1–5.3.43. doi:10.1002/cpcy.9</unknown>

<bok>Walters, M.C., F. Penn, F. Du Toit, T.C. Botha, K. Aalberrsberg, P.H. Mewitt, and S.W. Broodryk. 1980. Russian wheat aphid. Farming in South Africa. Leaflet Series. Wheat G3: 1-6. South Africa: Dep. of Agriculture.</bok>

<jrn>Yazdani, M., G. Baker, H. DeGraaf, K. Henry, K. Hill, B. Kimber, et al. 2017. First detection of Russian wheat aphid *Diuraphis noxia* Kordjumov (Hemiptera: Aphididae) in Australia: A major threat to cereal production. Aust. Entomol. 57:410–417. doi:10.1111/aen.12292</jrn>

<jrn>Zhang, H., Y. Yang, C. Wang, M. Liu, H. Li, Y. Fu, et al. 2014. Large-scale transcriptome comparison reveals distinct gene activations in wheat responding to stripe rust and powdery mildew. BMC Genomics 15:898. doi:10.1186/1471-2164-15-898</jrn>

<jrn>Zhu, T., Z. Hu, J.C. Rodriguez, K.R. Deal, J. Dvorak, J.P. Vogel, Z. Liu, and M.C. Luo. 2018. Analysis of *Brachypodium* genome with genome-wide optical maps. Genome 61(8):559–565.

Zimin, A.V., G. Marcais, D. Puiu, M. Roberts, S.L. Salzberg, and J.A. Yorke. 2013. The MaSuRCA genome assembler. Bioinformatics 29:2669–2677.</jrn>

<jrn>Zimin, A.V., D. Puiu, R. Hall, S. Kingan, B.J. Clavijo, and S.L. Salzberg. 2017. The first near-complete assembly of the hexaploid bread wheat genome, *Triticum aestivum*. GigaSci. 6:1–7. doi:10.1093/gigascience/gix097</jrn>

Publisher: AGRONOMY; Journal: TPG:The Plant Genome; Copyright: Will notify...
Volume: Will notify...; Issue: Will notify...; Manuscript: tpg-2018-09-0065; DOI: ; PII: <txtPII>
TOC Head: ; Section Head: ; Article Type: Original Research

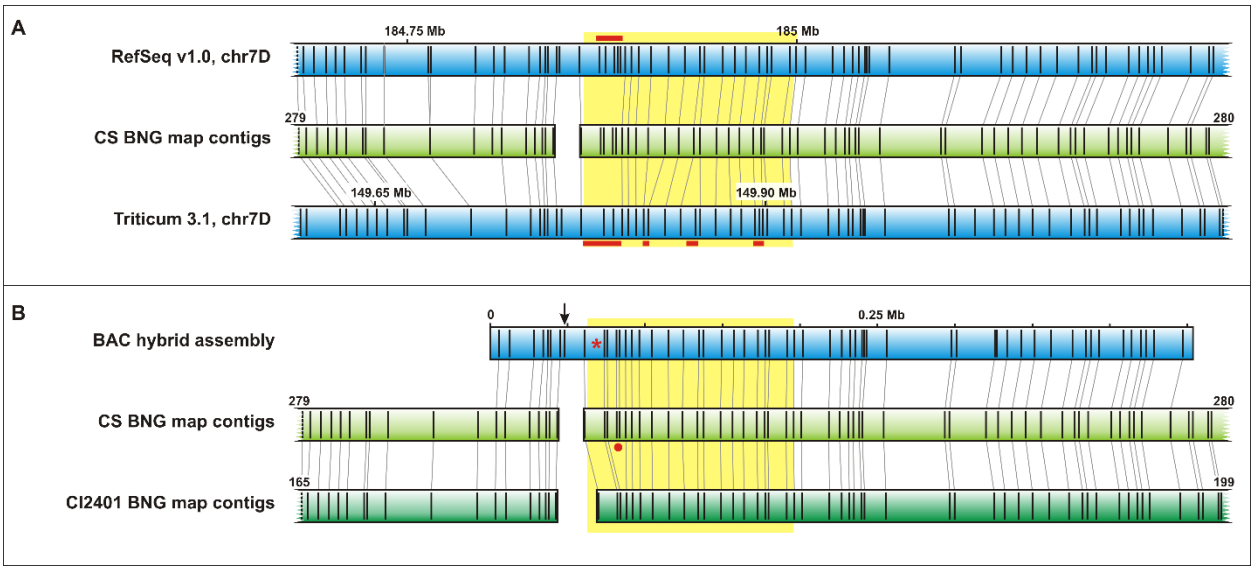


Fig. 1. Assembly and structural variation of the *Dn2401* region in wheat. (A) Alignments of the Bionano genome map (BNG) map of the short arm of wheat chromosome 7D (7DS) Chinese Spring (light green bar) (CS) to in silico Nt.BspQI digestion of whole-genome wheat assemblies (blue bars) showed misassembled regions (red lines) both in the IWGSC RefSeq v1.0 (IWGSC, 2018) and Triticum 3.1 (Zimin et al., 2017) assemblies. Numbers above the blue bars indicate positions in the 7D pseudomolecules. (B) The hybrid assembly of CS bacterial artificial chromosome (BAC) clones (blue bars) is in agreement with the CS BNG map. The 15-kb gap between BNG map contigs 279 to 280 (CS) and 165 to 199 (CI2401) is a consequence of a fragile site in the DNA molecule (arrow). Alignment of 7DS BNG maps from susceptible CS and resistant CI2401 indicated structural variation (red dot) upstream of the *EH2* gene (red asterisk). The yellow box highlights the *Dn2401* region as being delimited by *Xowm716* and *Xowm713*.

Publisher: AGRONOMY; Journal: TPG:The Plant Genome; Copyright: Will notify...
Volume: Will notify...; Issue: Will notify...; Manuscript: tpg-2018-09-0065; DOI: ; PII: <txtPII>
TOC Head: ; Section Head: ; Article Type: Original Research

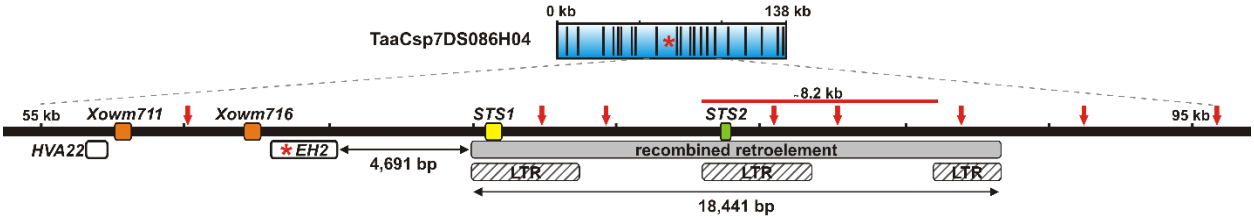


Fig. 2. Annotation and comparative analysis of the *HVA22–EH2* region in wheat. The analysis was performed on a hybrid assembly of the bacterial artificial chromosome (BAC) clone *TaaCsp086H04*, shown as Nt.BspQI in silico digestion (blue bar). Nt.BspQI recognition sites are depicted as black vertical lines (top) or red arrows (bottom). White boxes stand for coding sequences; the orange, yellow and green boxes correspond to markers applied in the study. The red line indicates the approximate position and size of the variable region missing in CI2401, as deduced from a comparison of Bionano genome map (BNG) maps and a polymerase chain reaction assay for the *STS*2 marker.

Publisher: AGRONOMY; Journal: TPG:The Plant Genome; Copyright: Will notify...
Volume: Will notify...; Issue: Will notify...; Manuscript: tpg-2018-09-0065; DOI: ; PII: <txtPII>
TOC Head: ; Section Head: ; Article Type: Original Research

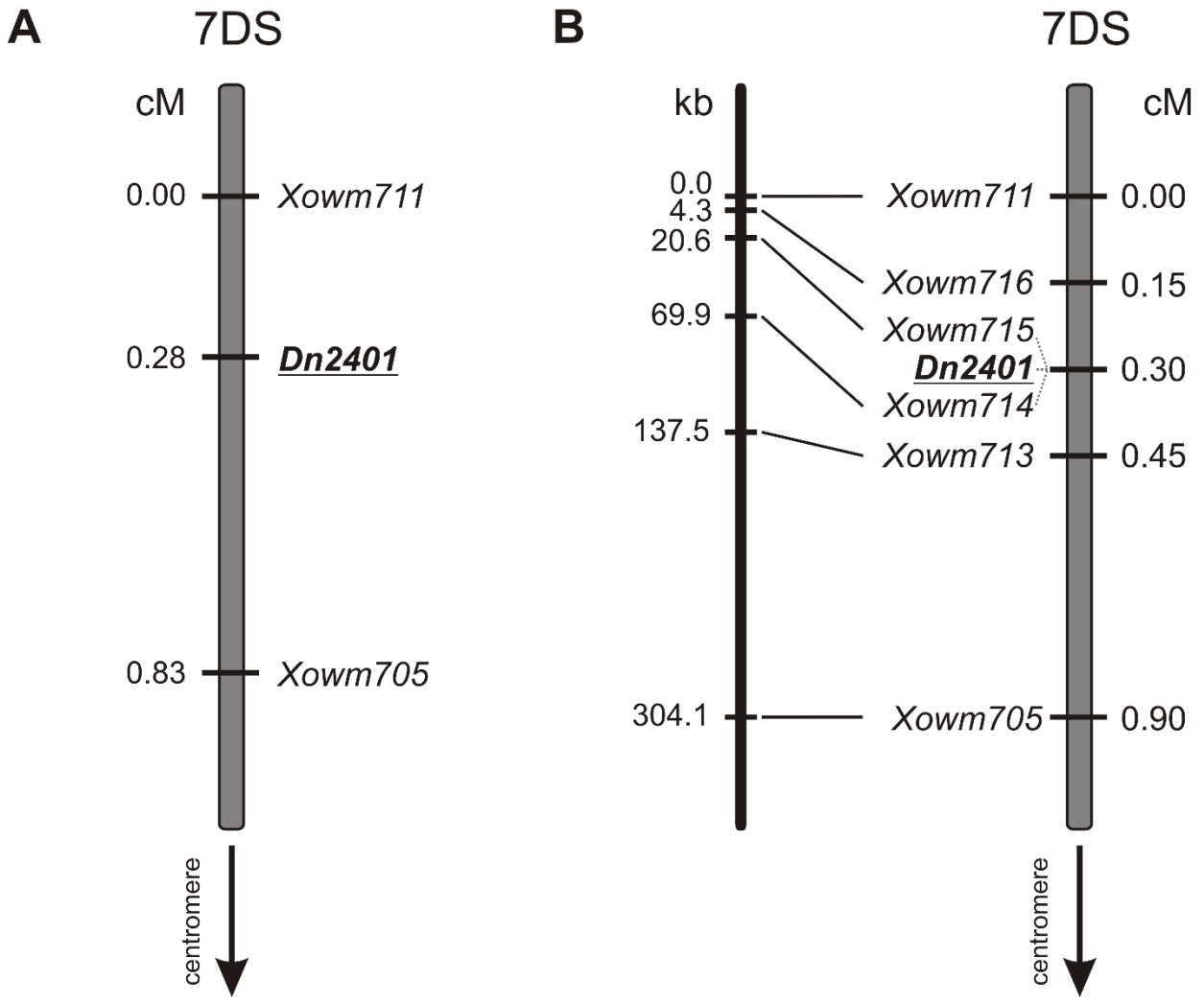


Fig. 3. Genetic and physical maps of the *Dn2401* region in wheat. (A) The *Dn2401* interval, as delimited in a previous study (Staňková et al., 2015). (B) The genetic map generated in the current study with an extended mapping population and new markers. The physical distances of the markers are shown on the left.

Publisher: AGRONOMY; Journal: TPG:The Plant Genome; Copyright: Will notify...
Volume: Will notify...; Issue: Will notify...; Manuscript: tpg-2018-09-0065; DOI: ; PII: <txtPII>
TOC Head: ; Section Head: ; Article Type: Original Research

Table 1. Predicted high-confidence coding sequences in the wider *Dn2401* region of Chinese Spring wheat

Gene No.	IWGSC RefSeq ID†	Functional analysis	Domain ID
1	*TraesCS7D01G224100.1* *TraesCS7D01G224100.2*	HVA22	PF03134, IPR004345
2	***TraesCS7D01G224200.1***	**Epoxide hydrolase 2‡**	**PF00561**, **IPR000073**
3	***TraesCS7D01G224300.1***	**Aspartic peptidase family**	**PF14543**, **PF14541**, **IPR001461**
4	***TraesCS7D01G224400.1***	**Bacterial trigger factor protein**	**PF05697**, **PF05698**, **IPR005215**
5	***TraesCS7D01G224500.1***	**2-oxoglutarate and Fe(II)-dependent oxygenase superfamily protein**	**PF14226**, **PF03171**, **IPR005123**
6	***TraesCS7D01G224600.1***	**Nuclear pore complex protein Nup98-Nup96**	–
7	***TraesCS7D01G224700.1***	**Dual specificity protein phosphatase, putative**	**PF00782**, **IPR000340**
8	*TraesCS7D01G224800.1*	Microtubule associated family protein	PF03999
9	*TraesCS7D01G224900.1*	RNA recognition motif	IPR000504
10	*TraesCS7D01G225000.1*	Membrane protein	PF09991, IPR018710
11	*TraesCS7D01G225100.1*	Acyl-coenzyme A oxidase	PF02770, PF00441, PF01756, IPR002655
12	*TraesCS7D01G225200.1*	Vacuolar protein sorting-associated protein VTA1	PF04652, IPR023175
13	*TraesCS7D01G225300.1*	PHD finger protein	PF00628, PF01426, IPR001025

† International Wheat Genome Sequencing Consortium RefSeq ID as in International Wheat Genome Sequencing Consortium **(**2018).

‡ The wider region was delimited by markers *Xowm705* and *Xowm711*. Coding sequences from the reduced interval delimited by markers *Xowm716* and *Xowm713* are in bold.

Publisher: AGRONOMY; Journal: TPG:The Plant Genome; Copyright: Will notify...
Volume: Will notify...; Issue: Will notify...; Manuscript: tpg-2018-09-0065; DOI: ; PII: <txtPII>
TOC Head: ; Section Head: ; Article Type: Original Research

Table 2. Markers mapped in the *Dn2401* region of wheat

Marker ID	Polymorphism	Primers	Ta °C	Amplicon size bp
Xowm713	SNP G/A	Forward: CGTGCATGATCCTCGACTATGAT Reverse: TTGCCTATTTTAACAATGCTCGT	67	467
Xowm714	SNP A/G	Forward: TCTGTAATGTGGAATGTTGTCTTAGT Reverse: GGCTGAAACAAAGAATCCATC	64	574
Xowm715	PAV	Forward: CCCTCGATACGAGCTGGA Reverse: GAGGGAGGGAGGTTGTCA	62	232
Xowm716	SNP C/T	Forward: GGGCAAAATGGTCTTTTCAC Reverse: AAGAAATTCGACTGAAATGAGGA	62	579

† SNP, single nucleotide polymorphism; PAV, presence–absence variation

APPENDIX III

Shifting the limits in wheat research and breeding using a fully annotated reference genome

International Wheat Genome Consortium

Science 361, eaar7191, 2018

doi: 10.1126/science.aar7191

IF: 41.058

RESEARCH ARTICLE SUMMARY

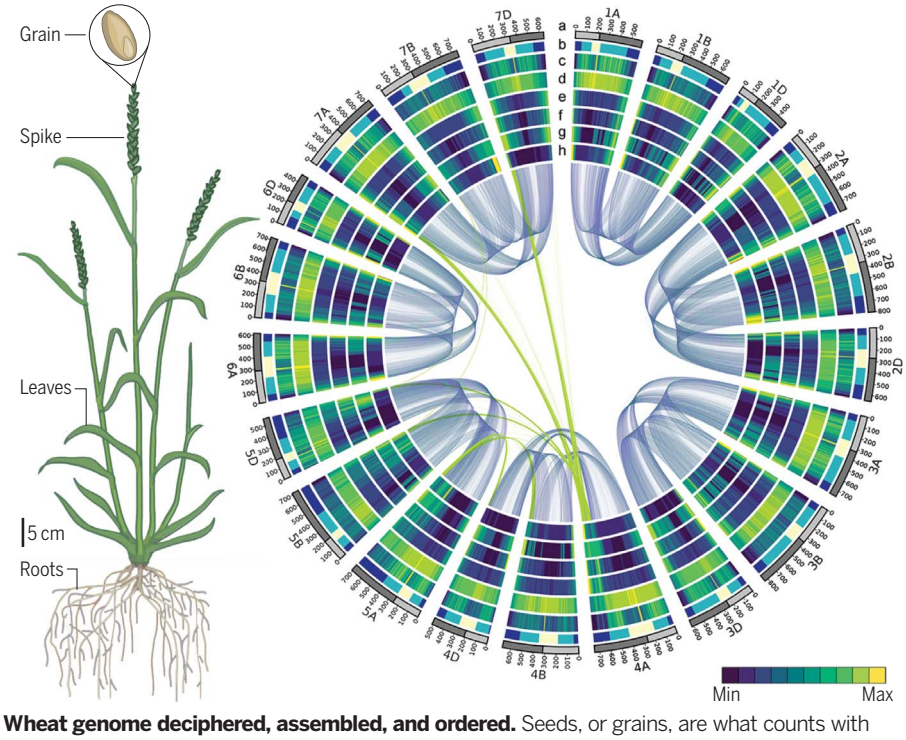
WHEAT GENOME

Shifting the limits in wheat research and breeding using a fully annotated reference genome

International Wheat Genome Sequencing Consortium (IWGSC)*

INTRODUCTION: Wheat (*Triticum aestivum* L.) is the most widely cultivated crop on Earth, contributing about a fifth of the total calories consumed by humans. Consequently, wheat yields and production affect the global economy, and failed harvests can lead to social unrest. Breeders continuously strive to develop improved varieties by fine-tuning genetically complex yield and end-use quality parameters while maintaining stable yields and adapting the crop to regionally specific biotic and abiotic stresses.

RATIONALE: Breeding efforts are limited by insufficient knowledge and understanding of wheat biology and the molecular basis of central agronomic traits. To meet the demands of human population growth, there is an urgent need for wheat research and breeding to accelerate genetic gain as well as to increase and protect wheat yield and quality traits. In other plant and animal species, access to a fully annotated and ordered genome sequence, including regulatory sequences and genome-diversity information, has promoted the development of systematic and more time-efficient approaches for the selection and understanding of important traits. Wheat has lagged behind, primarily owing to the challenges of assembling a genome that is more than five times as large as the human genome, polyploid, and complex, containing more than 85% repetitive DNA. To provide a foundation for improvement through molecular breeding, in 2005, the International Wheat Genome Sequencing Consortium set out to deliver a high-quality annotated reference genome sequence of bread wheat.



Wheat genome deciphered, assembled, and ordered. Seeds, or grains, are what counts with respect to wheat yields (left panel), but all parts of the plant contribute to crop performance. With complete access to the ordered sequence of all 21 wheat chromosomes, the context of regulatory sequences, and the interaction network of expressed genes—all shown here as a circular plot (right panel) with concentric tracks for diverse aspects of wheat genome composition—breeders and researchers now have the ability to rewrite the story of wheat crop improvement. Details on value ranges underlying the concentric heatmaps of the right panel are provided in the full article online.

RESULTS: An annotated reference sequence representing the hexaploid bread wheat genome in the form of 21 chromosome-like sequence assemblies has now been delivered, giving access to 107,891 high-confidence genes, including their genomic context of regulatory sequences. This assembly enabled the discovery of tissue- and developmental stage–related gene coexpression networks using a transcriptome atlas representing all stages of wheat development. The dynamics of change in complex gene families involved in environmental adaptation and end-use quality were revealed at subgenome resolution and contextualized to known agronomic single-gene or quantitative trait loci. Aspects of the future value of the annotated assembly for molecular breeding and research were exemplarily illustrated by resolving the genetic basis of a quantitative trait locus conferring resistance to abiotic stress and insect damage as well as by serving as the basis for genome editing of the flowering-time trait.

ON OUR WEBSITE

Read the full article at http://dx.doi.org/10.1126/science.aar7191

CONCLUSION: This annotated reference sequence of wheat is a resource that can now drive disruptive innovation in wheat improvement, as this community resource establishes the foundation for accelerating wheat research and application through improved understanding of wheat biology and genomics-assisted breeding. Importantly, the bioinformatics capacity developed for model-organism genomes will facilitate a better understanding of the wheat genome as a result of the high-quality chromosome-based genome assembly. By necessity, breeders work with the genome at the whole chromosome level, as each new cross involves the modification of genome-wide gene networks that control the expression of complex traits such as yield. With the annotated and ordered reference genome sequence in place, researchers and breeders can now easily access sequence-level information to precisely define the necessary changes in the genomes for breeding programs. This will be realized through the implementation of new DNA marker platforms and targeted breeding technologies, including genome editing. ■

The list of author affiliations is available in the full article online.
*Corresponding authors: Rudi Appels (rudi.appels@unimelb.edu.au); Kellye Eversole (eversole@eversoleassociates.com); Nils Stein (stein@ipk-gatersleben.de)
Cite this article as International Wheat Genome Sequencing Consortium, *Science* **361**, eaar7191 (2018). DOI: 10.1126/science.aar7191

RESEARCH ARTICLE

WHEAT GENOME

Shifting the limits in wheat research and breeding using a fully annotated reference genome

International Wheat Genome Sequencing Consortium (IWGSC)*

An annotated reference sequence representing the hexaploid bread wheat genome in 21 pseudomolecules has been analyzed to identify the distribution and genomic context of coding and noncoding elements across the A, B, and D subgenomes. With an estimated coverage of 94% of the genome and containing 107,891 high-confidence gene models, this assembly enabled the discovery of tissue- and developmental stage–related coexpression networks by providing a transcriptome atlas representing major stages of wheat development. Dynamics of complex gene families involved in environmental adaptation and end-use quality were revealed at subgenome resolution and contextualized to known agronomic single-gene or quantitative trait loci. This community resource establishes the foundation for accelerating wheat research and application through improved understanding of wheat biology and genomics-assisted breeding.

Wheat (*Triticum aestivum* L.), the most widely cultivated crop on Earth, contributes about a fifth of the total calories consumed by humans and provides more protein than any other food source (*1*, *2*). Breeders strive to develop improved varieties by fine-tuning genetically complex yield and end-use quality parameters while maintaining yield stability and regional adaptation to specific biotic and abiotic stresses (*3*). These efforts are limited, however, by insufficient knowledge and understanding of the molecular basis of key agronomic traits. To meet the demands of human population growth, there is an urgent need for wheat research and breeding to accelerate genetic gain while increasing wheat yield and protecting quality traits. In other plant and animal species, access to a fully annotated and ordered genome sequence, including regulatory sequences and genome-diversity information, has promoted the development of systematic and more time-efficient approaches for the selection and understanding of important traits (*4*). Wheat has lagged behind other species, primarily owing to the challenges of assembling a large (haploid genome, 1C = 16 Gb) (*5*), hexaploid, and complex genome that contains more than 85% repetitive DNA.

To provide a foundation for improvement through molecular breeding, the International Wheat Genome Sequencing Consortium (IWGSC) established a road map to deliver a high-quality reference genome sequence of the bread wheat cultivar Chinese Spring (CS). A chromosome survey sequence (CSS) intermediate product assigned 124,201 gene loci across the 21 chromosomes and revealed the evolutionary dynamics of the wheat genome through gene loss, gain, and duplication (*6*). The lack of global sequence contiguity and incomplete coverage (only 10 Gb were assembled), however, did not provide the wider regulatory genomic context of genes. Subsequent whole-genome assemblies improved contiguity (*7*–*9*) but lacked full annotation and did not resolve the intergenic space or present the genome in the correct physical order.

Here we report an ordered and annotated assembly (IWGSC RefSeq v1.0) of the 21 chromosomes of the allohexaploid wheat cultivar CS, an achievement that is built on a rich history of chromosome studies in wheat (*10*–*12*), which allowed the integration of genetic and genomic resources. The completeness and accuracy of IWGSC RefSeq v1.0 provide insights into global genome composition and enable the construction of complex gene coexpression networks to identify central regulators in critical pathways, such as flowering-time control. The ability to resolve the inherent complexity of gene families related to important agronomic traits demonstrates the impact of IWGSC RefSeq v1.0 on dissecting quantitative traits genetically and implementing modern breeding strategies for future wheat improvement.

Chromosome-scale assembly of the wheat genome

Pseudomolecule sequences representing the 21 chromosomes of the bread wheat genome were assembled by integrating a draft de novo whole-genome assembly (WGA), built from Illumina short-read sequences using NRGene deNovoMagic2 (Fig. 1A, Table 1, and tables S1 and S2), with additional layers of genetic, physical, and sequence data (tables S3 to S8 and figs. S1 and S2). In the resulting 14.5-Gb genome assembly, contigs and scaffolds with N50s of 52 kb and 7 Mb, respectively, were linked into superscaffolds (N50 = 22.8 Mb), with 97% (14.1 Gb) of the sequences assigned and ordered along the 21 chromosomes and almost all of the assigned sequence scaffolds oriented relative to each other (13.8 Gb, 98%). Unanchored scaffolds comprising 481 Mb (2.8% of the assembly length) formed the "unassigned chromosome" (ChrUn) bin. The quality and contiguity of the IWGSC RefSeq v1.0 genome assembly were assessed through alignments with radiation hybrid maps for the A, B, and D subgenomes [average Spearman's correlation coefficient (r) of 0.98], the genetic positions of 7832 and 4745 genotyping-by-sequencing derived genetic markers in 88 double haploid and 993 recombinant inbred lines (Spearman's r of 0.986 and 0.987, respectively), and 1.24 million pairs of neighbor insertion site–based polymorphism (ISBP) markers (*13*), of which 97% were collinear and mapped in a similar size range (difference of <2 kb) between the de novo WGA and the available bacterial artificial chromosome (BAC)–based sequence assemblies. Finally, IWGSC RefSeq v1.0 was assessed with independent data derived from coding and noncoding sequences, revealing that 99 and 98% of the previously known coding exons (*6*) and transposable element (TE)–derived (ISBP) markers (table S9), respectively, were present in the assembly. The approximate 1-Gb size difference between IWGSC RefSeq v1.0 and the new genome size estimates of 15.4 to 15.8 Gb (*14*) can be accounted for by collapsed or unassembled sequences of highly repeated clusters, such as ribosomal RNA coding regions and telomeric sequences.

A key feature distinguishing the IWGSC RefSeq v1.0 from previous draft wheat assemblies (*6*–*9*) is the long-range organization, with 90% of the genome represented in superscaffolds larger than 4.1 Mb and with each chromosome represented, on average, by only 76 superscaffolds (Table 1). The largest superscaffold spans 166 Mb, which is half the size of the rice (*Oryza sativa* L.) genome and is larger than the *Arabidopsis thaliana* L. genome (*15*, *16*). Moreover, the 21 pseudomolecules position molecular markers for wheat research and breeding [504 single-stranded repeats (SSRs), 3025 diversity array technologies (DArTs), 6689 expressed sequence tags (ESTs), 205,807 single-nucleotide polymorphisms (SNPs), and 4,512,979 ISBPs] (table S9), thus providing a direct link between the genome sequence and genetic loci and genes underlying traits of agronomic importance.

The composition of the wheat genome

Analyses of the components of the genome sequence revealed the distribution of key elements and enabled detailed comparisons of the homeologous A, B, and D subgenomes. Accounting for 85% of the genome, with a relatively equal distribution across the three subgenomes (Table 2), 3,968,974 copies of TEs belonging to 505 families were annotated. Many (112,744) full-length long terminal repeat (LTR)–retrotransposons were identified that have been difficult to define from short-read sequence assemblies (fig. S3). Although the TE content has been extensively rearranged

*All authors with their affiliations are listed at the end of this paper.

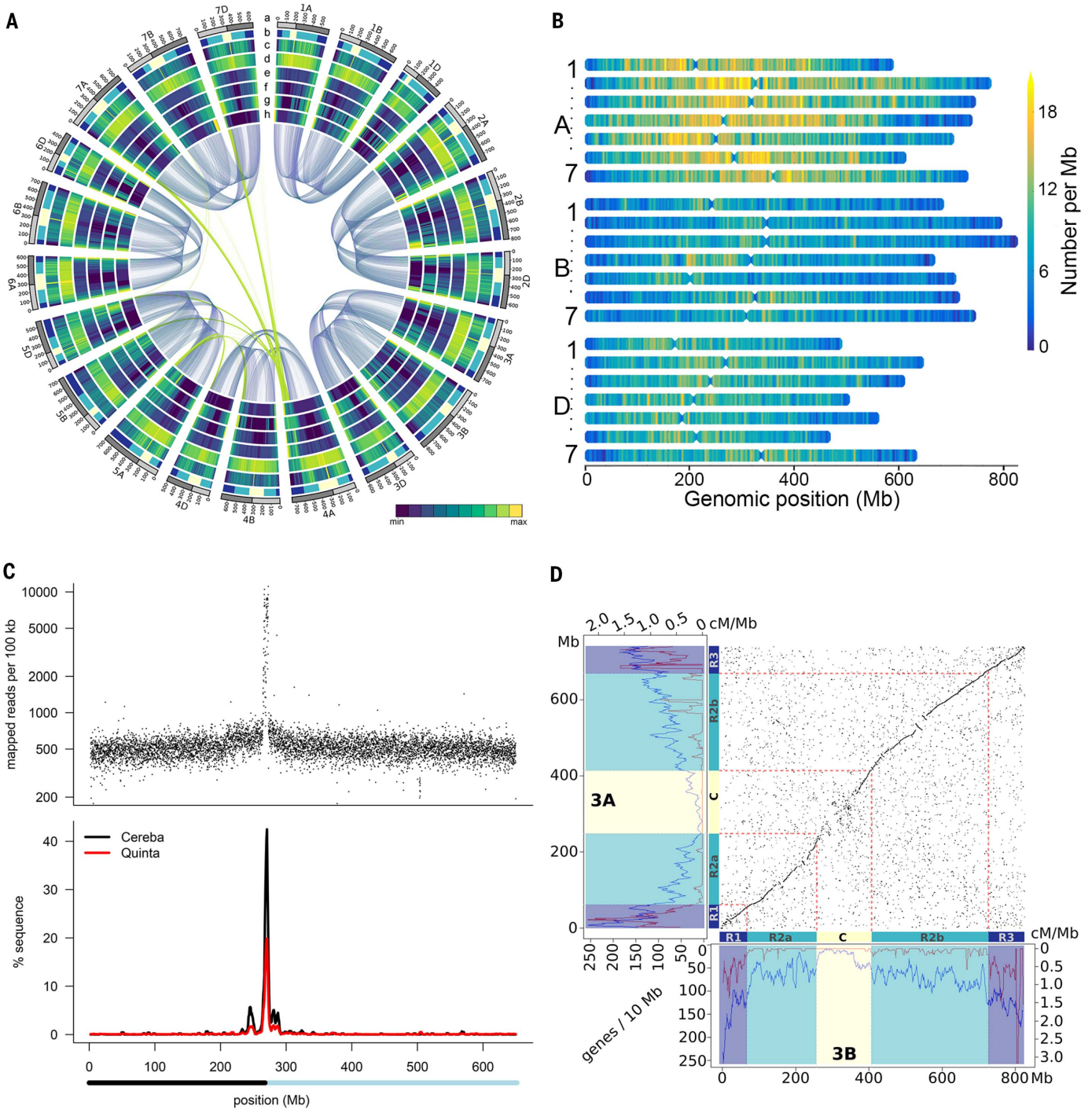


Fig. 1. Structural, functional, and conserved synteny landscape of the 21 wheat chromosomes. (**A**) Circular diagram showing genomic features of wheat. The tracks toward the center of the circle display (a) chromosome name and size (100-Mb tick size; light gray bar indicates the short arm and dark gray indicates the long arm of the chromosome); (b) dimension of chromosomal segments R1, R2a, C, R2b, and R3 [(*18*) and table S29]; (c) *K*-mer 20-frequencies distribution; (d) LTR-retrotransposons density; (e) pseudogenes density (0 to 130 genes per Mb); (f) density of HC gene models (0 to 32 genes per Mb); (g) density of recombination rate; and (h) SNP density. Connecting lines in the center of the diagram highlight homeologous relationships of chromosomes (blue lines) and translocated regions (green lines). (**B**) Distribution of Pfam domain PF08284 "retroviral aspartyl protease" signatures across the different wheat chromosomes. (**C**) Positioning of the centromere in the 2D pseudomolecule. Top panel shows density of CENH3 ChIP-seq data along the wheat chromosome. Bottom panel shows distribution and proportion of the total pseudomolecule sequence composed of TEs of the Cereba and Quinta families. The bar below the bottom panel indicates pseudomolecule scaffolds assigned to the short (black) or long (blue) arm on the basis of CSS data (*6*) mapping. (**D**) Dot-plot visualization of collinearity between homeologous chromosomes 3A and 3B in relation to distribution of gene density and recombination frequency (left and bottom panel boxes: blue and purple lines, respectively). Chromosomal zones R1, R2a, C, R2b, and R3 are colored as in (A). cM, centimorgan.

through rounds of deletions and amplifications since the divergence of the A, B, and D subgenomes about 5 million years ago, the TE families that shaped the Triticeae genomes have been maintained in similar proportions: 76% of the 165 TE families present in a cumulative length greater than 1 Mb contributed similar proportions (less than a twofold difference between subgenomes), and only 11 families, accounting for 2% of total TEs, showed a higher than threefold difference between two subgenomes (*17*). TE abundance accounts, in part, for the size differences between subgenomes—for example, 64% of the 1.2-Gb size difference between the B and D subgenomes can be attributed to lower gypsy retrotransposon content. Low-copy DNA content (primarily unclassified sequences) also varied between subgenomes, accounting, for example, for 97 Mb of the 245-Mb size difference between A and B subgenomes (fig. S4). As reported (*18*), no evidence was found for a major burst of transposition after polyploidization. The independent evolution in the diploid lineages was reflected in differences in the specific composition of the A, B, and D subgenomes at the subfamily (variants) level, as evidenced by subgenome-specific over-representation of individual transposon domain signatures (Fig. 1B). See (*17*) for a more detailed analysis of the TE content and its impact on the evolution of the wheat genome.

In addition to TEs, annotation of the intergenic space included noncoding RNAs. We identified eight new microRNA families (fig. S5 and table S10) and the entire complement of tRNAs (which showed an excess of lysine tRNAs, fig. S6). Around 8000 nuclear-inserted plastid DNA segments and 11,000 nuclear-inserted mitochondrial DNA segments representing, respectively, 5 and 17 Mb were also revealed by comparing the genome assembly with complete plastid and mitochondrial genomes assembled from the IWGSC RefSeq v1.0 raw read data (*14*).

Precise positions for the centromeres were defined by integrating Hi-C, CSS (*6*), and published chromatin immunoprecipitation sequencing (ChIP-seq) data for CENH3, a centromere-specific histone H3 variant (*19*). Clear ChIP-seq peaks were evident in all chromosomes and coincided with the centromere-specific repeat families (Fig. 1C, fig. S7, and table S11). CENH3 targets were also found in unassigned sequence scaffolds (ChrUn), indicating that centromeres of several chromosomes are not yet completely resolved. On the basis of these data, a conservative estimate for the minimal average size of a wheat centromere is 4.9 Mb (6.7 Mb, if including ChrUn; table S11), compared with an average centromere size of ~1.8 Mb in maize (*20*, *21*) and 0.4 to 0.8 Mb in rice (*22*).

Gene models were predicted with two independent pipelines previously utilized for wheat genome annotation and then consolidated to produce the RefSeq Annotation v1.0 (fig. S8). Subsequently, a set of manually curated gene models was integrated to build RefSeq Annotation v1.1 (fig. S9 and tables S12 to S17). In total, 107,891 high-confidence (HC) protein-coding loci were identified, with relatively equal distribution across the A, B, and D subgenomes (35,345, 35,643, and 34,212, respectively; Figs. 1D and 2A, fig. S10, and table S18). In addition, 161,537 other protein-coding loci were classified as low-confidence (LC) genes, representing partially supported gene models, gene fragments, and orphans (table S18). A predicted function was assigned to 82.1% (90,919) of HC genes in RefSeq Annotation v1.0 (tables S19 and S20), and evidence for transcription was found for 85% (94,114) of the HC genes versus 49% of the LC genes (*23*). Within the pseudogene category, 25,419 (8%) of 303,818 candidates matched LC gene models. The D subgenome contained significantly fewer pseudogenes than the A and B subgenomes (81,905 versus 99,754 and 109,097, respectively; χ^2 test, $P < 2.2 \times 10^{-16}$) (tables S21 and S22 and fig. S10). In ChrUn, 2691 HC and 675 LC gene models were identified.

The quality of the RefSeq Annotation v1.1 gene set was benchmarked against BUSCO v3 (*24*), representing 1440 Embryophyta near-universal single-copy orthologs and published annotated wheat gene sets (Fig. 2B and fig. S11). Of the BUSCO v3 genes, 99% (1436) were represented in at least one complete copy in RefSeq Annotation v1.1 and 90% (1292) in three complete copies, an improvement over the 25% (353) and 70% (1014) of BUSCO v3 genes that were identified in the IWGSC (*6*) and TGACv1 (*8*) gene sets, respectively (Fig. 2B). Improved contiguity of sequences in the immediate vicinity of genes was also found: 61% of the HC and LC genes were flanked by at least 10 kb of sequence without ambigous bases (Ns), in contrast to 37% and only 5% of the HC and LC genes in the TGACv1 and IWGSC CSS gene models, respectively (fig. S12).

To further characterize the gene space, a phylogenomic approach was applied to identify gene homeologs and paralogs between and within the wheat subgenomes and orthologs in other plant genomes (table S23 and figs. S13 to S15). Analysis of a subset of 181,036 genes ["filtered gene set," (*14*) and Table 3] comprising 103,757 HC and 77,279 LC genes identified 39,238 homeologous groups—that is, clades of A, B, and D subgenome orthologs deduced from gene trees—containing a total of 113,653 genes (63% of the filtered set). Gene losses or retention and gene gains (gene duplications) were determined for all homeologous loci of IWGSC RefSeq v1.0 (Table 3), assuming the presence of a single gene copy at

Table 1. Assembly statistics of IWGSC RefSeq v1.0.

Assembly characteristics	Values
Assembly size	14.5 Gb
Number of scaffolds	138,665
Size of assembly in scaffolds ≥ 100 kb	14.2 Gb
Number of scaffolds ≥ 100 kb	4,443
N50 contig length	51.8 kb
Contig L50 number	81,427
N90 contig length	11.7 kb
Contig L90 number	294,934
Largest contig	580.5 kb
Ns in contigs	0
N50 scaffold length	7.0 Mb
Scaffold L50 number	571
N90 scaffold length	1.2 Mb
Scaffold L90 number	2,390
Largest scaffold	45.8 Mb
Ns in scaffolds	261.9 Mb
Gaps filled with BAC sequences	183 (1.7 Mb)
Average size of inserted BAC sequence	9.5 kb
N50 superscaffold length	22.8 Mb
Superscaffold L50 number	166
N90 superscaffold length	4.1 Mb
Superscaffold L90 number	718
Largest superscaffold	165.9 Mb
Sequence assigned to chromosomes	14.1 Gb (96.8%)
Sequence ≥ 100 kb assigned to chromosomes	14.1 Gb (99.1%)
Number of superscaffolds on chromosomes	1,601
Number of oriented superscaffolds	1,243
Length of oriented sequence	13.8 Gb (95%)
Length of oriented sequence ≥ 100 kb	13.8 Gb (97.3%)
Smallest number of superscaffolds per subgenome chromosome	35 (7A), 68 (2B), 36 (1D)
Largest number of superscaffolds per subgenome chromosome	111 (4A), 176 (3B), 90 (3D)
Average number of superscaffolds per chromosome	76

every homeologous locus (referred to as a "triad"). The percentage of genes in homeologous groups for all configurations (ratios) is highly similar, hence balanced, across the three subgenomes: 63% (A), 61% (B), and 66% (D). The slightly higher percentage of homeologs in the D subgenome, together with the lower number of pseudogenes (table S22), is consistent with its more recent hybridization with the AABB tetraploid genome progenitor. Although most of the genes are present in homeologous groups, only 18,595 (47%) of the groups contained triads with a single gene copy per subgenome (an A:B:D configuration of 1:1:1). Of the groups of homeologous genes, 5673 (15%) exhibited at least one subgenome inparalog, that is, a gene copy resulting from a tandem or a segmental trans duplication (1:1:N A:B:D configuration; N indicates a minimum of one additional paralog per respective subgenome). The three genomes exhibited similar levels of loss of individual homeologs, affecting 10.7% (0:1:1), 10.3% (1:0:1), and 9.5% (1:1:0) of the homeologous groups in the A, B, and D subgenomes, respectively (Table 3 and tables S24 and S25).

Of the 67,383 (37%) genes of the filtered set not present in homeologous groups, 31,140 genes also had no orthologs in species included in the comparisons outside of bread wheat and mainly comprised gene fragments, non–protein-coding loci with open reading frames, or other gene-calling artifacts. The remaining 36,243 genes had homologs outside of bread wheat and appeared to be subgenome specific (Table 3). Two of the genes in this category were *granule bound starch synthase* (*GBSS*) on chromosome 4A (1:0:0, a gene that is a key determinant of udon noodle quality) and *ZIP4* within the *pairing homeologous 1* (*Ph1*) locus on chromosome 5B [0:1:0, a locus critical for the diploid meiotic behavior of the wheat homeologous chromosomes (*25*)]. The phylogenomic analysis indicated that the *GBSS* on 4A is a divergent translocated homeolog originally located on chromosome 7B (fig. S16), whereas *ZIP4* is a transduplication of a chromosome 3B locus (table S26). Both genes confer important properties on wheat and illustrate the diversity in origin and function of gene models that are not in a 1:1:1 configuration. No evidence was found for biased partitioning. Rather, our analyses support gradual gene loss and gene movement among the subgenomes that may have occurred in either the diploid progenitor species or the tetraploid ancestor or following the final hexaploidization event in modern bread wheat (Table 3 and figs. S24 and S25). Together with the equal contribution of the three homeologous genomes to the overall gene expression (*23*), this demonstrates the absence of subgenome dominance (*26*).

Of the bread wheat HC genes, 29,737 (27%) are present as tandem duplicates, which is up to 10% higher than that found for other monocotyledonous species (table S27). Tandemly repeated genes are most prevalent in the B subgenome (29%), contributing to its higher gene content and larger number of 1:N:1 homeologous groups (Table 3). The postulated hybrid origin of the D subgenome, as a result of interspecific crossing with AABB tetraploid genome progenitors 1 to 2 million years after they diverged (*27*), is consistent with the synonymous substitution rates of homeologous gene pairs (fig. S17). Homeologous groups with gene duplicates in at least one subgenome (1:1:N, 1:N:1, or N:1:1) showed elevated evolutionary rates (for the subgenome carrying the duplicate) as compared with strict 1:1:1 or 1:1 groups (figs. S18 to S22). Homeologs with recent duplicates also showed higher levels of expression divergence (fig. S23), consistent with gene and genome duplications acting as a driver of functional innovation (*28*, *29*).

Analysis of synteny between the seven triplets of homeologous chromosomes showed high levels

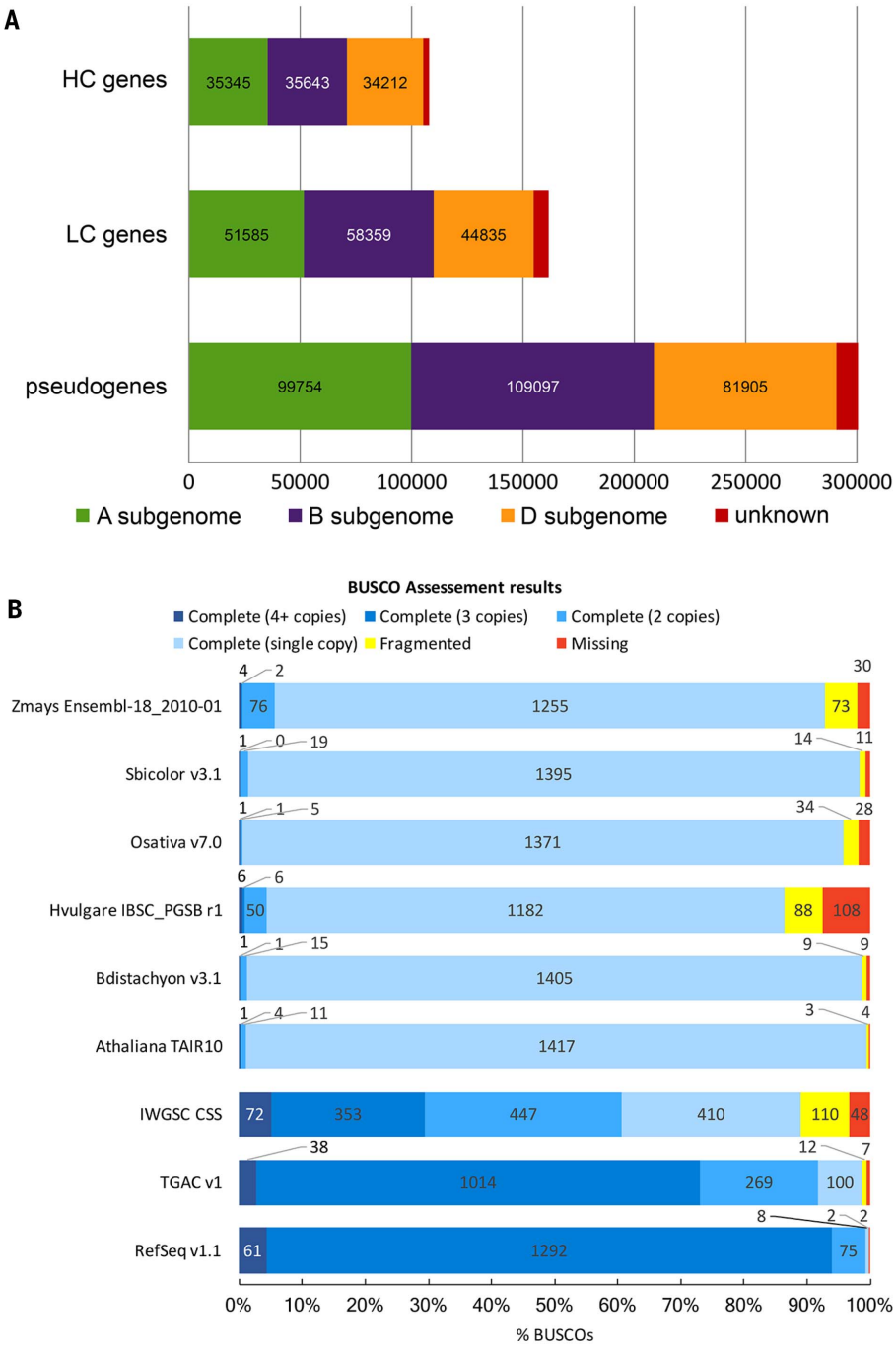


Fig. 2. Evaluation of automated gene annotation. (**A**) Selected gene prediction statistics of IWGSC RefSeq Annotation v1.1, including number and subgenome distribution of HC and LC genes as well as pseudogenes. (**B**) BUSCO v3 gene model evaluation comparing IWGSC RefSeq Annotation v1.1 to earlier published bread wheat whole-genome annotations, as well as to annotations of related grass reference-genome sequences. BUSCO provides a measure for the recall of highly conserved gene models.

of conservation. There was no evidence that any major rearrangements occurred since the A, B, and D subgenomes diverged ~5 million years ago (Fig. 1D), although collinearity between homeologs was disturbed by inversions occurring, on average, every 74.8 Mb, involving blocks of 10 genes or more (mean gene number of 48.2 with a mean size of 10.5 Mb) (Fig. 1D and table S28). Macrosynteny was conserved across centromere (C) regions, but collinearity (microsynteny) broke down specifically in these recombination-free, gene-poor regions for all seven sets of homeologous chromosomes (Fig. 1D, figs. S24 to S26, and table S29). Of the 113,653 homeologous genes, 80% (90,232) were found organized in macrosynteny, that is, still present at their ancestral position (table S24). At the microsynteny scale, 72% (82,308) of the homeologs were organized in collinear blocks, that is, intervals with a highly conserved gene order (Fig. 1D). A higher proportion of syntenic genes was found in the interstitial regions [short arm, R2a (*18*), 46% and long arm, R2b (*18*), 61%] than in the distal telomeric [short arm, R1 (*18*), 39% and long arm, R3 (*18*), 51%] and centromere regions [C (*18*), 29%], and the interstitial compartments harbored larger syntenic blocks (figs. S27 and S28). The higher proportions of duplicated genes in distal-terminal regions (34 and 27% versus 13 to 15% in the other regions; fig. S29) exerted a strong influence on the decay of syntenic block size and contributed to the higher sequence variability in these regions. Overall, distal chromosomal regions are the preferential targets of meiotic recombination and the fastest evolving compartments. As such, they represent the genomic environment for creating sequence, hence allelic, diversity, providing the basis for adaptability to changing environments.

Atlas of transcription reveals trait-associated gene co-regulation networks

The gene annotation, coupled with identification of homeologs and paralogs in IWGSC RefSeq v1.0, provides a resource to study gene expression in genome-wide and subgenome contexts. A total of 850 RNA-seq samples derived from 32 tissues at different growth stages and/or challenged by different stress treatments were mapped to RefSeq Annotation v1.0 (Fig. 3A, database S1, and tables S30 to S32). Expression was observed for 94,114 (84.9%) HC genes (fig. S30) and for 77,920 (49.1%) LC genes, the latter showing lower expression breadth and level [median six tissues; average 2.9 transcripts per million (tpm)] than the HC genes (median 20 tissues, average 8.2 tpm) (fig. S31). This correlated with the higher average methylation status of LC genes (figs. S32 and S33). A principal component analysis identified tissue (Fig. 3B), rather than growth stage or stress (fig. S34), as the main factor driving differential expression between samples, consistent with studies in other organisms (*30*–*33*). Of the total number of genes, 31.0% are expressed in more than 90% of tissues (average 16.9 tpm, ≥30 tissues), and 21.5% are expressed in 10% or fewer tissues (average 0.22 tpm, ≤3 tissues; fig. S31).

Of the HC genes, 8231 showed tissue-exclusive expression (fig. S35). About half of these were associated with reproductive tissues (microspores, anther, and stigma or ovary), consistent with observations in rice (*34*). The tissue-exclusive genes were enriched for response to extracellular stimuli and reproductive processes (database S2). By contrast, 23,146 HC genes expressed across all 32 tissues were enriched for biological processes associated with housekeeping functions such as protein translation and protein metabolic processes. Tissue-specific genes were shorter [1147 ± 8 base pairs (bp)], had fewer exons (2.76 ± 0.3), and were expressed at lower levels (3.4 ± 0.1 tpm) compared with ubiquitous genes (1429 ± 7 bp, 7.87 ± 0.4 exons, and 17.9 ± 0.4 tpm) (fig. S35).

Genes located in distal regions R1 and R3 (fig. S25 and table S29) showed lower expression breadth than those in the proximal regions (15.7 and 20.7 tissues, respectively) (Fig. 3C and fig. S36). This correlated with enrichment of Gene Ontology (GO) slim terms such as "cell cycle," "translation," and "photosynthesis" for genes in the proximal regions, whereas genes enriched for "response to stress" and "external stimuli" were found in the highly recombinant distal R1 and R3 regions (database S3, fig. S36, and table S33). The expression breadth pattern was also correlated with the distribution of the repressive H3K27me3 (trimethylated histone H3 lysine 27) (Pearson $r = -0.76$, $P < 2.2 \times 10^{-16}$) and with the active H3K36me3 and H3K9ac (acetylated H3K9) (Pearson $r = 0.9$ and 0.83, respectively; $P < 2.2 \times 10^{-16}$) histone marks (fig. S37).

Global patterns of coexpression (*35*) were determined with a weighted gene coexpression network analysis (WGCNA) on 94,114 expressed HC genes. Of these genes, 58% (54,401) could be assigned to 38 modules (Fig. 3D and database S4), and, consistent with the principal component analysis, tissues were the major driver of module identity (Fig. 3D and figs. S38 to S40). The analysis focused initially on the 9009 triads (syntenic and nonsyntenic) with a 1:1:1 A:B:D relationship and for which all homeologs were assigned to a module. Of the triads, 16.4% had at least one homeolog in a divergent module, with the B homeolog most likely to be divergent (37.4% B-divergent versus 31.7% A-divergent and 30.9% D-divergent triads, χ^2 test $P = 0.007$). However, the expression profiles of most (83.6%) of the triads were relatively consistent with all homeologs in the same (57.6%) or a closely related (26.0%) module. The proportion of homeologs found within the same module was higher than expected, pointing to a highly conserved expression pattern of homeologs across the 850 RNA-seq samples (Fig. 3E and table S34). Triads with at least one gene in a nonsyntenic position had a higher amount of

Table 2. Relative proportions of the major elements of the wheat genome. Proportions of TEs are given as the percentage of sequences assigned to each superfamily relative to genome size. Abbreviations in parentheses under the headings "Class 1" and "Class 2" indicate transposon types.

Major elements	Wheat subgenome			
	AA	BB	DD	Total
Assembled sequence assigned to chromosomes (Gb)	4.935	5.180	3.951	14.066
Size of TE-related sequences (Gb)	4.240	4.388	3.285	11.913
TEs (%)	85.9	84.7	83.1	84.7
Class 1				
LTR-retrotransposons				
Gypsy (RLG)	50.8	46.8	41.4	46.7
Copia (RLC)	17.4	16.2	16.3	16.7
Unclassified LTR-retrotransposons (RLX)	2.6	3.5	3.7	3.2
Non-LTR-retrotransposons				
Long interspersed nuclear elements (RIX)	0.81	0.96	0.93	0.90
Short interspersed nuclear elements (SIX)	0.01	0.01	0.01	0.01
Class 2				
DNA transposons				
CACTA (DTC)	12.8	15.5	19.0	15.5
Mutator (DTM)	0.30	0.38	0.48	0.38
Unclassified with terminal inverted repeats	0.21	0.20	0.22	0.21
Harbinger (DTH)	0.15	0.16	0.18	0.16
Mariner (DTT)	0.14	0.16	0.17	0.16
Unclassified class 2	0.05	0.08	0.05	0.06
hAT (DTA)	0.01	0.01	0.01	0.01
Helitrons (DHH)	0.0046	0.0044	0.0036	0.0042
Unclassified repeats	0.55	0.85	0.63	0.68
Coding DNA	0.89	0.89	1.11	0.95
Unannotated DNA	13.2	14.4	15.7	14.4
(Pre)-microRNAs	0.039	0.057	0.046	0.047
tRNAs	0.0056	0.0050	0.0068	0.0057

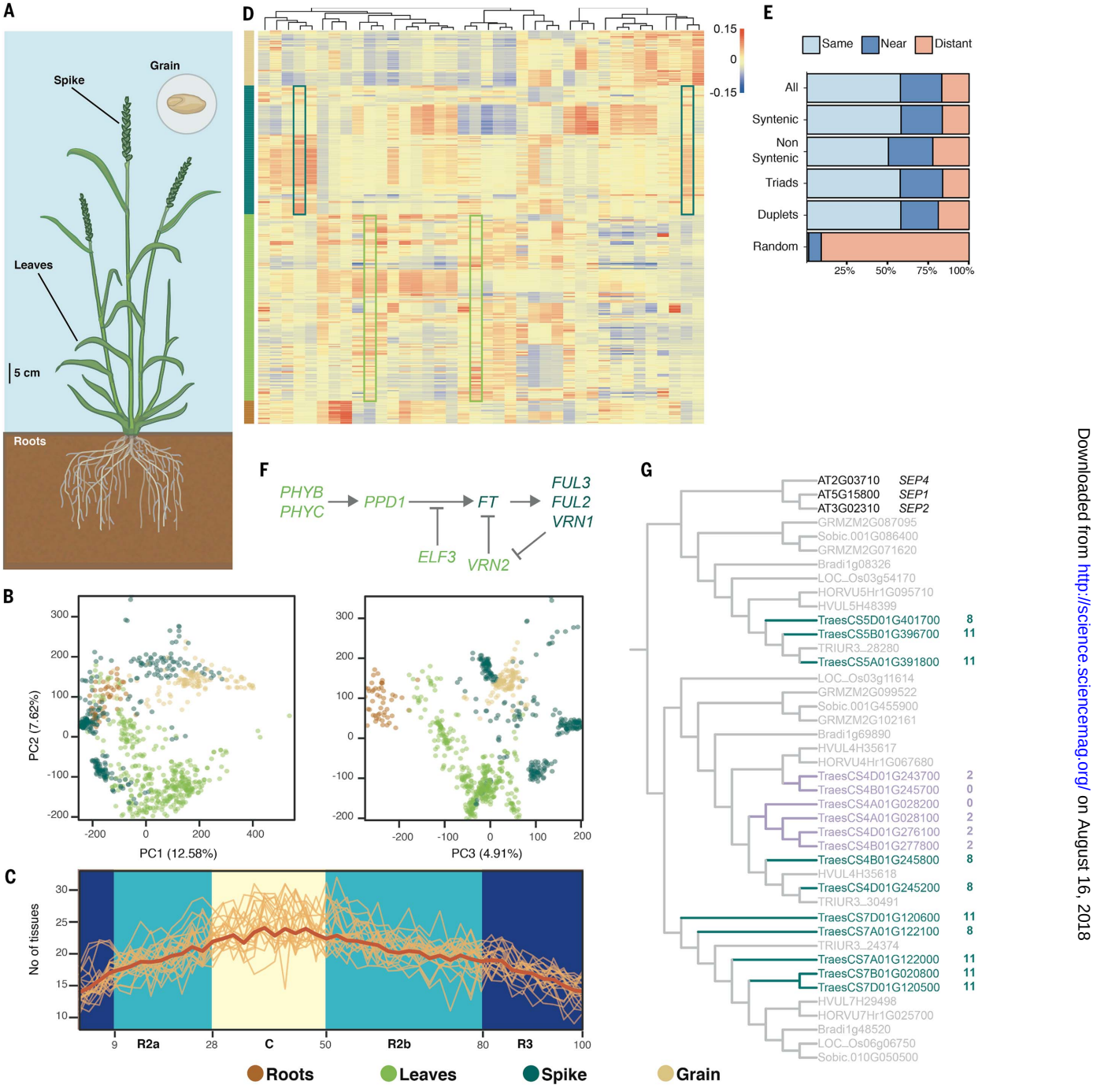


Fig. 3. Wheat atlas of transcription. (**A**) Schematic illustration of a mature wheat plant and high-level tissue definitions for "roots," "leaves," "spike," and "grain" used in the further analysis. (**B**) Principal component (PC) analysis plots for similarity of overall transcription, with samples colored according to their high-level tissue of origin [as introduced in (A)]. The color key for tissue is shown at the bottom of the figure under (C). (**C**) Chromosomal distribution of the average expression breadth [number of tissues in which genes are expressed (total number of tissues, $n = 32$)]. The average (dark orange line) is calculated on the basis of a scaled position of each gene within the corresponding genomic compartment (blue, aqua, and light yellow background) across the 21 chromosomes (orange lines). (**D**) Heatmap illustrating the expression of a representative gene (eigengene) for the 38 coexpression modules defined by WGCNA. Modules are represented as columns, with the dendrogram illustrating eigengene relatedness. Each row represents one sample. Colored bars to the left indicate the high-level tissue of origin; the color key is shown at the bottom of the figure under (C). DESeq2-normalized expression levels are shown. Modules 1 and 5 (light green boxes) were most correlated with high-level leaf tissue, whereas modules 8 and 11 (dark green boxes) were most correlated with spike. (**E**) Bar plot of module assignment (same, near, or distant) of homeologous triads and duplets in the WGCNA network. (**F**) Simplified flowering pathway in polyploid wheat. Genes are colored according to their assignment to leaf (light green)– or spike (dark green)–correlated modules. (**G**) Excerpt from phylogenetic tree for MADS transcription factors, including known *Arabidopsis* flowering regulators *SEP1*, *SEP2*, and *SEP4* (black) (for the full phylogenetic tree, see fig. S38). Green branches represent wheat orthologs of modules 8 and 11, whereas purple branches are wheat orthologs assigned to other modules (0 and 2). Gray branches indicate non-wheat genes.

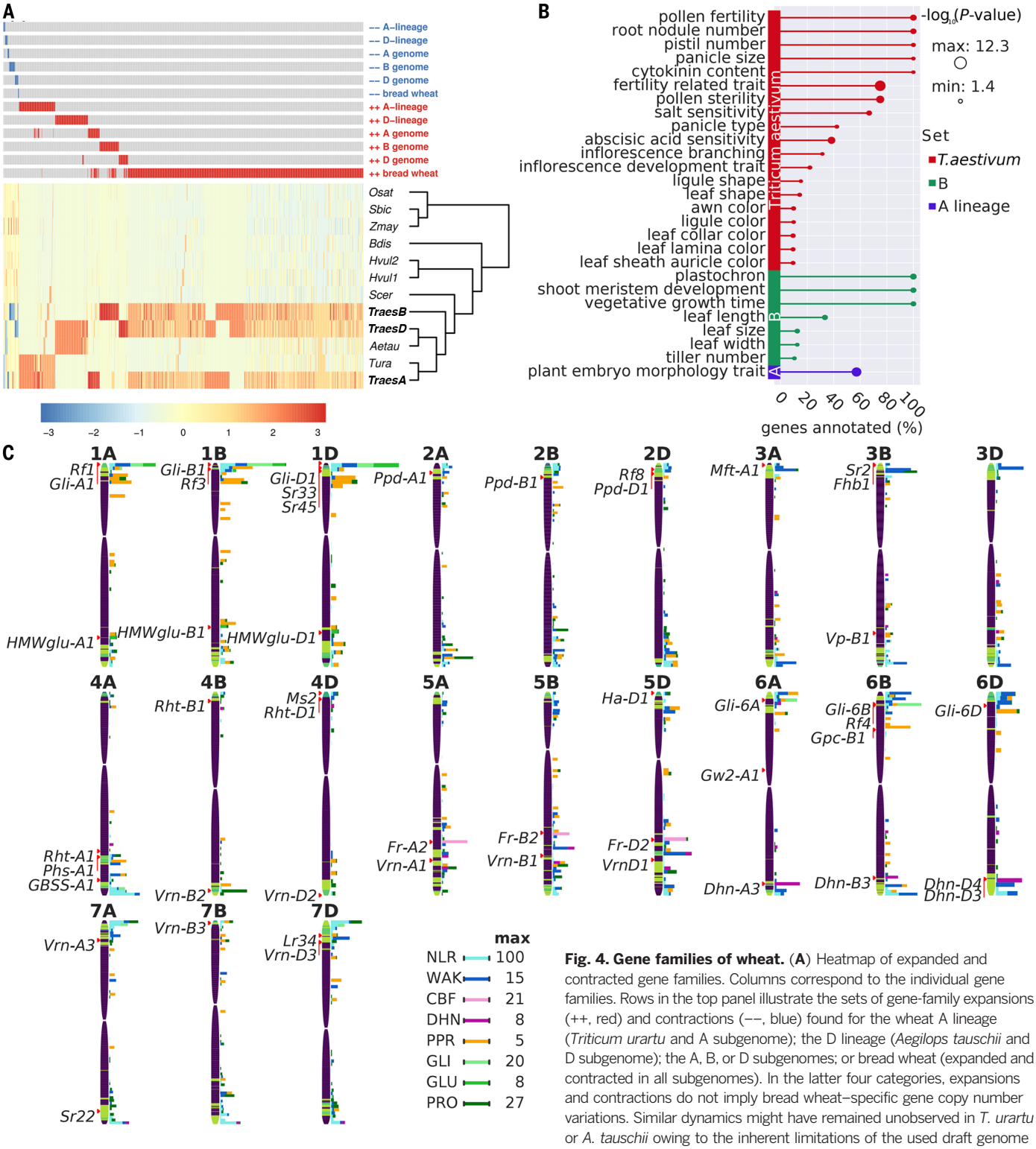


Fig. 4. Gene families of wheat. (**A**) Heatmap of expanded and contracted gene families. Columns correspond to the individual gene families. Rows in the top panel illustrate the sets of gene-family expansions (++, red) and contractions (−−, blue) found for the wheat A lineage (*Triticum urartu* and A subgenome); the D lineage (*Aegilops tauschii* and D subgenome); the A, B, or D subgenomes; or bread wheat (expanded and contracted in all subgenomes). In the latter four categories, expansions and contractions do not imply bread wheat–specific gene copy number variations. Similar dynamics might have remained unobserved in *T. urartu* or *A. tauschii* owing to the inherent limitations of the used draft genome assemblies (*53*, *54*). Rows in the bottom panel heatmap (color scheme on *z*-score scale) indicate the fold expansion and contraction of gene families for the taxa and species included in the analysis [*Oryza sativa* (Osat), *Sorghum bicolor* (Sbic), *Zea mays* (Zmay), *Brachypodium distachyon* (Bdis), *Hordeum vulgare* (Hvul1/2), *Secale cereale* (Scer), *A. tauschii* (Aetau), *T. urartu* (Tura), and wheat A (TraesA), B (TraesB), and D (TraesD) subgenomes]. (**B**) All enriched TO terms for the gene families depicted in (A). Overrepresented TO terms were found for expanded families in bread wheat (all subgenomes, red), the B subgenome (green), and the A lineage (*T. urartu* and A subgenome, blue) only, respectively. The *x* axis represents the percentage of genes annotated with the respective TO term that were contained in the gene set in question. The size of the bubbles corresponds to the *P* ($-\log_{10}$) significance of expansion. (**C**) Genomic distribution of gene families associated with adaptation to biotic (light and dark blue) or abiotic stress (light and dark pink), RNA metabolism in organelles and male fertility (orange), or end-use quality (light, medium, and dark green). Known positions of agronomically important genes and loci are indicated by red arrows and arrowheads to the left of the chromosome bars. Recombination rates are displayed as heatmaps in the chromosome bars [7.2 cM/Mb (light green) to 0 cM/Mb (black)].

divergent expression patterns compared to syntenic triads (21.2 versus 16.2%, χ^2 test $P < 0.001$) and fewer such triads shared all homeologs in the same module (48.7%) compared to syntenic triads (58.0%, chi-square test $P = 0.009$). Similar patterns were observed in the 1933 duplets that have a 1:1 relationship between only two homeologs (table S34). These results are consistent with syntenic homeologs showing similar expression patterns, whereas more dramatic changes in chromosome context associate with divergent expression and possible sub- or neofunctionalization. These trends were also found across diverse tissue-specific networks (*23*).

To explore the potential of the WGCNA network for identifying previously uncharacterized pathways in wheat, a search was undertaken for modules containing known regulators of wheat flowering time [e.g., *PPD1* (*36*) and *FT* (*37*); Fig. 3F]. Genes belonging to this pathway were grouped into specific modules. The upstream genes (*PHYB*, *PHYC*, *PPD1*, *ELF3*, and *VRN2*) were present mainly in modules 1 and 5 and were most highly correlated with expression in leaf and shoot tissues (0.68 and 0.67, respectively; adjusted $P < 1 \times 10^{-108}$). By contrast, the integrating gene *FT* and downstream genes *VRN1*, *FUL2*, and *FUL3* were found in modules 8 and 11, most highly correlated with expression in spikes (0.69 and 0.65, respectively; adjusted $P < 1 \times 10^{-101}$; table S35). The MADS_II transcription factor family that is generally associated with the above pathways was examined more closely, with a focus on the gene tree OG0000041, which contains 54 of the 118 MADS_II genes in wheat. Twenty-four MADS_II genes from modules 8 and 11 were identified within this gene tree, clustering into two main clades along with *Arabidopsis* and rice orthologs associated with floral patterning (fig. S41 and database S5). Within these clades, other MADS_II genes were found that were not in modules 8 or 11 (Fig. 3G), indicating a different pattern of coexpression. None of the 24 MADS_II genes had a simple 1:1 ortholog in *Arabidopsis*, suggesting that some wheat orthologs function in flowering (those within modules 8 and 11), whereas others could have developed different functions, despite being phylogenetically closely related. Thus, these data provide a framework to identify and prioritize the most likely functional orthologs of known model system genes within polyploid wheat, to characterize them functionally (*38*), and to dissect genetic factors controlling important agronomic traits (*39*, *40*). A more detailed analysis of tissue-specific and stress-related networks (*23*) provides a framework for defining quantitative variation and interactions between homeologs for many agronomic traits (*41*).

Gene-family expansion and contraction with relevance to wheat traits

Gene duplication and gene-family expansion are important mechanisms of evolution and environmental adaptation, as well as major contributors to phenotypic diversity (*42*, *43*). In a phylogenomic comparative analysis, wheat gene-family size and wheat-specific gene-family expansion and contraction were benchmarked against nine other grass genomes, including five closely related diploid Triticeae species (table S23 and figs. S13 to S15 and S42). A total of 30,597 gene families (groups of orthologous genes traced to a last common ancestor in the evolutionary hierarchy of the compared taxa) were defined, with 26,080

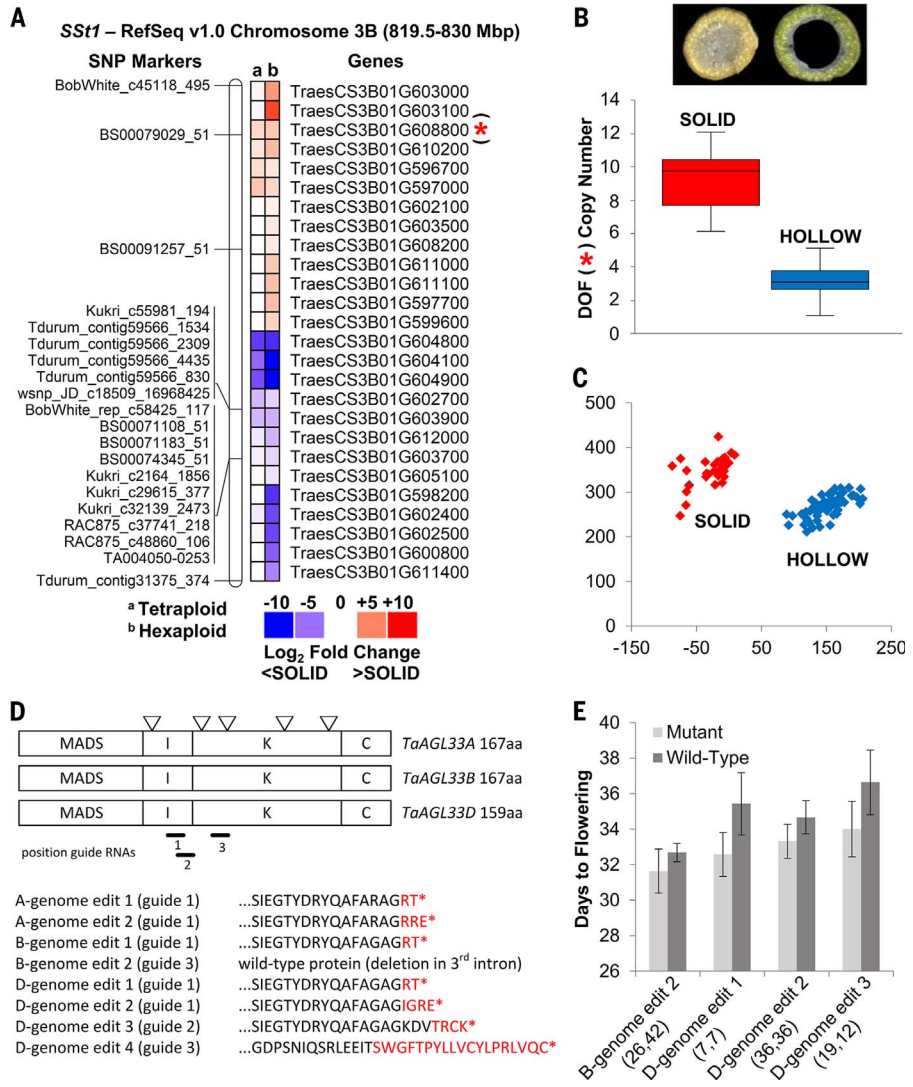


Fig. 5. IWGSC RefSeq v1.0–guided dissection of *SSt1* and *TaAGL33*. (**A**) The Lillian-Vesper population genetic map was anchored to IWGSC RefSeq v1.0 (left), and differentially expressed genes were identified between solid- and hollow-stemmed lines of hexaploid (bread) and tetraploid (durum) wheat (right). (**B**) Cross-sectioned stems of Lillian (solid) and Vesper (hollow) are shown as a phenotypic reference (top). Increased copy number of *TraesCS3B01G608800* [annotated as a DOF (DNA-binding one-zinc finger) transcription factor] is associated with stem phenotypic variation (bottom). (**C**) A high-throughput SNP marker tightly linked to *TraesCS3B01G608800* reliably discriminates solid- from hollow-stemmed wheat lines. Relative intensity of the fluorophores (FAM and HEX) used in KASPar analysis are shown. Vertical axis shows FAM signal; horizontal axis shows HEX signal. (**D**) Schematic of the three TaAGL33 proteins, showing the typical MADS, I, K, and C domains. Triangles indicate the position of the five introns that occur in all three homeologs. Bars indicate the position of single-guide RNAs designed for exons 2 and 3. Three T-DNA vectors—each containing the *bar* selectable marker gene, CRISPR nuclease, and one of three single-guide RNA sequences—were used for *Agrobacterium*-mediated wheat transformation, essentially as described earlier (*55*). Transgenic plants were obtained with edits at the targeted positions in all TaAGL33 homeologs. The putatively resulting protein sequence is displayed starting close to the edits, with wild-type amino acids (aa) in black font and amino acids resulting from the induced frame shifts in red font. * indicates premature termination codons. (**E**) Mean days to flowering (after 8 weeks of vernalization) for progeny of four homozygous edited plants (light gray bars) and the respective homozygous wild-type segregants (dark gray bars). Numbers in parentheses refer to the number of edited and wild-type plants examined, respectively. Error bars display SEM. Growth conditions were as described in (*50*).

families containing gene members from at least one of the three wheat subgenomes (tables S36 to S39). Among the 8592 expanded wheat gene families (33% of all families), 6216 were expanded in all three A, B, and D subgenomes (24%; either shared with the wild ancestor or specific to bread wheat, Fig. 4A). Another 1109 were expanded in only one of the wheat subgenomes, and 2102 gene families were expanded in either the A or the D genome lineages (Fig. 4A, fig. S43, and table S36). Overall, only 78 gene families were contracted in wheat. The number of gene families that are only expanded in wheat may be overestimated owing to limited completeness of the draft progenitor wheat genome assemblies used in this study (*14*) (table S39). Gene Ontology (GO; ontology of biomedical terms for the areas "cellular component," "biological process," and "molecular function"), Plant Ontology (PO; ontology terms describing anatomical structures and growth and developmental stages across Viridiplantae), and Plant Trait Ontology [TO; ontology of controlled vocabulary to describe phenotypic traits and quantitative trait loci (QTLs) that were physically mapped to a gene in flowering plant species] analyses identified 1169 distinct GO, PO, and TO terms (15% of all assigned terms) enriched in genes belonging to expanded wheat gene families (Fig. 4B and figs. S44 and S45). "A-subgenome" or "A-lineage" expanded gene families showed a bias for terms associated with seed formation [overrepresentation of the TO term "plant embryo morphology" (TO:0000064) and several seed, endosperm, and embryo-developmental GO terms] (fig. S46). Similarly, "B-subgenome" expanded gene families were enriched for TO terms related to plant vegetative growth and development (database S6 and fig. S47). Gene families that were expanded in all wheat subgenomes were enriched for 14 TO terms associated with yield-affecting morphological traits and five terms associated with fertility and abiotic-stress tolerance (Fig. 4B), which was also mirrored by enrichment for GO and PO terms associated with adaptation to abiotic stress ("salt stress" and "cold stress") and grain yield and quality ("seed maturation," "dormancy," and "germination"). The relationship between the patterns of enriched TO, PO, and GO terms for expanded wheat gene families and key characteristics of wheat performance (figs. S45 to S51) provides a resource (database S6) to explore future QTL mapping and candidate gene identification for breeding.

Many gene families with high relevance to wheat breeding and improvement were among the expanded group, and their genomic distribution was analyzed in greater detail (Fig. 4C and figs. S52 to S54). Disease resistance–related NLR (nucleotide-binding site leucine-rich repeat)–like loci and WAK (wall-associated receptor)–like genes were clustered in high numbers at the distal (R1 and R3) regions of all chromosome arms, with NLRs often co-localizing with known disease resistance loci (Fig. 4C). The restorer-of-fertility–like (RFL) subclade of P class pentatricopeptide repeat (PPR) proteins, potentially of interest for hybrid wheat production, comprised 207 genes, nearly threefold more per haploid subgenome than have been identified in any other plant genome analyzed to date (*44*, *45*). They localized mainly as clusters of genes in regions on the group 1, 2, and 6 chromosomes, which carry fertility-restoration QTLs in wheat (Fig. 4C and fig. S54). Within the dehydrin gene family, implicated with drought tolerance in plants, 25 genes that formed well-defined clusters on chromosomes 6A, 6B, and 6D (figs. S53 and S55) showed early increased expression under severe drought stress. As the structural variation in the *CBF* genes of wheat is known to be associated with winter survival (*46*), the array of *CBF* paralogs at the *Fr-2* locus (fig. S56) revealed by IWGSC RefSeq v1.0 provides a basis for targeted allele mining for previously uncharacterized *CBF* haplotypes from highly frost-tolerant wheat genetic resources. Lastly, high levels of expansion and variation in members of grain prolamin gene families [fig. S52 and (*47*)] either related to the response to heat stress or whose protein epitopes are associated with levels of celiac disease and food allergies (*47*) provide candidates for future selection in breeding programs. From these few examples, it is evident that flexibility in gene copy numbers within the wheat genome has contributed to the adaptability of wheat to produce high-quality grain in diverse climates and environments (*48*). Knowledge of the complex picture of the genome-wide distribution of gene families (Fig. 4C), which needs to be considered for selection in breeding programs in the context of

Table 3. Groups of homeologous genes in wheat. Homeologous genes are "subgenome orthologs" and were inferred by species tree reconciliation in the respective gene family. Numbers include both HC and LC genes filtered for TEs (filtered gene set). Conserved subgenome-specific (orphan) genes are found only in one subgenome but have homologs in other plant genomes used in this study. This includes orphan outparalogs resulting from ancestral duplication events and conserved only in one of the subgenomes. Nonconserved orphans are either singletons or duplicated in the respective subgenome, but neither have obvious homologs in the other subgenomes or the other plant genomes studied. Microsynteny is defined as the conservation and collinearity of local gene ordering between orthologous chromosomal regions. Macrosynteny is defined as the conservation of chromosomal location and identity of genetic markers like homeologs but may include the occurrence of local inversions, insertions, or deletions. Additional data are presented in table S24.

Homeologous group (A:B:D)	Number in wheat genome	Composition of groups (%)	Number of genes in A	Number of genes in B	Number of genes in D	Total number of genes
1:1:1	21,603	55.1	21,603	21,603	21,603	64,809
1:1:N	644	1.6	644	644	1,482	2,770
1:N:1	998	2.5	998	2,396	998	4,392
N:1:1	761	1.9	1,752	761	761	3,274
1:1:0	3,708	9.5	3,708	3,708	0	7,416
1:0:1	4,057	10.3	4,057	0	4,057	8,114
0:1:1	4,197	10.7	0	4,197	4,197	8,394
Other ratios	3,270	8.3	4,999	5,371	4,114	14,484
1:1:1 in microsynteny	18,595	47.4	18,595	18,595	18,595	55,785
Total in microsynteny	30,339	77.3	27,240	27,063	28,005	82,308
1:1:1 in macrosynteny	19,701	50.2	19,701	19,701	19,701	59,103
Total in macrosynteny	32,591	83.1	29,064	30,615	30,553	90,232
Total in homeologous groups	**39,238**	**100.0**	**37,761**	**38,680**	**37,212**	**113,653**
Conserved subgenome orphans			12,412	12,987	10,844	36,243
Nonconserved subgenome singletons			10,084	12,185	8,679	30,948
Nonconserved subgenome duplicated orphans			71	83	38	192
Total (filtered)			**60,328**	**63,935**	**56,773**	**181,036**

distribution of recombination and allelic diversity, can now be applied to wheat improvement strategies. This is especially true if "must-have traits" that are allocated in chromosomal compartments with highly contrasting characteristics are fixed in repulsion or are found only in incompatible gene pools of the respective breeding germplasm.

Rapid trait improvement using physically resolved markers and genome editing

The selection and modification of genetic variation underlying agronomic traits in breeding programs is often complicated if phenotypic selection depends on the expression of multiple loci with quantitative effects that can be strongly influenced by the environment. This dilemma can be overcome if DNA markers in strong linkage disequilibrium with the phenotype are identified through forward genetic approaches or if the underlying genes can be targeted through genome editing. The potential for IWGSC RefSeq v1.0, together with the detailed genome annotation, to accelerate the identification of potential candidate genes underlying important agronomic traits was exemplified for two targets. A forward genetics approach was used to fully resolve a QTL for stem solidness (*SSt1*) conferring resistance to drought stress and insect damage (*49*) that was disrupted in previous wheat assemblies by a lack of scaffold ordering and annotation, partial assembly, and/or incomplete gene models (fig. S57 and tables S40 and S41). In IWGSC RefSeq v1.0, *SSt1* contains 160 HC genes (table S42), of which 26 were differentially expressed (DESeq2, Benjamini-Hochberg adjusted $P < 0.01$) between wheat lines with contrasting phenotypes. One of the differentially expressed genes, *TraesCS3B01G608800*, was present as a single copy in IWGSC RefSeq v1.0 but showed copy number variation associated with stem solidness in a diverse panel of hexaploid cultivars (Fig. 5A, fig. S58, and table S43). Using IWGSC RefSeq v1.0, we developed a diagnostic SNP marker physically linked to the copy number variation that has been deployed to select for stem solidness in wheat breeding programs (Fig. 5B).

Knowledge from model species can also be used to annotate genes and provide a route to trait enhancement through reverse genetics. The approach here targeted flowering time, which is important for crop adaptation to diverse environments and is well studied in model plants. Six wheat homologs of the *FLOWERING LOCUS C* (*FLC*) gene have been identified as having a role in the vernalization response, a critical process regulating flowering time (*50*). IWGSC RefSeq v1.0 was used to refine the annotation of these six sequences to identify four HC genes and then to design guide RNAs to specifically target, with CRISPR-Cas9–based gene editing, one of these genes, *TaAGL33*, on all subgenomes [*TraesCS3A01G435000* (A), *TraesCS3B01G470000* (B), and *TraesCS3D01G428000* (D)] [Fig. 5C and (*14*)]. Editing was obtained at the targeted gene and led to truncated proteins after the MADS box through small deletions and insertions (Fig. 5D). Expression of all homeologs was high before vernalization, dropped during vernalization, and remained low post-vernalization, implying a role for this gene in flowering control. This expression pattern was not strongly affected by the genome edits (fig. S59). Plants with the editing events in the D subgenome flowered 2 to 3 days earlier than controls (Fig. 5E). Further refinement should help to fully elucidate the importance of the *TaAGL33* gene for vernalization in monocots. These results exemplify how the IWGSC RefSeq v1.0 could accelerate the development of diagnostic markers and the design of targets for genome editing for traits relevant to breeding.

Conclusions

IWGSC RefSeq v1.0 is a resource that has the potential for disruptive innovation in wheat improvement. By necessity, breeders work with the genome at the whole-chromosome level, as each new cross involves the modification of genome-wide gene networks that control the expression of complex traits such as yield. With the annotated and ordered reference genome sequence in place, researchers and breeders can now easily access sequence-level information to define changes in the genomes of lines in their programs. Although several hundred wheat QTLs have been published, only a small number of genes have been cloned and functionally characterized. IWGSC RefSeq v1.0 underpins immediate application by providing access to regulatory regions, and it will serve as the backbone to anchor all known QTLs to one common annotated reference. Combining this knowledge with the distribution of meiotic recombination frequency and genomic diversity will enable breeders to more efficiently tackle the challenges imposed by the need to balance the parallel selection processes for adaptation to biotic and abiotic stress, end-use quality, and yield improvement. Strategies can now be defined more precisely to bring desirable alleles into coupling phase, especially in less-recombinant regions of the wheat genome. Here the full potential of the newly available genome information may be realized through the implementation of DNA-marker platforms and targeted breeding technologies, including genome editing (*51*).

Methods summary

Whole-genome sequencing of cultivar Chinese Spring by short-read sequencing-by-synthesis provided the data for de novo genome assembly and scaffolding with the software package DenovoMAGIC2. The assembly was superscaffolded and anchored into 21 pseudomolecules with high-density genetic (POPSEQ) and physical (Hi-C and 21 chromosome-specific physical maps) mapping information and by integrating additional genomic resources. Validation of the assembly used independent genetic (de novo genotyping-by-sequencing maps) and physical mapping evidence (radiation hybrid maps, BioNano "optical maps" for group 7 homeologous chromosomes). The genome assembly was annotated for genes, repetitive DNA, and other genomic features, and in-depth comparative analyses were carried out to analyze the distribution of genes, recombination, position, and size of centromeres and the expansion and contraction of wheat gene families. An atlas of wheat gene transcription was built from an extensive panel of 850 independent transcriptome datasets and was then used to study gene coexpression networks. Furthermore, the assembly was used for the dissection of an important stem-solidness QTL and to design targets for genome editing of genes implicated in flowering-time control in wheat. Detailed methodological procedures are described in the supplementary materials.

REFERENCES AND NOTES

1. Food and Agriculture Organization of the United Nations, FAOSTAT statistics database, Food balance sheets (2017); www.fao.org/faostat/en/#data/FBS.
2. Food and Agriculture Organization of the United Nations, FAOSTAT statistics database, Crops (2017); www.fao.org/faostat/en/#data/QC.
3. G. N. Atlin, J. E. Cairns, B. Das, Rapid breeding and varietal replacement are critical to adaptation of cropping systems in the developing world to climate change. *Glob. Food Sec.* **12**, 31–37 (2017). doi: 10.1016/j.gfs.2017.01.008; pmid: 28580238
4. J. M. Hickey, T. Chiurugwi, I. Mackay, W. Powell; Implementing Genomic Selection in CGIAR Breeding Programs Workshop Participants, Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery. *Nat. Genet.* **49**, 1297–1303 (2017). doi: 10.1038/ng.3920; pmid: 28854179
5. K. Arumuganathan, E. D. Earle, Nuclear DNA content of some important plant species. *Plant Mol. Biol. Report.* **9**, 208–218 (1991). doi: 10.1007/BF02672069
6. International Wheat Genome Sequencing Consortium (IWGSC), A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science* **345**, 1251788 (2014). doi: 10.1126/science.1251788; pmid: 25035500
7. J. A. Chapman *et al*., A whole-genome shotgun approach for assembling and anchoring the hexaploid bread wheat genome. *Genome Biol.* **16**, 26 (2015). doi: 10.1186/s13059-015-0582-8; pmid: 25637298
8. B. J. Clavijo *et al*., An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. *Genome Res.* **27**, 885–896 (2017). doi: 10.1101/gr.217117.116; pmid: 28420692
9. A. V. Zimin *et al*., The first near-complete assembly of the hexaploid bread wheat genome, *Triticum aestivum*. *Gigascience* **6**, 1–7 (2017). doi: 10.1093/gigascience/gix097; pmid: 29069494
10. T. R. Endo, B. S. Gill, The deletion stocks of common wheat. *J. Hered.* **87**, 295–307 (1996). doi: 10.1093/oxfordjournals.jhered.a023003
11. M. E. Sorrells *et al*., Comparative DNA sequence analysis of wheat and rice genomes. *Genome Res.* **13**, 1818–1827 (2003). pmid: 12902377
12. K. Eversole, J. Rogers, B. Keller, R. Appels, C. Feuillet, in *Breeding, Quality Traits, Pests and Diseases*, vol. 1 of *Achieving Sustainable Cultivation of Wheat*, P. Langridge, Ed. (Burleigh-Dodds Science Publishing, 2017), chap. 2.
13. E. Paux *et al*., Insertion site-based polymorphism markers open new perspectives for genome saturation and marker-assisted selection in wheat. *Plant Biotechnol. J.* **8**, 196–210 (2010). doi: 10.1111/j.1467-7652.2009.00477.x; pmid: 20078842
14. See supplementary materials.
15. Arabidopsis Genome Initiative, Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796–815 (2000). doi: 10.1038/35048692; pmid: 11130711
16. International Rice Genome Sequencing Project, The map-based sequence of the rice genome. *Nature* **436**, 793–800 (2005). doi: 10.1038/nature03895; pmid: 16100779
17. T. Wicker *et al*., International Wheat Genome Sequencing Consortium, Impact of transposable elements on genome

structure and evolution in wheat. *Genome Biol.* doi: 10.1186/s13059-018-1479-0 (2018).
18. F. Choulet *et al*., Structural and functional partitioning of bread wheat chromosome 3B. *Science* **345**, 1249721 (2014). doi: 10.1126/science.1249721; pmid: 25035497
19. X. Guo *et al*., De novo centromere formation and centromeric sequence expansion in wheat and its wide hybrids. *PLOS Genet.* **12**, e1005997 (2016). doi: 10.1371/journal.pgen.1005997; pmid: 27110907
20. K. Wang, Y. Wu, W. Zhang, R. K. Dawe, J. Jiang, Maize centromeres expand and adopt a uniform size in the genetic background of oat. *Genome Res.* **24**, 107–116 (2014). doi: 10.1101/gr.160887.113; pmid: 24100079
21. Y. Jiao *et al*., Improved maize reference genome with single-molecule technologies. *Nature* **546**, 524–527 (2017). pmid: 28605751
22. H. Yan *et al*., Intergenic locations of rice centromeric chromatin. *PLOS Biol.* **6**, e286 (2008). doi: 10.1371/journal.pbio.0060286; pmid: 19067486
23. R. H. Ramírez-González *et al*., The transcriptional landscape of polyploid wheat. *Science* **361**, eaar6089 (2018). doi: 10.1126/science.aar6089
24. F. A. Simão, R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, E. M. Zdobnov, BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015). doi: 10.1093/bioinformatics/btv351; pmid: 26059717
25. M.-D. Rey *et al*., Exploiting the *ZIP4* homologue within the wheat *Ph1* locus has identified two lines exhibiting homoeologous crossover in wheat-wild relative hybrids. *Mol. Breed.* **37**, 95 (2017). doi: 10.1007/s11032-017-0700-2; pmid: 28781573
26. F. Cheng *et al*., Gene retention, fractionation and subgenome differences in polyploid plants. *Nat. Plants* **4**, 258–268 (2018). doi: 10.1038/s41477-018-0136-7; pmid: 29725103
27. T. Marcussen *et al*., Ancient hybridizations among the ancestral genomes of bread wheat. *Science* **345**, 1250092 (2014). pmid: 25035499
28. Y. Van de Peer, S. Maere, A. Meyer, The evolutionary significance of ancient genome duplications. *Nat. Rev. Genet.* **10**, 725–732 (2009). doi: 10.1038/nrg2600; pmid: 19652647
29. P. S. Soltis, D. E. Soltis, Ancient WGD events as drivers of key innovations in angiosperms. *Curr. Opin. Plant Biol.* **30**, 159–165 (2016). doi: 10.1016/j.pbi.2016.03.015; pmid: 27064530
30. M. Melé *et al*., The human transcriptome across tissues and individuals. *Science* **348**, 660–665 (2015). doi: 10.1126/science.aaa0355; pmid: 25954002
31. S. C. Stelpflug *et al*., An expanded maize gene expression atlas based on RNA sequencing and its use to explore root development. *Plant Genome* **9**, 0025 (2016). pmid: 27898762
32. F. He *et al*., Large-scale atlas of microarray data reveals the distinct expression landscape of different tissues in Arabidopsis. *Plant J.* **86**, 472–480 (2016). doi: 10.1111/tpj.13175; pmid: 27015116
33. X. Wang *et al*., Comparative genomic analysis of C4 photosynthetic pathway evolution in grasses. *Genome Biol.* **10**, R68 (2009). doi: 10.1186/gb-2009-10-6-r68; pmid: 19549309
34. L. Xia *et al*., Rice Expression Database (RED): An integrated RNA-Seq-derived gene expression database for rice. *J. Genet. Genomics* **44**, 235–241 (2017). doi: 10.1016/j.jgg.2017.05.003; pmid: 28529082
35. R. J. Schaefer, J.-M. Michno, C. L. Myers, Unraveling gene function in agricultural species using gene co-expression networks. *Biochim. Biophys. Acta* **1860**, 53–63 (2017). pmid: 27485388
36. J. Beales, A. Turner, S. Griffiths, J. W. Snape, D. A. Laurie, A *Pseudo-Response Regulator* is misexpressed in the photoperiod insensitive *Ppd-D1a* mutant of wheat (*Triticum aestivum* L.). *Theor. Appl. Genet.* **115**, 721–733 (2007). doi: 10.1007/s00122-007-0603-4; pmid: 17634915
37. L. Yan *et al*., The wheat and barley vernalization gene *VRN3* is an orthologue of *FT*. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 19581–19586 (2006). doi: 10.1073/pnas.0607142103; pmid: 17158798
38. K. V. Krasileva *et al*., Uncovering hidden variation in polyploid wheat. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E913–E921 (2017). doi: 10.1073/pnas.1619268114; pmid: 28096351
39. S. Wang *et al*., Cytological and transcriptomic analyses reveal important roles of *CLE19* in pollen exine formation. *Plant Physiol.* **175**, 1186–1202 (2017). doi: 10.1104/pp.17.00439; pmid: 28916592
40. M. Pfeifer *et al*., Genome interplay in the grain transcriptome of hexaploid bread wheat. *Science* **345**, 1250091 (2014). doi: 10.1126/science.1250091; pmid: 25035498
41. P. Borrill, N. Adamski, C. Uauy, Genomics as the key to unlocking the polyploid potential of wheat. *New Phytol.* **208**, 1008–1022 (2015). doi: 10.1111/nph.13533; pmid: 26108556
42. F. A. Kondrashov, Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proc. Biol. Sci.* **279**, 5048–5057 (2012). doi: 10.1098/rspb.2012.1108; pmid: 22977152
43. P. H. Schiffer, J. Gravemeyer, M. Rauscher, T. Wiehe, Ultra large gene families: A matter of adaptation or genomic parasites? *Life* **6**, 32 (2016). doi: 10.3390/life6030032; pmid: 27509525
44. T. Sykes *et al*., In silico identification of candidate genes for fertility restoration in cytoplasmic male sterile perennial ryegrass (*Lolium perenne* L.). *Genome Biol. Evol.* **9**, 351–362 (2017). pmid: 26951780
45. J. Melonek, J. D. Stone, I. Small, Evolutionary plasticity of restorer-of-fertility-like proteins in rice. *Sci. Rep.* **6**, 35152 (2016). doi: 10.1038/srep35152; pmid: 27775031
46. T. Würschum, C. F. H. Longin, V. Hahn, M. R. Tucker, W. L. Leiser, Copy number variations of *CBF* genes at the *Fr-A2* locus are essential components of winter hardiness in wheat. *Plant J.* **89**, 764–773 (2017). doi: 10.1111/tpj.13424; pmid: 27859852
47. A. Juhász *et al*., Genome mapping of seed-borne allergens and immune-responsive proteins in wheat. *Sci. Adv.* **4**, eaar8602 (2018). doi: 10.1126/sciadv.aar8602
48. M. Feldman, A. A. Levy, in *Alien Introgression in Wheat: Cytogenetics, Molecular Biology, and Genomics*, M. Molnár-Láng, C. Ceoloni, J. Doležel, Eds. (Springer, 2015), pp. 21–76.
49. K. T. Nilsen *et al*., High density mapping and haplotype analysis of the major stem-solidness locus *SSt1* in durum and common wheat. *PLOS ONE* **12**, e0175285 (2017). doi: 10.1371/journal.pone.0175285; pmid: 28399136
50. N. Sharma *et al*., A flowering locus C homolog is a vernalization-regulated repressor in *Brachypodium* and is cold regulated in wheat. *Plant Physiol.* **173**, 1301–1315 (2017). doi: 10.1104/pp.16.01161; pmid: 28034954
51. H. Puchta, Applying CRISPR/Cas for genome engineering in plants: The best is yet to come. *Curr. Opin. Plant Biol.* **36**, 1–8 (2017). doi: 10.1016/j.pbi.2016.11.011; pmid: 27914284
52. International Wheat Genome Sequencing Consortium, Gene family expansion and contraction in the genome of bread wheat cv. Chinese Spring. e!DAL (2018). doi: 10.5447/IPK/2018/5
53. H.-Q. Ling *et al*., Draft genome of the wheat A-genome progenitor *Triticum urartu*. *Nature* **496**, 87–90 (2013). doi: 10.1038/nature11997; pmid: 23535596
54. J. Jia *et al*., *Aegilops tauschii* draft genome sequence reveals a gene repertoire for wheat adaptation. *Nature* **496**, 91–95 (2013). doi: 10.1038/nature12028; pmid: 23535592
55. Y. Ishida, M. Tsunashima, Y. Hiei, T. Komari, in *Agrobacterium Protocols: Volume 1*, K. Wang, Ed. (Springer, 2015), pp. 189–198.

ACKNOWLEDGMENTS

The IWGSC would like to thank the following individuals: M. Burrell and C. Bridson (Norwich Biosciences Institute) for computational support of RNA-seq data; I. Christie (Graminor AS) and H. Rudi (Norwegian University of Life Sciences) for assistance with chromosome 7B; R. P. Davey (Earlham Institute) for assistance with RNA-seq data; J. Deek (Tel Aviv University) for growing the source plants and DNA extraction used for whole-genome sequencing; Z. Dubská, E. Jahnová, M. Seifertová, R. Šperková, R. Tušková, and J. Weiserová (Institute of Experimental Botany, Olomouc) for assistance with flow cytometric chromosome sorting, BAC library construction, and estimation of genome size; S. Durand, V. Jamilloux, M. Lainé, and C. Michotey (URGI, INRA) for assistance with and access to the IWGSC sequence repository; A. Fiebig of the Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) for submitting the Hi-C data; T. Florio for the design of the wheat schematic for the expression atlas and *SSt1* figure (www.flozbox.com/Science_Illustrated); C. Karunakaran and T. Bond of the Canadian Light Source for performing CT imaging; J. Kawai, N. Kondo, H. Sano, N. Suzuki, M. Tagami, and H. Tarui of RIKEN for assistance with deep sequencing of chromosome 6B; H. Fujisawa, Y. Katayose, K. Kurita, S. Mori, Y. Mukai, and H. Sasaki of the Institute of Crop Science, NARO for assistance with deep sequencing of chromosome 6B; T. Matsumoto of Tokyo University of Agriculture for assistance with deep sequencing of chromosome 6B; P. Lenoble and C. Orvain of Genoscope for assistance in the sequencing of chromosome 1B; A. J. Lukaszewski of the University of California, Riverside, and B. Friebe and J. Raupp of Kansas State University for providing seeds of wheat telosomic lines for chromosome sorting; C. Maulis (https://polytypo.design; https://propepper.net) for design and graphics of the prolamin superfamily chromosome map; M. Seifertová and H. Tvardíková of the Institute of Experimental Botany for assistance with BAC DNA extraction and sequencing for chromosomes 3DS, 4A, and 7DS; and I. Willick and K. Tanino of the University of Saskatchewan for their assistance in sample preparation and the use of lab facilities.
Funding: The authors would like to thank the following for their financial support of research that enabled the completion of the IWGSC RefSeq v1.0 Project: Agence Nationale pour la Recherche (ANR), ANR-11-BSV5-0015–Ploid-Ploid Wheat–Unravelling bases of polyploidy success in wheat and ANR-16-TERC-0026-01–3DWHEAT; Agriculture and Agri-Food Canada National Wheat Improvement Program and the AgriFlex Program; Alberta Wheat Development Commission through the Canadian Applied Triticum Genomics (CTAG2); Australian Government, Department of Industry, Innovation, Climate Change, Science, Research, and Tertiary Education, Australia China Science and Research Fund Group Mission (Funding Agreement ACSRF00542); Australian Research Council Centre of Excellence in Plant Energy Biology (CE140100008); Australian Research Council Laureate Fellowship (FL140100179); Bayer CropScience; Biotechnology and Biological Sciences Research Council (BBSRC) 20:20 Wheat (project number BB/J00426X/1), Institute Strategic Programme grant (BB/J004669/1), Designing Future Wheat (DFW) Institute Strategic Programme (BB/P016855/1), the Wheat Genomics for Sustainable Agriculture (BB/J003557/1), and the Anniversary Future Leader Fellowship (BB/M014045/1); Canada First Research Excellence Fund through the Designing Crops for Global Food Security initiative at the University of Saskatchewan; Council for Agricultural Research and Economics, Italy, through CREA-Interomics; Department of Biotechnology, Ministry of Science and Technology, Government of India File No. F grant no. BT/IWGSC/03/TF/2008; DFG (SFB924) for support of KFXM; European Commission through the *TriticeaeGenome* (FP7-212019); France Génomique (ANR-10-INBS-09) Genome Canada through the CTAG2 project; Genome Prairie through the CTAG2 project; German Academic Exchange Service (DAAD) PPP Australien 1j16; German Federal Ministry of Food and Agriculture grant 2819103915 WHEATSEQ; German Ministry of Education and Research grant 031A536 de.NBI; Global Institute for Food Security Genomics and Bioinformatics fund; Gordon and Betty Moore Foundation grant GBMF4725 to Two Blades Foundation; Grain Research Development Corporation (GRDC) Australia; Graminor AS NFR project 199387, Expanding the technology base for Norwegian wheat breeding; Sequencing wheat chromosome 7B; illumina; INRA, French National Institute for Agricultural Research; International Wheat Genome Sequencing Consortium and its sponsors; Israel Science Foundation grants 999/12, 1137/17, and 1824/12; Junta de Andalucía, Spain, project P12-AGR-0482; MINECO (Spanish Ministry of Economy, Industry, and Competitiveness) project BIO2011-15237-E; Ministry of Agriculture, Forestry, and Fisheries of Japan through Genomics for Agricultural Innovation, KGS-1003 and through Genomics-based Technology for Agricultural Improvement, NGB-1003; Ministry of Education and Science of Russian Federation project RFMEFI60414X0106 and project RFMEFI60414 X0107; Ministry of Education, Youth, and Sport of the Czech Republic award no. LO1204 (National Program of Sustainability I); Nisshin Flour Milling, Inc.; National Research Council of Canada Wheat Flagship program; Norwegian University of Life Sciences (NMBU) NFR project 199387, Expanding the technology base for Norwegian wheat breeding, Sequencing wheat chromosome 7B; National Science Foundation, United States, award (FAIN) 1339389, GPF-PG: Genome Structure and Diversity of Wheat and Its Wild Relatives, award DBI-0701916, and award IIP-1338897; Russian Science Foundation project 14-14-00161; Saskatchewan Ministry of Agriculture through the CTAG2 project; Saskatchewan Wheat Development Commission through the CTAG2 project; The Czech Science Foundation award no. 521/06/1723 (Construction of BAC library and physical mapping of the wheat chromosome 3D), award no. 521-08-1629 (Construction of BAC DNA libraries specific for chromosome 4AL and positional cloning of gene for adult plant resistance to powdery mildew in wheat), award no. P501/10/1740 (Physical map of the wheat chromosome 4AL and positional cloning of a gene for yield), award no. P501/12/2554 (Physical map of wheat chromosome arm 7DS and its use to clone a Russian wheat aphid-resistance gene), award no. P501/12/G090 (Evolution and function of complex plant genomes), award no. 14-07164S (Cloning and molecular characterization of wheat *QPm-tut-4A* gene conferring seedling and adult plant race-nonspecific powdery-mildew resistance), and award no. 13-08786S (Chromosome arm 3DS of bread wheat: Its

sequence and function in allopolyploid genome); The Research Council of Norway (NFR) project 199387, Expanding the technology base for Norwegian wheat breeding; Sequencing wheat chromosome 7B; U.S. Department of Agriculture NIFA 2008-35300-04588, the University of Zurich; Western Grains Research Foundation through the CTAG2 project; Western Grains Research Foundation National Wheat Improvement Program; and the Winifred-Asbjornson Plant Science Endowment Fund. The research leading to these results has also received funding from the French Government managed by the ANR under the Investment for the Future program (BreedWheat project ANR-10-BTBR-03), from FranceAgriMer (2011-0971 and 2013-0544), French Funds to support Plant Breeding (FSOV), and INRA. Axiom genotyping was conducted on the genotyping platform GENTYANE at INRA Clermont-Ferrand (gentyane.clermont.inra.fr). This research was supported in part by the NBI Computing infrastructure for Science (CiS) group through the HPC cluster.
Author contributions: See below, where authors are arranged by working group and contributions; leaders, co-leaders, and major contributors are listed alphabetically first and then other contributors follow alphabetically. **Competing interests:** The authors declare no competing interests. Bayer CropScience holds a patent application (WO2015000914A1) that covers the modulation of flowering time in monocots using the *FLC* gene. **Data and materials availability:** The IWGSC RefSeq v1.0 assembly and annotation data, physical maps for all chromosomes and chromosome arms, and all data related to this study are available in the IWGSC Data Repository hosted at URGI: https://wheat-urgi.versailles.inra.fr/Seq-Repository, Assembly and annotation data are also available at ENSEMBL-Plants: https://plants.ensembl.org/Triticum_aestivum/Info/Index. The BAC libraries for all chromosomes and chromosomes arms are available at the CNRGV-INRA: https://cnrgv.toulouse.inra.fr/en/Library/Wheat. Details on gene-family expansion and contraction in the genome of bread wheat cultivar Chinese Spring are provided in database S6 at http://dx.doi.org/10.5447/IPK/2018/5 (52). The raw sequencing data used for de novo whole-genome assembly are available from the Sequence Read Archive under accession number SRP114784. RNA-seq data are available at SRA under accession numbers PRJEB25639, PRJEB23056, PRJNA436817, PRJEB25640, SRP133837, and PRJEB25593. Hi-C sequence data are available under accession number PRJEB25248. ChIP-seq data are available under SRA study PRJNA420988 (SRP1262229). CS bisulfite sequencing data are available under project number SRP133674 (SRR6792673 to SRR6792689). Organellar DNA sequences were deposited at NCBI GenBank (MH051715 and MH051716). Further details on data accessibility are outlined in the supplementary materials and methods.

The International Wheat Genome Sequencing Consortium (IWGSC)

IWGSC RefSeq principal investigators: Rudi Appels[1,36*]‡, Kellye Eversole[2,3*]‡, Catherine Feuillet[17], Beat Keller[41], Jane Rogers[6]‡, Nils Stein[4,5*]‡.

IWGSC whole-genome assembly principal investigators: Curtis J. Pozniak[11]‡, Nils Stein[4,5*]‡, Frédéric Choulet[7], Assaf Distelfeld[25], Kellye Eversole[2,3*], Jesse Poland[28], Jane Rogers[6], Gil Ronen[12], Andrew G. Sharpe[43].

Whole-genome sequencing and assembly: Curtis Pozniak[11]‡, Gil Ronen[12]‡, Nils Stein[4,5*]‡, Omer Barad[12]†, Kobi Baruch[12]†, Frédéric Choulet[7]†, Gabriel Keeble-Gagnère[1]†, Martin Mascher[4,67]†, Andrew G. Sharpe[43]†, Gil Ben-Zvi[12]†, Ambre-Aurore Josselin[7].

Hi-C data-based scaffolding: Nils Stein[4,5*]‡, Martin Mascher[4,67]‡, Axel Himmelbach[4].

Whole-genome assembly quality control and analyses: Frédéric Choulet[7]‡, Gabriel Keeble-Gagnère[1]‡, Martin Mascher[4,67]‡, Jane Rogers[6]‡, François Balfourier[7], Juan Gutierrez-Gonzalez[30], Matthew Hayden[1], Ambre-Aurore Josselin[7], ChuShin Koh[43], Gary Muehlbauer[30], Raj K. Pasam[1], Etienne Paux[7], Curtis J. Pozniak[11], Philippe Rigault[39], Andrew G. Sharpe[43], Josquin Tibbits[1], Vijay Tiwari[54].

Pseudomolecule assembly: Frédéric Choulet[7]‡, Gabriel Keeble-Gagnère[1]‡, Martin Mascher[4,67]‡, Ambre-Aurore Josselin[7], Jane Rogers[6].

RefSeq genome structure and gene analyses: Manuel Spannagl[9]‡, Frédéric Choulet[7]‡, Daniel Lang[9]†, Heidrun Gundlach[9], Georg Haberer[9], Gabriel Keeble-Gagnère[1], Klaus F. X. Mayer[9,44], Danara Ormanbekova[9,48], Etienne Paux[7], Verena Prade[9], Hana Šimková[8], Thomas Wicker[41].

Automated annotation: Frédéric Choulet[7]‡, Manuel Spannagl[9]‡, David Swarbreck[50]‡, Hélène Rimbert[7]†, Marius Felder[9], Nicolas Guilhot[7], Heidrun Gundlach[9], Georg Haberer[9], Gemy Kaithakottil[50], Jens Keilwagen[40], Daniel Lang[9], Philippe Leroy[7], Thomas Lux[9], Klaus F. X. Mayer[9,44], Sven Twardziok[9], Luca Venturini[50].

Manual gene curation: Rudi Appels[1,36*]‡, Hélène Rimbert[7]†, Frédéric Choulet[7], Angéla Juhász[36,37], Gabriel Keeble-Gagnère[1].

Subgenome comparative analyses: Frédéric Choulet[7]‡, Manuel Spannagl[9]‡, Daniel Lang[9]†, Michael Abrouk[8,19], Georg Haberer[9], Gabriel Keeble-Gagnère[1], Klaus F. X. Mayer[9,44], Thomas Wicker[41].

Transposable elements: Frédéric Choulet[7]‡, Thomas Wicker[41]†, Heidrun Gundlach[9]†, Daniel Lang[9], Manuel Spannagl[9].

Phylogenomic analyses: Daniel Lang[9]‡, Manuel Spannagl[9]‡, Rudi Appels[1,36*], Iris Fischer[9].

Transcriptome analyses and RNA-seq data: Cristobal Uauy[10]‡, Philippa Borrill[10]†, Ricardo H. Ramirez-Gonzalez[10]†, Rudi Appels[1,36*], Dominique Arnaud[63], Smahane Chalabi[63], Boulos Chalhoub[62,63], Frédéric Choulet[7], Aron Cory[11], Raju Datla[22], Mark W. Davey[18], Matthew Hayden[1], John Jacobs[18], Daniel Lang[9], Stephen J. Robinson[52], Manuel Spannagl[9], Burkhard Steuernagel[10], Josquin Tibbits[1], Vijay Tiwari[54], Fred van Ex[18], Brande B. H. Wulff[10].

Whole-genome methylome: Curtis J. Pozniak[11]‡, Stephen J. Robinson[52]‡, Andrew G. Sharpe[43]‡, Aron Cory[11].

Histone mark analyses: Moussa Benhamed[15]‡, Etienne Paux[7]‡, Abdelhafid Bendahmane[15], Lorenzo Concia[15], David Latrasse[15].

BAC chromosome MTP IWGSC–Bayer Whole-Genome Profiling (WGP) tags: Jane Rogers[6]‡, John Jacobs[18]‡, Michael Alaux[13], Rudi Appels[1,36*], Jan Bartoš[8], Arnaud Bellec[20], Hélène Berges[20], Jaroslav Doležel[8], Catherine Feuillet[17], Zeev Frenkel[26], Bikram Gill[28], Abraham Korol[26], Thomas Letellier[13], Odd-Arne Olsen[56], Hana Šimková[8], Kuldeep Singh[65], Miroslav Valárik[8], Edwin van der Vossen[64], Sonia Vautrin[20], Song Weining[66].

Chromosome LTC mapping and physical mapping quality control: Abraham Korol[26]‡, Zeev Frenkel[26]†, Tzion Fahima[26]†, Vladimir Glikson[29], Dina Raats[50], Jane Rogers[6].

RH mapping: Vijay Tiwari[54]‡, Bikram Gill[28], Etienne Paux[7], Jesse Poland[28].

Optical mapping: Jaroslav Doležel[8]‡, Jarmila Čihalíková[8], Hana Šimková[8], Helena Toegelová[8], Jan Vrána[8].

Recombination analyses: Pierre Sourdille‡[7], Benoit Darrier[7].

Gene family analyses: Rudi Appels[1,36*]‡, Manuel Spannagl[9]‡, Daniel Lang[9]†, Iris Fischer[9], Danara Ormanbekova[9,48], Verena Prade[9].

CBF gene family: Delfina Barabaschi[16]‡, Luigi Cattivelli[16].

Dehydrin gene family: Pilar Hernandez[33]‡, Sergio Galvez[27]‡, Hikmet Budak[14].

NLR gene family: Burkhard Steuernagel[10]‡, Jonathan D. G. Jones[35], Kamil Witek[35], Brande B. H. Wulff[10], Guotai Yu[10].

PPR gene family: Ian Small[45]‡, Joanna Melonek[45]†, Ruonan Zhou[4].

Prolamin gene family: Angéla Juhász[36,37]‡, Tatiana Belova[56]†, Rudi Appels[1,36*], Odd-Arne Olsen[56].

WAK gene family: Kostya Kanyuka[38]‡, Robert King[42]†.

Stem solidness (*SSt1*) QTL team: Kirby Nilsen[11]‡, Sean Walkowiak[11]‡, Curtis J. Pozniak[11]‡, Richard Cuthbert[21], Raju Datla[22], Ron Knox[21], Krysta Wiebe[11], Daoquan Xiang[22].

Flowering locus C (*FLC*) gene team: Antje Rohde[72]‡, Timothy Golds[18]‡.

Genome size analysis: Jaroslav Doležel[8]‡, Jana Čížková[8], Josquin Tibbits[1].

MicroRNA and tRNA annotation: Hikmet Budak[14]‡, Bala Ani Akpinar[14], Sezgi Biyiklioglu[14].

Genetic maps and mapping: Gary Muehlbauer[30]‡, Jesse Poland[28]‡, Liangliang Gao[28], Juan Gutierrez-Gonzalez[30], Amidou N'Daiye[11].

BAC libraries and chromosome sorting: Jaroslav Doležel[8]‡, Hana Šimková[8]†, Jarmila Čihalíková[8], Marie Kubaláková[8], Jan Šafář[8], Jan Vrána[8].

BAC pooling, BAC library repository, and access: Hélène Berges[20]‡, Arnaud Bellec[20], Sonia Vautrin[20].

IWGSC sequence and data repository and access: Michael Alaux[13]‡, Françoise Alfama[13], Anne-Françoise Adam-Blondon[13], Raphael Flores[13], Claire Guerche[13], Thomas Letellier[13], Mikaël Loaec[13], Hadi Quesneville[13].

Physical maps and BAC-based sequences:

1A BAC sequencing and assembly: Curtis J. Pozniak[11]‡, Andrew G. Sharpe[22,43]‡, Sean Walkowiak‡[11], Hikmet Budak[14], Janet Condie[22], Jennifer Ens[11], ChuShin Koh[43], Ron Maclachlan[11], Yifang Tan[22], Thomas Wicker[41].

1B BAC sequencing and assembly: Frédéric Choulet[7]‡, Etienne Paux[7]‡, Adriana Alberti[61], Jean-Marc Aury[61], François Balfourier[7], Valérie Barbe[61], Arnaud Couloux[61], Corinne Cruaud[61], Karine Labadie[61], Sophie Mangenot[61], Patrick Wincker[61,68,69].

1D, 4D, and 6D physical mapping: Bikram Gill[28]‡, Gaganpreet Kaur[28], Mingcheng Luo[34], Sunish Sehgal[53].

2AL physical mapping: Kuldeep Singh[65]‡, Parveen Chhuneja[65], Om Prakash Gupta[65], Suruchi Jindal[65], Parampreet Kaur[65], Palvi Malik[65], Priti Sharma[65], Bharat Yadav[65].

2AS physical mapping: Nagendra K. Singh[70]‡, Jitendra P. Khurana[71]‡, Chanderkant Chaudhary[71], Paramjit Khurana[71], Vinod Kumar[70], Ajay Mahato[70], Saloni Mathur[71], Amitha Sevanthi[70], Naveen Sharma[71], Ram Sewak Tomar[70].

2B, 2D, 4B, 5BL, and 5DL IWGSC–Bayer Whole-Genome Profiling (WGP) physical maps: Jane Rogers[6]‡, John Jacobs[18]‡, Michael Alaux[13], Arnaud Bellec[20], Hélène Berges[20], Jaroslav Doležel[8], Catherine Feuillet[17], Zeev Frenkel[26], Bikram Gill[28], Abraham Korol[26], Edwin van der Vossen[64], Sonia Vautrin[20].

3AL physical mapping: Bikram Gill[28]‡, Gaganpreet Kaur[28], Mingcheng Luo[34], Sunish Sehgal[53].

3DS physical mapping and BAC sequencing and assembly: Jan Bartoš[8]‡, Kateřina Holušová[8], Ondřej Plíhal[49].

3DL BAC sequencing and assembly: Matthew D. Clark[50,73], Darren Heavens[50], George Kettleborough[50], Jon Wright[50].

4A physical mapping, BAC sequencing, assembly, and annotation: Miroslav Valárik[8]‡, Michael Abrouk[8,19], Barbora Balcárková[8], Kateřina Holušová[8], Yuqin Hu[34], Mingcheng Luo[34].

5BS BAC sequencing and assembly: Elena Salina[47]‡, Nikolai Ravin[23,51]‡, Konstantin Skryabin[23,51]‡, Alexey Beletsky[23], Vitaly Kadnikov[23], Andrey Mardanov[23], Michail Nesterov[47], Andrey Rakitin[23], Ekaterina Sergeeva[47].

6B BAC sequencing and assembly: Hirokazu Handa[31]‡, Hiroyuki Kanamori[31], Satoshi Katagiri[31], Fuminori Kobayashi[31], Shuhei Nasuda[46], Tsuyoshi Tanaka[31], Jianzhong Wu[31].

7A physical mapping and BAC sequencing: Rudi Appels[1,36*]‡, Matthew Hayden[1], Gabriel Keeble-Gagnère[1], Philippe Rigault[39], Josquin Tibbits[1].

7B physical mapping, BAC sequencing, and assembly: Odd-Arne Olsen[56]‡, Tatiana Belova[56]†, Federica Cattonaro[58], Min Jiumeng[60], Karl Kugler[9], Klaus F. X. Mayer[9,44], Matthias Pfeifer[9], Simen Sandve[57], Xu Xun[59], Bujie Zhan[56]†.

7DS BAC sequencing and assembly: Hana Šimková[8]‡, Michael Abrouk[8,19], Jacqueline Batley[24], Philipp E. Bayer[24], David Edwards[24], Satomi Hayashi[32], Helena Toegelová[8], Zuzana Tulpová[8], Paul Visendi[55].

7DL physical mapping and BAC sequencing: Song Weining[66]‡, Licao Cui[66], Xianghong Du[66], Kewei Feng[66], Xiaojun Nie[66], Wei Tong[66], Le Wang[66].

Figures: Philippa Borrill[10], Heidrun Gundlach[9], Sergio Galvez[27], Gemy Kaithakottil[50], Daniel Lang[9], Thomas Lux[9], Martin Mascher[4,67], Danara Ormanbekova[9,48], Verena Prade[9], Ricardo H. Ramirez-Gonzalez[10], Manuel Spannagl[9], Nils Stein[4,5*], Cristobal Uauy[10], Luca Venturini[50].

Manuscript writing team: Nils Stein[4,5*]‡, Rudi Appels[1,36*]‡, Kellye Eversole[2,3*], Jane Rogers[6]†, Philippa Borrill[10], Luigi Cattivelli[16], Frédéric Choulet[7], Pilar Hernandez[33], Kostya Kanyuka[38], Daniel Lang[9], Martin Mascher[4,67], Kirby Nilsen[11], Etienne Paux[7], Curtis J. Pozniak[11], Ricardo H. Ramirez-Gonzalez[10], Hana Šimková[8], Ian Small[45], Manuel Spannagl[9], David Swarbreck[50], Cristobal Uauy[10].

[1]AgriBio, Centre for AgriBioscience, Department of Economic Development, Jobs, Transport, and Resources, 5 Ring Road, La Trobe University, Bundoora, VIC 3083, Australia. [2]International Wheat Genome Sequencing Consortium (IWGSC), 5207 Wyoming Road, Bethesda, MD 20816, USA. [3]Eversole Associates, 5207 Wyoming Road, Bethesda, MD 20816, USA. [4]Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Genebank, Corrensstr. 3, 06466 Stadt Seeland, Germany. [5]The University of Western Australia (UWA), School of Agriculture and Environment, 35 Stirling Highway, Crawley, WA 6009, Australia. [6]International Wheat Genome Sequencing Consortium (IWGSC), 18 High Street, Little Eversden, Cambridge CB23 1HE, UK. [7]GDEC (Genetics, Diversity and Ecophysiology of Cereals), INRA, Université Clermont Auvergne (UCA), 5 chemin de Beaulieu, 63039 Clermont-Ferrand, France. [8]Institute of Experimental Botany, Centre of the Region Haná for Biotechnological and Agricultural Research, Šlechtitelů 31, CZ-78371 Olomouc, Czech Republic. [9]Helmholtz Center Munich, Plant Genome and Systems Biology (PGSB), Ingolstaedter Landstr. 1, 85764 Neuherberg, Germany. [10]John Innes Centre, Crop Genetics, Norwich Research Park, Norwich NR4 7UH, UK. [11]University of Saskatchewan, Crop Development Centre, Agriculture Building, 51 Campus Drive, Saskatoon, SK S7N 5A8, Canada. [12]NRGene Ltd., 5 Golda Meir Street, Ness Ziona 7403648, Israel. [13]URGI, INRA, Université Paris-Saclay, 78026 Versailles, France. [14]Plant Sciences and Plant Pathology, Cereal Genomics Lab, Montana State University, 412 Leon Johnson Hall, Bozeman, MT 59717, USA. [15]Biology Department, Institute of Plant Sciences–Paris-Saclay, Bâtiment 630, rue de Noetzlin, Plateau du Moulon, CS80004, 91192 Gif-sur-Yvette Cedex, France. [16]Council for Agricultural Research and Economics (CREA), Research Centre for Genomics and Bioinformatics, via S. Protaso, 302, I -29017

Fiorenzuola d'Arda, Italy. [17]Bayer CropScience, Crop Science Division, Research and Development, Innovation Centre, 3500 Paramount Parkway, Morrisville, NC 27560, USA. [18]Bayer CropScience, Trait Research, Innovation Center, Technologiepark 38, 9052 Gent, Belgium. [19]Biological and Environmental Science and Engineering Division, King Abdullah University of Science and Technology, Thuwal 23955-6900, Kingdom of Saudi Arabia. [20]INRA, CNRGV, chemin de Borde Rouge, CS 52627, 31326 Castanet-Tolosan Cedex, France. [21]Agriculture and Agri-Food Canada, Swift Current Research and Development Centre, Box 1030, Swift Current, SK S9H 3X2, Canada. [22]National Research Council Canada, Aquatic and Crop Resource Development, 110 Gymnasium Place, Saskatoon, SK S7N 0W9, Canada. [23]Research Center of Biotechnology of the Russian Academy of Sciences, Institute of Bioengineering, Leninsky Avenue 33, Building 2, Moscow 119071, Russia. [24]School of Biological Sciences and Institute of Agriculture, University of Western Australia, Perth, WA 6009, Australia. [25]School of Plant Sciences and Food Security, Tel Aviv University, Ramat Aviv 69978, Israel. [26]University of Haifa, Institute of Evolution and the Department of Evolutionary and Environmental Biology, 199 Abba-Hushi Avenue, Mount Carmel, Haifa 3498838, Israel. [27]Universidad de Málaga, Lenguajes y Ciencias de la Computación, Campus de Teatinos, 29071 Málaga, Spain. [28]Plant Pathology, Throckmorton Hall, Kansas State University, Manhattan, KS 66506, USA. [29]MultiQTL Ltd., University of Haifa, Haifa 3498838, Israel. [30]Department of Agronomy and Plant Genetics, University of Minnesota, 411 Borlaug Hall, St. Paul, MN 55108, USA. [31]Institute of Crop Science, NARO, 2-1-2 Kannondai, Tsukuba, Ibaraki 305-8518, Japan. [32]Queensland University of Technology, Earth, Environmental and Biological Sciences, Brisbane, QLD 4001, Australia. [33]Instituto de Agricultura Sostenible (IAS-CSIC), Consejo Superior de Investigaciones Científicas, Alameda del Obispo s/n, 14004 Córdoba, Spain. [34]Department of Plant Sciences, University of California, Davis, One Shield Avenue, Davis, CA 95617, USA. [35]The Sainsbury Laboratory, Norwich Research Park, Norwich NR4 7UH, UK. [36]Murdoch University, Australia-China Centre for Wheat Improvement, School of Veterinary and Life Sciences, 90 South Street, Murdoch, WA 6150, Australia. [37]Agricultural Institute, MTA Centre for Agricultural Research, Applied Genomics Department, 2 Brunszvik Street, Martonvásár H 2462, Hungary. [38]Rothamsted Research, Biointeractions and Crop Protection, West Common, Harpenden AL5 2JQ, UK. [39]GYDLE, Suite 220, 1135 Grande Allée, Ouest, Québec, QC G1S 1E7, Canada. [40]Julius Kühn-Institut, Institute for Biosafety in Plant Biotechnology, Erwin-Baur-Str. 27, 06484 Quedlinburg, Germany. [41]Department of Plant and Microbial Biology, University of Zurich, Zollikerstrasse 107, 8008 Zurich, Switzerland. [42]Rothamsted Research, Computational and Analytical Sciences, West Common, Harpenden AL5 2JQ, UK. [43]University of Saskatchewan, Global Institute for Food Security, 110 Gymnasium Place, Saskatoon, SK S7N 4J8, Canada. [44]School of Life Sciences Weihenstephan, Technical University of Munich, 85354 Freising, Germany. [45]School of Molecular Sciences, ARC Centre of Excellence in Plant Energy Biology, The University of Western Australia, 35 Stirling Highway, Crawley, WA 6009, Australia. [46]Graduate School of Agriculture, Kyoto University, Kitashirakawaoiwake-cho, Sakyo-ku, Kyoto 606-8502, Japan. [47]The Federal Research Center Institute of Cytology and Genetics, SB RAS, pr. Lavrentyeva 10, Novosibirsk 630090, Russia. [48]Department of Agricultural Sciences, University of Bologna, Viale Fanin, 44 40127 Bologna, Italy. [49]Department of Molecular Biology, Centre of the Region Haná for Biotechnological and Agricultural Research, Palacký University, Šlechtitelů 27, CZ-78371 Olomouc, Czech Republic. [50]Earlham Institute, Core Bioinformatics, Norwich NR4 7UZ, UK. [51]Faculty of Biology, Moscow State University, Leninskie Gory, 1, Moscow 119991, Russia. [52]Agriculture and Agri-Food Canada, Saskatoon Research and Development Centre, 107 Science Place, Saskatoon, SK S7N 0X2, Canada. [53]Agronomy Horticulture and Plant Science, South Dakota State University, 2108 Jackrabbit Drive, Brookings, SD 57006, USA. [54]Plant Science and Landscape Architecture, University of Maryland, 4291 Fieldhouse Road, 2102 Plant Sciences Building, College Park, MD 20742, USA. [55]University of Greenwich, Natural Resources Institute, Central Avenue, Chatham, Kent ME4 4TB, UK. [56]Faculty of Bioscience, Department of Plant Science, Norwegian University of Life Sciences, Arboretveien 6, 1433 Ås, Norway. [57]Faculty of Bioscience, Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, Arboretveien 6, 1433 Ås, Norway. [58]Instituto di Genomica Applicata, Via J. Linussio 51, Udine 33100, Italy. [59]BGI-Shenzhen, BGI Genomics, Yantian District, Shenzhen 518083, Guangdong, China. [60]BGI-Shenzhen, BGI Genomics, Building No. 7, BGI Park, No. 21 Hongan 3rd Street, Yantian District, Shenzhen 518083, China. [61]CEA–Institut de Biologie François-Jacob, Genoscope, 2 rue Gaston Crémieux, 91057 Evry Cedex, France. [62]Monsanto SAS, 28000 Boissay, France. [63]Institut National de la Recherche Agronomique (INRA), 2 rue Gaston Crémieux, 91057 Evry Cedex, France. [64]Keygene, N.V., Agro Business Park 90, 6708 PW Wageningen, Netherlands. [65]Punjab Agricultural University, Ludhiana, School of Agricultural Biotechnology, ICAR–National Bureau of Plant Genetic Resources, Dev Prakash Shastri Marg, New Delhi 110012, India. [66]State Key Laboratory of Crop Stress Biology in Arid Areas, College of Agronomy, Northwest A&F University, Yangling, Shaanxi 712101, China. [67]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Deutscher Platz 5e, 04103 Leipzig, Germany. [68]CNRS, UMR 8030, CP5706, 91057 Evry, France. [69]Université d'Evry, UMR 8030, CP5706, 91057 Evry, France. [70]ICAR–National Research Centre on Plant Biotechnology, LBS Building, Pusa Campus, New Delhi 110012, India. [71]University of Delhi South Campus, Interdisciplinary Center for Plant Genomics and Department of Plant Molecular Biology, Benito Juarez Road, New Delhi 110021, India. [72]Bayer CropScience, Breeding and Trait Development, Technologiepark 38, 9052 Gent, Belgium. [73]Department of Lifesciences, Natural History Museum, Cromwell Road, London SW7 5BD, UK.

Authorship of this paper should be cited as "International Wheat Genome Sequencing Consortium" (IWGSC, 2018).
*Corresponding author. Email: rudi.appels@unimelb.edu.au (R.A.); eversole@eversoleassociates.com (K.E.); stein@ipk-gatersleben.de (N.S.)
†Major contributors
‡Working group leader(s) or co-leaders

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6403/eaar7191/suppl/DC1
Materials and Methods
Figs. S1 to S59
Tables S1 to S43
References (*56*–*186*)
Databases S1 to S5

13 December 2017; accepted 11 July 2018
10.1126/science.aar7191

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005.

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005.

APPENDIX IV

Integrated physical map of bread wheat chromosome arm 7DS to support evolutionary studies and genome sequencing

Tulpová Z, Luo MC, Toegelová H, Visendi P, Hayashi S, Vojta P, Hastie A, Kilian A, Tiwari VK, Bartoš J, Batley J, Edwards D, Doležel J, Šimková H

In: Abstracts of the "Olomouc Biotech 2017. Plant Biotechnology: Green for Good IV". Olomouc, Czech Republic 2017

Integrated physical map of bread wheat chromosome arm 7DS to support evolutionary studies and genome sequencing

Zuzana Tulpová[1], Ming-Cheng Luo[2], Helena Toegelová[1], Paul Visendi[3], Satomi Hayashi[4], Petr Vojta[6], Alex Hastie[7], Andrzej Kilian[8], Vijay K. Tiwari[10], Jan Bartoš[1], Jacqueline Batley[4,5], David Edwards[5], Jaroslav Doležel[1], Hana Šimková[1]

[1] Institute of Experimental Botany, Centre of the Region Haná for Biotechnological and Agricultural Research, Šlechtitelů 31, 783 71 Olomouc, Czech Republic
[2] Department of Plant Sciences, University of California, Davis, USA
[3] Australian Centre for Plant Functional Genomics, University of Queensland, Brisbane, Australia
[4] School of Agriculture and Food Sciences, University of Queensland, Brisbane, Australia
[5] School of Plant Biology, University of Western Australia, Crawley, WA, Australia
[7] BioNano Genomics, 9640 Towne Centre Drive, San Diego, USA
[6] Laboratoty of Experimental Medicine, Institute of Molecular and Translational Medicine, Hněvotínská 5, 779 00 Olomouc, Czech Republic
[8] Diversity Arrays Technology Pty Ltd, University of Canberra, Bruce, Australia
[10] Wheat Genetics Resource Center, Department of Plant Pathology, Kansas State University

Summary

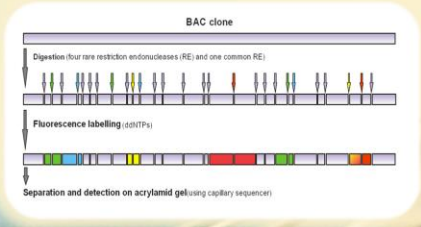
Bread wheat (*Triticum aestivum*) is a staple food for 40 % of human population. Polyploid nature of wheat genome (2n = 6x = 42, AABBDD) together with huge genome size (~17 Gb) and prevalence of repetitive sequences hamper its mapping and sequencing. To overcome these obstacles, BAC-by-BAC sequencing strategy based on chromosome-derived physical maps has been adopted by the International Wheat Genome Sequencing Consortium (IWGSC) as a key strategy towards obtaining a reference genome sequence.

Our work is focused on mapping and sequencing of the short arm of wheat chromosome 7D (7DS). Clones from a 7DS-specific BAC library were fingerprinted using SNaPshot-based HICF technology and automatically assembled into contigs using FPC software. Integration of initial 7DS physical map with that of *Aegilops tauschii* (D genome ancestor; Luo *et al*, 2013) provided a clue for further merging of automatically built contigs. The assembly was verified using LTC software. Physical map has been anchored on chromosome using different approaches of forward anchoring and also a 7DS specific optical map. Hitherto, we have positioned 1864 markers in 420 contigs and anchored another 37 marker-missing contigs through an optical map generated for 7DS. Thus we anchored 74 % of the 7DS physical map in total.

Our sequencing strategy involves pair-end sequencing pools of four non-overlapping BAC clones of the minimum tilling path (MTP) and mate-pair sequencng of MT plate pools (384 clones) using the Illumina HiSeq. Obtained pair-end reads were assembled into sequence contigs using assembler Sassy and mate-pair reads were applied to build scaffolds by SSPACE.

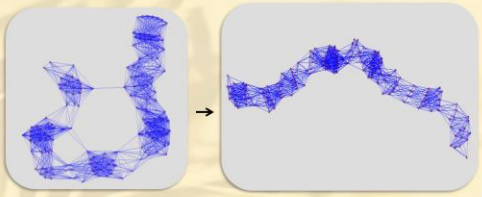
Physical map construction

1) SNaPshot fingeprinting of all BAC clones from the 7DS-specific BAC library



2) Automatic assembly using FPC software - Manual end-merging of automatically assembled contigs based on integration of *Ae. tauschii* whole-genome map with the map of the wheat 7DS chromosome arm

3) Verification of the assembly using **LTC** software – splitting incorrectly assembled (branching) contigs)



Physical map anchoring

Method		No. anch. markers
PCR screening of the 7DS BAC library		32
Integration with *Ae. tauschii*	• FPC graphical (prediction)	582
	• Based on 7DS sequences (validation)	461
Sequence based *in silico*	• SNP markers (Rimbert, unpublished)	708
	• SNP markers (Tiwari *et al.*, 2015)	170
	• DArTseq markers (Raman *et al.*, 2014)	493
Total*		**1864**

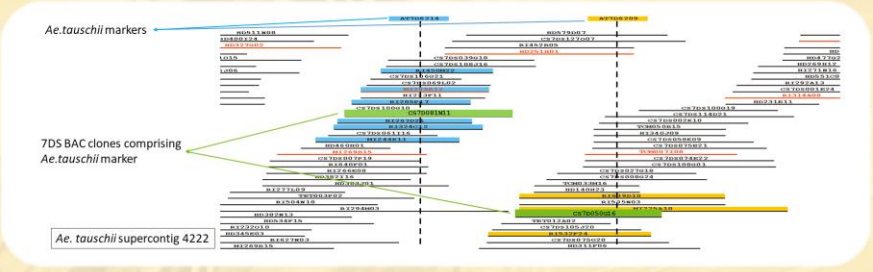
*) without markers anchored via graphical view in FPC

Ae. tauschii-guided *in silico* anchoring (FPC graphical)

- Based on FPC visualisation of overlaps between *Ae. tauschii* BAC clones comprising markers and clones of the wheat 7DS
- Marker assignment validated by aligning available sequences of the *Ae. tauschii* markers with the 7DS BAC clone assemblies

Parameters of the 7DS physical map

No. contigs > 2 BAC clones	931
No. clones in contigs	28 339
N50	528 kb
L50	205
7DS arm coverage	95%
No. MTP BAC clones	**4 608**

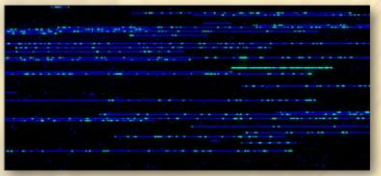


Sequencing of 7DS MTP

- 4 non-overlapping MTP BAC clones pooled → **1152 BAC pools**
- 96 pools sequenced in one lane of Illumina HiSeq → 100 bp pair-end reads, > 500x coverage
- BAC-end sequences are mapped to the contigs to identify ends of BAC assemblies and to assign contigs to specific BACs
- Scaffolding done using mate-pair data obtained from MTP-plate pools (384 clones)

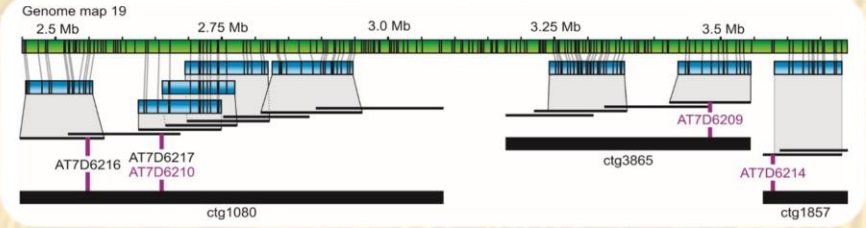
7DS optical map

- BioNano map was constructed as described by Hastie *et al.* (2013) from flow-sorted 7DS arm after digestion with Nt.BspQI (~12 recognition sites per 100 kb)
- The map is being, used for scaffolding physical map and sequence contigs, sizing gaps and identifying discrepancies in the physical map and sequence assembly



Optical map as a tool for anchoring and ordering contigs not only in non-recombining regions

- *Ae. tauschii* markers (AT7D…) are localized in the genetic map in a low-recombining region (black coloured markers co-segregate together, purple coloured markers co-segregate together)
- Sequences of MTP BAC clones of three contigs from the 7DS physical map were aligned to the optical map based on their Nt.BspQI pattern
- Optical map helped ordering and orienting 7DS BAC contigs comprising the clustering markers



- Another 37 marker-missing contigs of the 7DS physical map based on their position on the same optical map(s) as already anchored contigs through markers

Conclusions

- **Physical map covering 95 % of the 7DS chromosome has been assembled**
- **74 % of the physical map have been anchored → 1864 markers in 420 contigs**
- Optical map showed to be a useful tool to validate and anchor the physical map

References

Hastie *et al.* (2013): Rapid genome mapping in nanochannel arrays for highly complete and accurate de novo sequence assembly of the complex *Aegilops tauschii* genome. PLoS ONE 7(12): e55864
Luo *et al.* (2013): A 4-gigabase physical map unlocks the structure and evolution of the complex genome of Aegilops tauschii, the wheat D-genome progenitor. PNAS 110: 7940-7945
Raman *et al.* (2014): Genome-wild delineation of natural variation for pod shatter resistance in Brassica napus. PLoS One 9:7.
Tiwari *et al.* (2012): Endosperm tolerance of paternal aneuploidy allows radiation hybrid mapping of the wheat D-genome and measure of γ ray-induces chromosome breaks. PLoS ONE 7(11), e48815.

This work has been supported by Czech Science Foundation (P501/12/2554), MSMT CZ (National program of sustainability I, grant LO1204) and Internal Grant Agency PrF UP 2012-2013.

APPENDIX V

Completing reference sequence of the wheat chromosome arm 7DS

Tulpová Z, Toegelová H, Luo MC, Visendi P, Hayashi S, Kilian A, Tiwari VK, Kumar A, Hastie AR, Leroy P, Rimbert H, Abrouk M, Bartoš J, Batley J, Edwards D, Doležel J, Šimková H

In: Abstracts of the International Conference "Plant and Animal Genome XXIV". P.0826. Sherago International, Inc., San Diego, 2016.

Completing Reference Sequence of the Wheat Chromosome Arm 7DS

Zuzana Tulpová[1], Helena Staňková[1], Ming-Cheng Luo[2], Paul Visendi[3], Satomi Hayashi[3], Andrzej Kilian[4], Vijay K. Tiwari[5], Ajay Kumar[6], Alex Hastie[7], Philippe Leroy[8], Hélène Rimbert[8], Michael Abrouk[1], Jan Bartoš[1], Jacqueline Batley[4], David Edwards[3], Jaroslav Doležel[1], Hana Šimková[1]

[1] Institute of Experimental Botany, Centre of the Region Haná for Biotechnological and Agricultural Research, Šlechtitelů 31, CZ-783 71, Olomouc, Czech Republic
[2] Department of Plant Sciences, University of California, Davis, CA 95616, USA
[3] School of Agriculture and Food Sciences, University of Queensland, Brisbane, QLD 4072, Australia
[4] Diversity Arrays Technology, Ltd, University of Canberra, Bruce, ACT 2617, Australia
[5] Wheat Genetics Resource Center, Department of Plant Pathology, Kansas State University, Manhattan, KS, 66506
[6] Department of Plant Sciences, North Dakota State University, Fargo, ND, 58108, USA
[7] BioNano Genomics, 9640 Towne Centre Drive, San Diego, CA 92121, USA
[8] INRA/UBP, UMR1095, Genetics, Diversity and Ecophysiology of Cereals, F-63100 Clermont-Ferrand, France

Summary

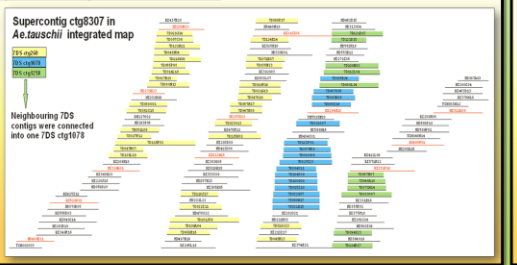
Bread wheat genome is characterized by a high complexity (~17 Gb), polyploidy (AABBDD) and a high content of repetitive sequences (>90%), which hampers genome mapping and sequencing. To overcome these obstacles, BAC-by-BAC sequencing strategy based on chromosome-derived physical maps has been adopted by the International Wheat Genome Sequencing Consortium (IWGSC) as the key strategy towards obtaining a reference genome sequence.

In the framework of this effort, we focus on mapping and sequencing of the 7DS chromosome arm (381 Mb). Clones from a 7DS-specific BAC library were fingerprinted and assembled into contigs using FPC and the map assembly was validated by LTC. Integration of the 7DS physical map with that of *Aegilops tauschii* (D genome progenitor) provided a clue for further merging and landing the contigs on the chromosome. Our BAC sequencing strategy involved sequencing pools of four non-overlapping clones using the Illumina HiSeq. Pair-end reads thus obtained were assembled using a custom assembler Sassy. Assembling and pool deconvolution were supported by BAC-end sequences and mate-pair data generated for pools of 384 clones. In the resulting assembly, we achieved 1.9 scaffolds per BAC clone on average and scaffold N50 of 116 kb. A BioNano genome map has been constructed for the 7DS and used to validate the physical map, anchor and orient BAC contigs, size gaps, deconvolute pools and support sequence scaffolding.

The current 7DS assembly consisting of 9,084 scaffolds was annotated using TriAnnot pipeline v5.1. Resulted data show that the 7DS assembly comprises 3,454 predicted genes (1,916 High Confidence genes; 1,347 Low Confidence genes and 201 Medium Confidence genes). In total 63,262 pseudogenes have been detected (frameshift, no start codon, fragments). 84% of total scaffold sequence length are transposable elements (TEs). These results provide the first complex insight into the structural and functional partitioning of the 7DS arm.

Physical map construction

- **SNaPshot fingeprinting** of **49,152 BAC clones from the 7DS-specific BAC library** (insert size 113 kb, coverage 12.1x)
  → **39, 765** useful fingerprints
- **Automatic assembly using FPC** up to $1e^{-45}$
- **Manual end-merging** of automatically assembled contigs based on **integration of *Aegilops tauschii*** whole-genome map (Luo *et al.*, 2013) **with** the map of the bread wheat **7DS** chromosome arm
- **Verification of the assembly using LTC** (*Linear Topological Contigs*) – splitting incorrectly assembled (branching) contigs



Physical map anchoring

Strategy	Marker type and map	No. markers	No. ctg anchored
• **Forward anchoring**			
- Manual (3D pools screening)	SSR genetic (GrainGenes)	23	27
	STS genetic (*owm*, Staňková *et al.*, 2015)	11	8
- *In silico* (sequence-based)	SNP genetic *Ae. tauschii* (Luo *et al.*, 2013)	516	302
	DArTseq genetic (Kilian, unpublished)	556	295
	RJM – RH CS (Tiwari, unpublished)	205	150
	SNP – RH CS (Tiwari, unpublished)	202	144
	DArT – RH *Ae. tauschii* (Kumar *et al.*, 2015)	51	14
• **Reverse anchoring**			
- Based on BAC sequences	STS – RH *Ae. tauschii*	29	26

Sequencing of 7DS MTP

- Pools of four non-overlapping MTP BAC clones → **1,152 BAC pools**
- 96 pools sequenced in one lane of Illumina HiSeq → 100 bp pair-end reads, > 500x coverage – **Sassy**
- BAC-end sequences are mapped to the contigs to identify ends of BAC assemblies and to assign contigs to specific BACs
- Scaffolding done using mate-pair data obtained from MTP-plate pools (384 clones) - **SSPACE**

Physical map statistics

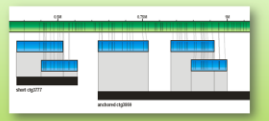
No. contigs > 2 BAC clones	904
No. contigs > 5 clones	652
Assembly length	360 Mb
7DS arm coverage	95%
Contig N50	548 kb
Contig L50	199
No. MTP BAC clones	**4,608**

Sequence assembly statistics

Contigs per BAC	1 - 17
Average No. contigs/BAC	1.9
Median No. contigs/BAC	1.5
Average scaffold size	56 kb
Scaffold N50	116 kb
% N in assembly	2.9 %

Sequences of short BAC contigs

Short BAC contigs (< 6 clones) are frequently excluded from the map assembly and are not sequenced, mainly because of difficulties with their positioning on the chromosome. Aiming to investigate potential contribution of short BAC contigs to the assembly, we included **clones from all contigs > 2 BAC clones in the 7DS MTP**. We considered as contributing those contigs, whose belonging to the 7DS was confirmed by markers or through anchoring to the 7DS BioNano map.

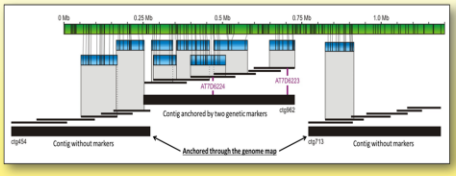


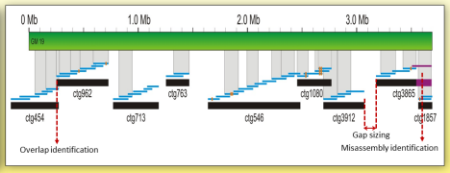
	3-BAC	4-BAC	5-BAC	Total
No. short contigs in 7DS physical map	169	58	25	252
No. anchored contigs	32	16	17	65
• through markers (+ BioNano map)	22	13	13	48
• through BioNano map only	10	3	4	17
Total length of short contigs [kbp]	25,800	10,605	4,214	40,873
• length of anchored (% assembly)	4,867 (1,4%)	3,254 (0,9%)	2,736 (0,8%)	11,111 (3%)
• length of non-anchored (% assem.)	21,187 (5,9%)	7,351 (2,0%)	1,478 (0,4%)	29,762 (8%)

Contigs containing 3-5 BAC clones represent 11% of the 7DS physical map assembly → 27% of them could be anchored to the 7DS. **If excluded → 3% gaps**
Contigs as short as 3 BAC clones (2 MTP clones) can be efficiently positioned through the BioNano map.

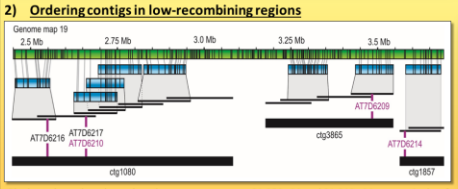
BioNano genome map to improve the physical map

- BioNano map was constructed as described by Staňková *et al.* (2016) from flow-sorted 7DS arm after labelling Nt.BspQI nicking sites (GCTCTTC motif, ~12 sites per 100 kb)
- Usefull tool for: **1) Scaffolding and validating the physical map**
  - Sizing gaps and identifying overlaps; anchoring and orienting contigs, identifying misassemblies
  - **36 BAC contigs representing 4% assembly length** were newly anchored through the BioNano map

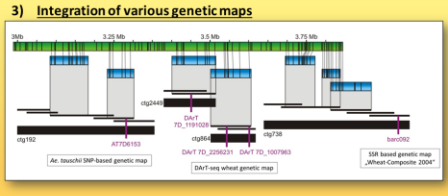




2) Ordering contigs in low-recombining regions



3) Integration of various genetic maps

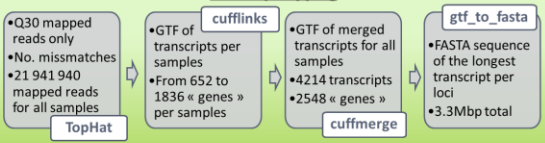


Black- and purple-coloured markers co-segregate at 81.858 cM and 81.812 cM, respectively, of the *Ae. tauschii* genetic map.

Annotation of the 7DS assembly

- **TriAnnot pipeline version v5.1p02**
- Supported by 7DS RNAseq data obtained from RNAseq mapping on the 7DS scaffolds
- RNAseq mapping was performed under 15 conditions (Pingault *et al.*, 2015)
- Read mapping & Transcript identification: Tuxedo suite

RNAseq mapping



Total No. 7DS transcripts	2,584
• No. transcripts mapped on 7DS scaffolds	2,364 (91%)
• No. transcripts mapped on 7DS CSS sequences (IWGSC 2014)	1,556 (60%)

Annotation statistics

No. scaffolds in 7DS assembly	9,084
• No. analyzed scaffolds (>10 kb)	6,152
• No. scaffolds with at least one gene	1,873 (~30%)
No. predicted protein-coding genes	3,454
• No. high confidence genes	1,916 (~55.5%)
No. pseudogenes	63,262
No. TEs	59,668 (15 families)
• Class I – LTR: *Copia & Gypsy*	9,218 + 15,796 (~42%)
• Class II – TIR: *CACTA*	13,908 (~23%)

Conclusions

- **Physical map covering 95% of the 7DS chromosome arm** has been assembled
- **77% of the physical map** have been **anchored** → 508 contigs through 1,601 markers, 36 contigs through BioNano map
- 4608 MTP BAC clones have been sequenced and **assembled into 9,084 scaffolds**
- Pseudomolecule construction and annotation are underway → **preliminarily 3,454 predicted protein-coding genes, 63,262 pseudogenes and 59,668 TEs** (84% sequence length)
- The sequence assembly comprises **91% of gene models identified in 7DS CSS sequences** (IWGSC, 2014)
- BioNano genome map showed a useful tool to validate and scaffold the physical map and sequence assemblies

References

Pingault et al. (2015): Deep transcriptome sequencing provides new insights into the structural and functional organization of the wheat genome. Genome Biol. 16: 29.
Kumar et al. (2015): Radiation hybrid maps of the D-genome of *Aegilops tauschii* and their application in sequence assembly of large and complex plant genomes. BMC Genomics 16: 800.
Luo et al. (2013): A 4-gigabase physical map unlocks the structure and evolution of the complex genome of Aegilops tauschii, the wheat D-genome progenitor. PNAS 110: 7940-7945
Staňková et al. (2016): BioNano Genome Mapping of Individual Chromosomes Supports Physical Mapping and Sequence Assembly in Complex Plant Genomes, Plant Biotech. J., doi: 10.1111/pbi.12513
Staňková et al. (2015): Chromosomal genomics facilitates fine mapping of a Russian wheat aphid resistance gene. Theor. Appl. Genet. 128 (7), 1373 – 1383.
IWGSC (2014): A chromosome-based draft sequence of the hexaploid bread wheat. Science 18: 345, 6194.

This work has been supported by the Czech Science Foundation (P501/12/2554) and MSMT CZ (National program of sustainability I, grant LO1204).

APPENDIX VI

Structural variation of a wheat chromosome arm revealed by optical mapping

Tulpová Z, Toegelová H, Vrána J, Hastie AR, Lukaszewski AJ, Kopecký D, Lapitan NLV, Batley J, Edwards D, IWGSC, Doležel J, Šimková H

In: Abstracts of the International Conference "Plant and Animal Genome XXV". P.0837. Sherago International, Inc., San Diego, 2017.

Structural Variation of a Wheat Chromosome Arm Revealed by Optical Mapping

Tulpová, Z.[1], Toegelová, H.[1], Vrána J.[1], Hastie, A.R.[2], Lukaszewski, A.J.[3], Kopecký, D.[1], Lapitan, N.L.V.[4], Batley, J.[5], Edwards, D.[5], IWGSC[6], Doležel, J.[1], Šimková, H.[1]

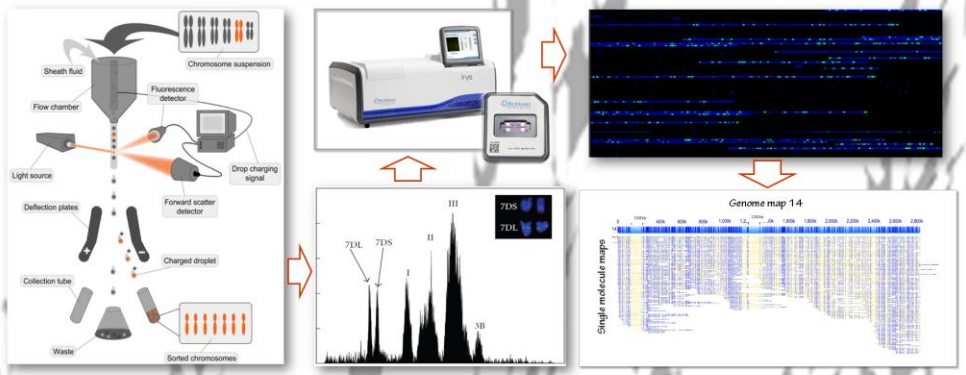
[1]Institute of Experimental Botany, Šlechtitelů 31, CZ-78371 Olomouc, Czech Republic
[2]BioNano Genomics, 9640 Towne Centre Drive, San Diego, CA 92121, USA
[3]Dept. of Botany & Plant Sciences, University of California Riverside ,CA 92521, USA
[4]Department of Soil and Crop Sciences, Colorado State University, Fort Collins, CO 80523, USA
[5]School of Plant Biology, University of Western Australia, Crawley, WA 6009, Australia
[6]IWGSC, 2841 NE Marywood Ct, Lee's Summit, MO 64086, USA

BACKGROUND

Optical mapping on the Irys platform of BioNano Genomics creates maps of short sequence motives (typically GCTCTTC) along DNA stretches hundreds to thousands kilobases in length that can be used to validating and improving DNA sequence assemblies and studying structural variation among individuals, cultivars or species. Coupling the optical mapping with flow sorting of individual chromosomes of bread wheat enables zooming in on particular regions of the complex (~15 Gb) hexaploid genome, making the studies more focused and cost-efficient.

CHROMOSOME MAPPING ON NANOCHANNEL ARRAY

The short arm of bread wheat (*Triticum aestivum*) chromosome 7D (7DS, 381 Mb) was purified by flow cytometry from ditelosomic lines of two accessions - cv. Chinese Spring (CS) and CI2401 – in which the 7DS arm is maintained as a pair of stable telocentric chromosomes. Using Irys platform, we generated 68 and 78Gb size-filtered (>150kb) data for CS and CI2401, respectively, which were assembled into genome maps as described in Staňková et al. (2016). Another optical map was prepared for the 7DL arm (346Mb) of cv. Chinese Spring from 117Gb size-filtered data.



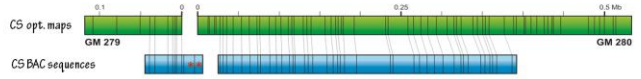
	Molecule coverage	Molecule N50 (kb)	No. genome maps	Avg. genome map length (kb)	Genome map N50 (kb)	Total map length (Mb)
7DS-CS	180x	344	371	900	1,300	350
7DS-CI2401	206x	219	468	765	1,355	358
7DL-CS	338x	210	364	869	1,288	316

STUDYING STRUCTURAL VARIATION BETWEEN 7DS ARMS FROM CV. CHINESE SPRING AND CI2401

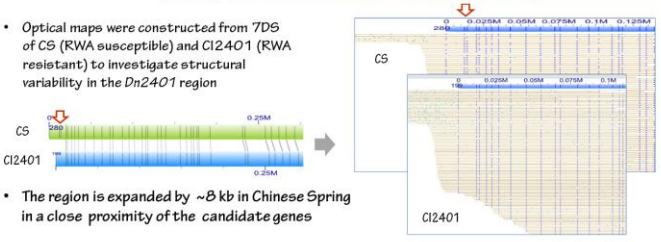
83% of the 7DS-CI2401 optical map length could be aligned to the 7DS-CS map at high stringency indicating high sequence similarity of the 7DS telosomics. Alignment of the optical maps to a reference genome generated from cv. Chinese Spring by the IWGSC (IWGSC v1.0) revealed a striking variability between the two accessions in (sub)telomeric and (peri)centromeric regions. Structural variation was also found around a Russian Wheat Aphid resistance (RWA) resistance gene.

LOCAL STRUCTURAL VARIATION AROUND A RWA RESISTANCE GENE

- Russian Wheat Aphid (*Diuraphis noxia*) resistance gene *Dn2401* has been mapped on 7DS to a region of 0.81 cM spanned in cv. Chinese Spring by five BAC clones delimiting a region of ~350kb (Staňková et al. 2015). The clones were sequenced and annotated → two candidate genes (red asterisks)

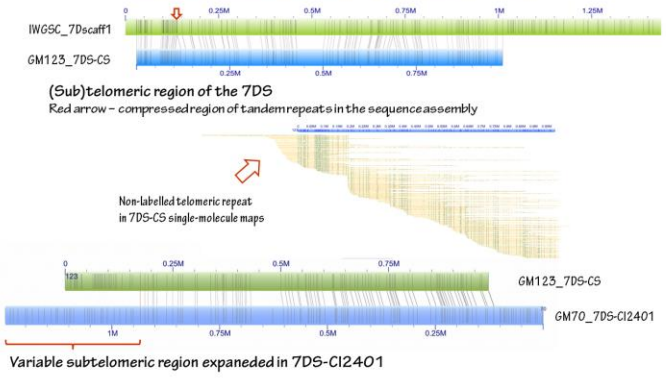


- Optical maps were constructed from 7DS of CS (RWA susceptible) and CI2401 (RWA resistant) to investigate structural variability in the *Dn2401* region

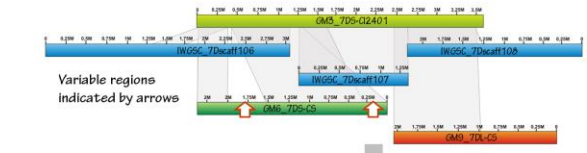


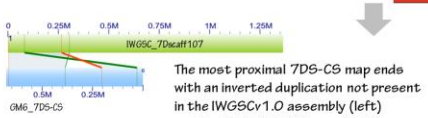
- The region is expanded by ~8 kb in Chinese Spring in a close proximity of the candidate genes

STRUCTURAL VARIATION IN THE 7DS TELOMERIC REGION

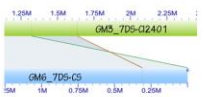


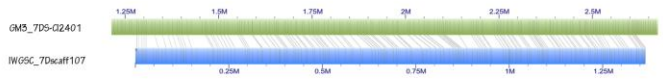
STRUCTURAL VARIATION IN THE 7D CENTROMERIC REGION



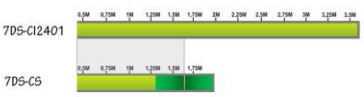


The most proximal 7DS-CS map ends with an inverted duplication not present in the IWGSCv1.0 assembly (left) and the 7DS-CI2401 map (right)





- The GM3_7DS-CI2401 map spans the whole 7D centromeric region. This validates the sequence assembly and suggests that the break giving rise to the 7DS telocentric in CI2401 was more distal from the centromere.



- Proposed structure of centromeric ends of the two 7DS telocentrics. The CS telocentric ends with inverted repeat.

REFERENCES:

Staňková, H. et al.: Chromosomal genomics facilitates fine mapping of a Russian wheat aphid resistance gene. – Theor. Appl. Genet. 128: 1373-1383, 2015.

Staňková, H. et al.: BioNano genome mapping of individual chromosomes supports physical mapping and sequence assembly in complex plant genomes. – Plant Biotech. J. 14: 1523-1531, 2016.

APPENDIX VII

Poziční klonování genu pro rezistenci ke mšici zhoubné (*Diuraphis noxia*): konstrukce vysokohustotní genetické mapy

Staňková H, Valárik M, Lapitan N, Berkman P, Edwards D, Luo MC, Tulpová Z, Kubaláková M, Stein N, Doležel J, Šimková H

In: Sborník abstrakt, Bulletin České společnosti experimentální biologie rostlin, "6. Metodické dny". Seč, Česká republika, 2014

[In Czech]

Poziční klonování genu pro rezistenci k mšici zhoubné (*Diuraphis noxia*): konstrukce vysokohustotní genetické mapy

Helena Staňková[1], Miroslav Valárik[1], Nora Lapitan[2], Paul Berkman[3], David Edwards[3], Ming-Cheng Luo[4], Zuzana Tulpová[1], Marie Kubaláková[1], Nils Stein[5], Jaroslav Doležel[1], Hana Šimková[1]

[1] Centrum regionu Haná pro biotechnologický a zemědělský výzkum, Ústav experimentální botaniky, Šlechtitelů 31, 783 71, Olomouc - Holice, Czech Republic
[2] Department of Soil and Crop Sciences, Colorado State University, Fort Collins, Colorado, 80524, USA
[3] Australian Centre for Plant Functional Genomics, University of Queensland, Brisbane, QLD 4072, Australia
[4] Department of Plant Sciences, University of California, Davis, CA 95616, USA
[5] Leibniz Institute of Plant Genetics and Crop Plant Research, Department Genebank, AG Genome Diversity, D-06466 Gatersleben, Germany

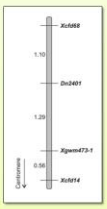
Úvod

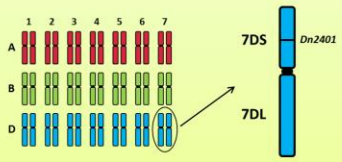
Pšenice setá (*Triticum aestivum* L.) představuje jednu z ekonomicky nejvýznamnějších kulturních plodin, jež poskytuje zdroj potravy pro 35% obyvatel světa. Jedná se o allohexaploidním druh (2n = 6x = 42) s celkovou velikostí genomu téměř 17 x 10^9 bp. Genom je tvořen třemi homeologními subgenomy (A, B a D) a jeho podstatnou část (přes 80%) tvoří repetitivní sekvence. Všechny výše zmíněné vlastnosti pšeničného genomu znesnadňují jeho analýzu, genetické i fyzické mapování, sekvenování či poziční klonování. Třídění jednotlivých chromozómů a jejich ramen pomocí průtokové cytometrie umožňuje rozložit tento obrovský genom na malé a snadno analyzovatelné části.

Na krátkém rameni chromozómu 7D (7DS) pšenice byl identifikován gen *Dn2401* pro rezistenci k mšici zhoubné (*Diuraphis noxia*). Mšice zhoubná je jedním z nejvýznamnějších škůdců pšenice a ječmene. Chemické i biologické postupy hubení nejsou v případě mšice zhoubné dostatečně účinné. Z tohoto důvodu se jeví jako nejvýhodnější způsob ochrany pěstování odrůd nesoucích geny pro rezistenci vůči tomuto škůdci. Konstrukce vysokohustotní genetické mapy pokrývající oblast zkoumaného genu je nezbytná pro jeho následné poziční klonování, tedy izolaci genu na základě jeho pozice na genetické či fyzické mapě. Za účelem konstrukce mapy byla vyvinuta metoda pro cílené odvozování vysoce specifických markerů ze zkoumané oblasti v podmínkách polyploidního genomu. Naše metoda využívá syntenie mezi pšenicí a jejími příbuznými druhy (ječmen, *Brachypodium*, rýže, čirok, *Aegilops tauschii*) v kombinaci se sekvencemi jednotlivých chromozómů skupiny 7, získanými celochromozómovým neuspořádaným (shotgun) sekvenováním. Tímto způsobem jsme získali nové markery specifické pro oblast studovaného genu. Následně byl za pomoci těchto markerů identifikován kontig ve fyzické mapě ramene 7DS, jenž překlenuje oblast genu *Dn2401*. Klony z tohoto kontigu byly osekvenovány a následně anotovány. Anotace sekvencí odhalila přítomnost několika kandidátních genů.

Gen *Dn2401*

- Podmiňuje rezistenci k mšici zhoubné (*Diuraphis noxia*)
- Identifikován v linii CI2401
- Lokalizován na krátkém rameni chromozómu 7D pšenice (7DS)
- Zamapován mezi markery *cfd68* a *gwm473* (2,39 cM)





Mapovací populace F2

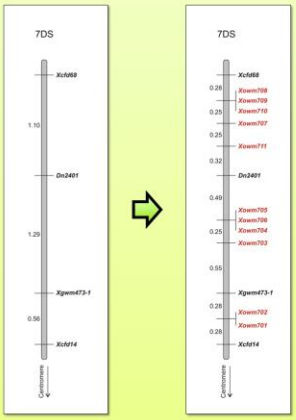
- Křížení **CI2401** (rezistentní) x **Glupro** (citlivá)
- 158 jedinců

Markery *owm*

- Celková spěšnost navrhování primerů specifických pro rameno 7DS – 86% (138 ze 161 párů)
- Pouze 11 párů primerů poskytovalo produkt polymorfní mezi rodiči mapovací populace

⇩

- Odvozeno 11 nových markerů
  - ***owm701*** – ***owm710*** – SNP markery
  - ***owm711*** – délkový polymorfismus
- Interval mezi hraničními markery zkrácen z 2,39 cM na 0,81 cM



Konstrukce vysokohustotní genetické mapy

- využití syntenie mezi pšenicí a příbuznými druhy - ječmen, *Brachypodium*, rýže, čirok, *Aegilops tauschii*

Genomické zdroje pro vývoj genetických markerů

- **GenomeZipper**
  - Soubor genů ječmene seřazených na základě kolinearity s osekvenovanými genomy rýže, *Brachypodia* a čiroku a ukotvený na genetickou mapu ječmene tvořenou EST markery
- **Genetická mapa *Aegilops tauschii***
  - Tvořena geneticky zamapovanými SNP markery
  - Většina markerů odvozena z genů
- **Sekvence chromozómů skupiny 7**
  - Získány celochromozómovým neuspořádaným (shotgun) sekvenováním

GenomeZipper

Marker	cM	Brachypodium	Rýže	Čirok	Ječmenový EST
-	-	Bradi1g44100.1	Os06g0557200	-	276
-	-	Bradi1g44110.1	Os06g0558400	Sb10g031100.1	8010.30524
-	-	-	Os06g0558500	-	-
0_0496	71.71	Bradi1g44200.1	Os06g0557300	-	1098
-	-	Bradi1g44250.1	Os06g0558200	Sb10g030040.1	18538.13626
-	-	Bradi1g44380.1	Os06g0558100	Sb10g030230.1	45494.25645
-	-	-	Os06g0557700	-	-
-	-	Bradi1g44380.1	Os06g0555300	Sb10g029200.1	43211.39112
-	-	-	Os06g0556200	-	-
0_1267	71.71	Bradi1g44100.1	Os06g0557100	Sb10g030180.1	37077.17759
-	-	Bradi1g44200.1	Os06g0554700	Sb10g029170.1	3189.3190
-	-	Bradi1g44220.1	Os06g0553900	Sb10g029120.1	-
-	-	Bradi1g44230.1	Os06g0553100	Sb10g029090.1	-
-	-	-	-	Sb10g029080.1	-
0_0963	71.71	Bradi1g44400.1	Os03g0267500	Sb10g028850.1	19371
-	-	-	Os06g0548700	-	50581.9698.39277
-	-	Bradi1g44480.1	Os06g0548500	-	26968
-	-	Bradi1g48720.1	-	Sb10g031800.1	-
-	-	Bradi1g48730.1	Os06g0525200	Sb10g031810.1	8409.31629
-	-	Bradi1g48740.1	Os06g0529900	Sb10g031820.1	3647
-	-	Bradi1g48750.1	-	Sb10g031830.1	23662
0_0832	71.71	Bradi1g48760.1	-	-	15554

Postup vývoje lokusově specifických genetických markerů

Identifikace kolineární oblasti v GenomeZipperu a na genetické mapě *Ae. tauschii*

⇩

Výběr ječmenových ESTů a genetických markerů *Ae. tauschii* z dané oblasti

⇩

Nalezení sekvencí homologních k ESTům a markerům *Ae. tauschii* v sekvencích chromozómů skupiny 7 (7A, 7B, 7D)

⇩

Návrh primerů specifických pro 7DS na základě polymorfismů mezi subgenomy A, B a D

⇩

Optimalizace podmínek PCR

⇩

PCR na rodičích mapovací populace (CI2401, Glupro) a sekvenování produktů PCR

⇩

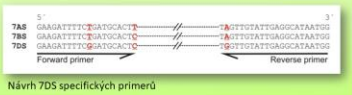
Identifikace polymorfismů mezi rodiči mapovací populace

⇩

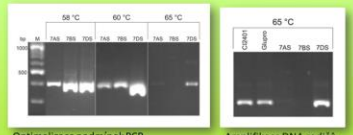
Mapování markerů v F2 mapovací populaci CI2401 x Glupro

⇩

Markery *owm*

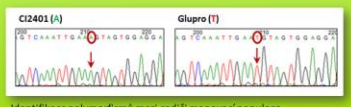


Návrh 7DS specifických primerů



Optimalizace podmínek PCR

Amplifikace DNA rodičů



Identifikace polymorfismů mezi rodiči mapovací populace

Fyzické mapování a sekvenování

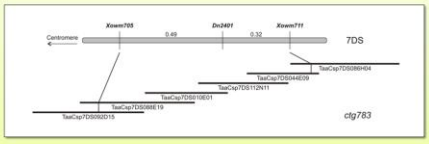
BAC knihovna DNA z ramenen 7DS

- Skríning 3D poolů hraničními markery *owm705* a *owm711*
- Identifikace pozitivních BAC klonů

⇩

Fyzická mapa 7DS

- Nalezení BAC kontigu, jenž obsahuje pozitivní BAC klony → *ctg783*



⇨

Sekvenování

- MTP klony mezi hraničními markery
- Sekvenování technologií Illumina
- Sekvence skládány programem Sassy

Anotace sekvencí BAC klonů

- Porovnání sekvencí s databází anotovaných genů ječmene
- Nalezeny dva kandidání geny
  - salycylát o-methyltransferáza (součást metabolické dráhy kyseliny salycilové)
  - epoxid hydroláza (pozorována zvýšená exprese při napadení rostliny mšicí)

⇩

- Kandidátní geny budou sekvenování za použití DNA obou rodičů mapovací populace → hledání polymorfismů

Souhrn

- Bylo odvozeno 11 nových markerů
  → Zamapovány do oblasti genu *Dn2401*
- Interval vymezující oblast genu byl zkrácen z 2,39 cM na 0,81 cM
- Nové hraniční markery umožnily identifikovat kontig *ctg783* překlenující studovanou oblast
- Sekvenování BAC klonů z *ctg 783* odhalila přítomnost dvou kandidátních genů

Tato práce je podporována grantem GA ČR P501/12/2554, MŠMT ČR a EU (Operační program Výzkum a Vývoj pro Inovace ED0007/01/01) a Interní Grantovou Agenturou PřF 2014-001.

Palacký University Olomouc

Faculty of Science

Department of Cell Biology and Genetics

and

Centre of Plant Structural and Functional Genomics

Institute of Experimental Botany CAS

Zuzana Tulpová

Sequence and functional analysis of the short arm of wheat chromosome 7D

P1527 - Molecular and Cellular Biology

Summary of Ph.D. thesis

Olomouc 2019

Ph.D. Thesis was carried out at the Department of Cellular biology and Genetics, Faculty of Science, Palacký university Olomouc, between years 2013-2019

Candidate: **Mgr. Zuzana Tulpová**

Supervisor: **Ing. Hana Šimková, CSc.**

Reviewers: **RNDr. Helena Štorchová, CSc.**

Institute of Experimental Botany CAS, Plant Reproduction Laboratory

RNDr. Aleš Kovařík, CSc.

Institute of Biophysics CAS, Department of Molecular Epigenetics

The evaluation of this Ph.D. thesis was written by ……
……...

The summary of the Ph.D. thesis was sent for distribution on ……

The oral defence will take place on …….…….in front of the commission for the Ph.D. study of the study program Molecular and Cellular Biology in ……...

The Ph.D. thesis is available in the Library of the biological departments of the Faculty of science at Palacký University Olomouc, Šlechtitelů 11, Olomouc-Holice.

Prof. RNDr. Zdeněk Dvořák, DrSc. et Ph.D.
Chairman of the Commission for the Ph.D.
Thesis of the Study Program Molecular and Cellular Biology
Department of Cellular biology and Genetics, Faculty of Science
Palacký University Olomouc

Content

1. Introduction

Bread wheat (*Triticum aestivum* L.) was one of the first domesticated plants. It is currently grown on the largest agricultural area and is a staple food for 40 % of the world's population. Bread wheat is a monocotyledonous species belonging to the family *Poaceae*, subfamily *Pooideae* and the tribe *Triticeae*. It is an allohexaploid species (2n = 6x = 42), whose genome arose through spontaneous hybridization of three progenitors resulting into a highly complex genome of ~16 Gb (Doležel *et al*., 2018) consisting of three homoeologous sub-genomes A, B and D and comprising over 85 % repetitive DNA (IWGSC, 2018). One of the major obstacles of studying the wheat genome was the high similarity of all three sub-genomes. However, it has been overcome by the use of flow cytometry, with which it is possible to sort not only individual chromosomes but also their arms (Doležel *et al.*, 2007). This can significantly reduce the complexity of the genome and simplify its analysis and assembling of a reference sequence, the knowledge of which can dramatically speed up breeding of new wheat varieties with improved agronomical traits.

A physical contig map consists of overlapping clones from long-insert DNA libraries, usually cloned in the bacterial artificial chromosome (BAC) vector. The overlaps are determined based on the similarity of restriction spectra generated for each clone. Using such a map, it is possible to select clones of minimal tilling path (MTP), consisting of BAC clones that continuously cover the entire chromosome(s). Positions of the physical map segments (contigs) on the chromosomes are determined by their anchoring, most often with the help of genetic markers. This allows subsequent arrangement of the whole genome sequence. In 2005, the International Wheat Genome Sequencing Consortium (IWGSC) adopted a strategy for generating a wheat genome reference sequence based on sorted chromosomes or their arms. Chromosome-specific BAC libraries served as the genomic resource for the construction of physical maps, starting with that of the largest wheat chromosome 3B (Paux *et al*., 2008). Within the framework of the IWGSC project, physical maps for all wheat chromosomes (IWGSC, 2018) were gradually constructed and many of them have been sequenced. Chromosomal physical maps and BAC clone sequences have become important components in assembling the reference wheat genome sequence (IWGSC, 2018). Although nowadays, the sequencing approach based on BAC libraries and physical maps is overcome, physical maps and BAC clones associated with them remain a favourable genomic resource for targeted analyses of narrowly delimited regions of the genome, particularly in gene cloning projects.

The knowledge of the wheat reference sequence has given an access to more than a hundred thousand genes, including description of their genomic context and identification of regulatory regions (IWGSC, 2018). This opened the door to a more efficient identification of genes for important agronomic traits and for faster breeding of new varieties. The benefit of the reference sequence can be demonstrated on the example of the short arm of wheat chromosome 7D, which is the subject of the presented work. This arm carries a variety of genes or loci underlying agronomically important traits, such as yield components

or resistances to a wide range of fungal pathogens or insect pests, including Russian wheat aphid.

Nowadays, the Russian wheat aphid (*Diuraphis noxia*, Kurdjumov) is spread all over the world and is one of the most important pests not only of wheat and barley, but also of other plants from 43 genera comprising up to 140 different cultivated and wild grass species (Yazdani *et al*., 2017). During its feeding, aphid is removing plant photoassimilates, which results in chlorosis, longitudinal streaking along the main leaf vein, head trapping, substantial reduction in biomass and, in severe cases, plant death (Burd and Burton, 1992). One of the typical symptoms of the aphid infestation is leaf rolling, which provides aphids an additional advantage, because it shelters their colonies against their natural predators as well as insecticide spraying, which reduces the effect of chemical pest treatment. Thus the most effective strategy to handle aphid attacks lies in using natural sources of aphid resistance to breeding of novel resistant cultivars.

Wheat chromosome arm 7DS carries several genes underlying resistance to Russian wheat aphid, namely *Dn1*, *Dn2* (Du Toit, 1987), *Dn5* (Du Toit *et al.*, 1995), *Dn6* (Nkongolo *et al*., 1991), *Dn8* (Liu *et al*., 2001), *Dnx* (Harvey and Martin, 1990), *Dn626580* (Valdez *et al*., 2012) and *Dn2401* (Voothuluru *et al*., 2006). The current work focused on the latter one, which has been identified in wheat line CI2401 and localized on the 7DS arm in a proximity of marker *gwm111* (Fazel-Najafabadi *et al*., 2015). In a previous study by Staňková *et al*. (2015), a high-density genetic map was constructed to facilitate positional cloning of the *Dn2401* gene. Newly derived markers delimited a gene interval that has been spanned by five BAC clones, which became the starting point for the study presented in this dissertation.

2. Aims of the thesis

2.1 Anchoring and validation of the physical map of the short arm of wheat chromosome 7D

The first aim of the thesis is to anchor a 7DS physical contig map to the chromosome through integration with several types of genetic and physical genomic resources, including a physical map of the D genome progenitor, *Aegilops tauschii*. This integration will be used to compare structure of the 7DS between bread wheat and its ancestor.

2.2 Sequencing of the chromosome arm 7DS and assembling of the reference sequence

The second aim of the work will be sequencing of a minimal set of BAC clones continuously covering the entire 7DS arm, assembling and annotating of the obtained sequence. The generated data and other genomic resources will be used to assemble and validate the reference sequence of the 7DS arm.

2.3 Positional cloning of a Russian wheat aphid resistance gene

The third aim of the thesis will be application of the BAC clone assembly in positional cloning of a Russian wheat aphid resistance gene and in analyses of other regions of interest.

3. Material and Methods

The presented work is focused on the short arm of the wheat chromosome 7D. One of the aims of the work was to obtain a quality reference sequence. For this purpose, a long-insert library cloned in the BAC vector was generated from sorted chromosome arms 7DS from wheat cultivar Chinese Spring (Šimková *et al*., 2011) and used to construct a physical contig map. Map contigs have been anchored using genetically mapped markers, either by manual PCR screening of the library on three-dimensional pooled samples, or by *in silico* anchoring using available BAC clone sequences or an optical map. Besides, an unconventional approach was applied, employing integration of physical maps of two related species.

Manual screening of libraries with mapped genetic markers was performed using PCR on 768 three-dimensional pooled samples prepared from 49,152 BAC clones of the 7DS arm-specific library. The use of pooled samples significantly improved library screening. Positive clones identified after deconvolution of pooled data were assigned to individual contigs, allowing their positioning on the genetic map. *In silico* anchoring was based on the search for homology between marker sequences and available BAC clone sequences from the 7DS MTP. SNPs and DArT markers with known positions in genetic maps of wheat (Rimbert *et al.*, 2017; Tulpová *et al.* 2019a) or *Ae. tauschii* (Luo *et al*., 2013) as well as SNP markers from a wheat radiation hybrid map were used (Tiwari *et al*., 2016). Sequencing of BAC clones was performed by Illumina technology on the HiSeq platform, generating pair-end (PE) as well as mate-pair (MP) reads. Sequencing was performed with pooled samples (four clones for PE and 384 clones for MP reads) and data from Sanger sequencing of the paired ends of individual BAC clones was used to sequence deconvolution.

An important addition to the mentioned anchoring strategies was the use of the integration of two physical maps and the 7DS optical map. The 7DS integrated physical map was created by integrating the 7DS physical map of hexaploid wheat with the anchored physical map of *Aegilops tauschii* (Luo *et al*., 2013), the donor of the D sub-genome. The integration of the maps allowed positioning of 7DS physical map contigs on a genetic map used for anchoring of the *Ae. tauschii* physical map and direct comparison of the structure of the arm 7DS in wheat and its progenitor. The optical map of the wheat 7DS (Staňková *et al*., 2016) was used to validate the physical map and anchoring of marker-free contigs. The possibility of direct comparison of the structure of 7DS in wheat and its progenitor also helped in a more detailed analysis of the (peri)centromeric region. In addition to centromere, another analysis employing available wheat reference sequence and published data from CENH3 ChIP-seq wheat analysis (Guo *et al*., 2016) was performed. This analysis resulted in an identification of the functional centromere of chromosome 7D.

The generated resources and data were further used in the work focused on positional cloning of the *Dn2401* gene underlying resistance to Russian wheat aphid. Five BAC clones spanning the gene interval (Staňková *et al.*, 2015) were individually sequenced by Illumina technology on the MiSeq platform, assembled, annotated using the TriAnnot pipeline and predicted gene models were manually edited.

Comparison of marker distances in the genetic and physical maps of the studied region indicated that the susceptible cultivar Chinese Spring (CS), from which the sequenced BAC clones originated, could have a larger deletion in the *Dn2401* gene region and consequently, its sequence would not provide complete information on the gene content of the analysed interval. To verify this hypothesis, an optical map of the 7DS arm derived from the line CI2401 carrying the resistance gene was created using the Bionano Genomics platform. Comparison of the 7DS optical maps from the susceptible and the resistant genotype allowed identifying structural variability in the vicinity of selected candidate genes. Comparison of the 7DS CS optical map with obtained BAC clone sequences also revealed a gap in the sequence of one of the BAC clones. Since this part was not properly assembled even in the available whole-genome wheat sequences (IWGSC, 2018; Zimin *et al*., 2017), we proceeded to sequencing the critical BAC clone with Oxford Nanopore Technologies. Some of the reads thus obtained spanned the entire length of the sequenced clone insert. These were combined with the Illumina reads and assembled using MaSuRCA software. The complete and accurate sequence of the region was used to develop new markers to further narrow down the interval and identify candidate genes for the *Dn2401*.

4. Summary of results

4.1 Anchoring and validation of the physical map of the short arm of wheat chromosome 7D

To anchor the 7DS physical map, several approaches were applied, including manual anchoring by PCR on 3D BAC pools or *in silico* anchoring that utilized the generated BAC sequences. Final version of the 7DS physical map (Tulpová *et al*., 2019a) integrated markers from a radiation hybrid map and three genetic maps, including one from the D-genome ancestor, *Ae. tauschii*. Besides, our approach to physical-map assembly included integration of the 7DS physical map with a whole-genome map of *Ae. tauschii* (Luo *et al.*, 2013). This together with involvement of a Bionano genome (BNG) map of the 7DS arm (Staňková *et al.*, 2016) facilitated ordering of physical-map contigs even in the non-recombining region of the genetic centromere.

4.2 Sequencing of the chromosome arm 7DS and assembling of the reference sequence

Within a joint effort coordinated by the IWGSC, 4,608 MTP BAC clones from a 7DS-specific BAC library were sequenced in BAC pools of four non-overlapping clones using Illumina HiSeq platform. Resulting sequences were assembled into contigs with N50 of 72 kb. Their scaffolding was done using mate-pair data obtained from MTP-plate pools (384 clones). The final 7DS BAC assembly, composed of 9,063 scaffolds with N50 of 117 kb (Tulpová *et al*., 2019a), became a valuable data source contributing to the reference sequence of the 7DS chromosome arm and delimiting of its functional centromere (IWGSC, 2018).

4.3 Positional cloning of a Russian wheat aphid resistance gene

The 7DS BAC assembly was also utilized in positional cloning project targeting *Dn2401* gene underlying resistance to RWA, a serious pest of small grain cereals and many grass species. In the previous study of Staňková *et al.* (2015), ~300-kb interval containing the *Dn2401* resistance gene was delimited and five BAC clones spanning this region were selected. Here we used a targeted strategy that combined traditional approaches towards gene cloning, comprising genetic mapping and Illumina sequencing of BAC clones, with novel technologies including optical mapping and long-read nanopore sequencing. Comparison of the obtained BAC hybrid assembly covering the gene region with corresponding parts in two wheat whole genome assemblies (IWGSC, 2018; Zimin *et al*, 2017b) revealed misassemblies in a close proximity of predicted candidate genes. The highly accurate BAC assembly facilitated precise annotation of the *Dn2401* region, saturation of the interval with new markers and proposing and resequencing of candidate genes. Identification of *Epoxide hydrolase 2* as the most likely *Dn2401* candidate opened an avenue to its validation by functional genomics approaches (Tulpová *et al.*, 2019b).

5. Summary

The presented work was focused on the study of the short arm of the wheat chromosome 7D, especially on the generation and exploitation of its reference sequence.

In the first part of the thesis, the physical contig map of the 7DS arm was anchored by various strategies. Using 1713 markers with a known position on three genetic maps or a radiation hybrid map, we were able to anchor contigs of BAC clones covering 73 % of the total length of the arm. In addition, an integrated map, created by combining physical maps of the wheat arm 7DS and whole genome of *Aegilops tauschii*, was employed. The integrated map together with an optical map of the 7DS arm revealed genomic reconstruction in the (peri)centromeric region between wheat and its ancestor. In parallel, BAC clones representing the MTP of the 7DS arm were sequenced, and a sequence consisting of 9063 scaffolds with an N50 length of 117 kb was assembled. This sequence, together with the physical map and the 7DS optical map, contributed to the construction of the reference sequence of the 7DS arm under the auspices of the International Wheat Genome Sequencing Consortium.

The second part of the work was focused on the project of positional cloning of the *Dn2401* gene underlying resistance to Russian wheat aphid. The presented work follows a previous study that delimited an interval, in which the *Dn2401* gene is located, and identified five BAC clones that span this interval. Within the framework of this thesis, the selected BAC clones were individually sequenced by Illumina technology and assembled. Since we were not able to obtain a complete sequence of the interval from the Illumina data, we approached to sequencing one of the clones by a long-read technology. The resulting complete and accurate sequence of the region enabled development of new markers that narrowed down the *Dn2401* gene interval, which contributed to a reduction in the number of candidate genes. After comparing sequence and optical maps of a susceptible cultivar and a resistant line, a gene encoding epoxide hydrolase 2 was selected as the most likely candidate for *Dn2401*.

The presented work provides valuable information for breeding of new wheat varieties with higher resistance to a pest or with other agronomically important traits.

6. References

Burd JD, Burton RL (1992) Characterization of plant damage caused by Russian wheat aphid (Homoptera: Aphididae). *Journal of Economic Entomology* **85**: 2017-2022.

Doležel J, Kubaláková M, Paux E *et al.* (2007) Chromosome-based genomics in cereals. *Chromosome Research* **15**: 51-66.

Doležel J, Čížková J, Šimková H, Bartoš J (2018) One major challenge of sequencing large plant genomes is to know how big they really are. *International Journal of Molecular Science* **19**(11): 3554.

Du Toit F (1987) Resistance in wheat (*Triticum aestivum*) to *Diuraphis no*xia (Homoptera: *Aphididae*). *Cereal Research Communication* **15:** 175-179.

Du Toit F, Wessel WG, Marais GF (1995) The chromosome arm location of the Russian wheat aphid resistance gene, Dn5. *Cereal Research Communication* **23**: 15-17.

Fazel-Najafabadi M, Peng J, Peairs FB *et al.* (2015) Genetic mapping of resistance to *Diuraphis noxia* (Kordjumov) biotype 2 in wheat (*Triticum aestivum* L) accession CI2401. *Euphytica* **203**: 607-614.

Guo X, Su H, Shi Q *et al.* (2016) *De novo* centromere formation and centromeric sequence expansion in wheat and its wide relatives. *PLOS Genetics* **12**(4): e1005997

Harvey TL, Martin TJ (1990) Resistance to Russian wheat aphid, *Diuraphis noxia*, in wheat (*Triticum aestivum*). *Cereal Research Communication* **18**: 127-129.

IWGSC – The International Wheat Genome Sequencing Consortium (2018) Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science*, doi: 10.1126/science.aar7191.

Liu XM, Smith CM, Gill BS, Tolmay V (2001) Microsatellite markers linked to six Russian wheat aphid resistance genes in wheat. *Theoretical and Applied Genetics* **102**: 500-511.

Luo MC, Gu YQ, You FM *et al.* (2013) A 4-gigabase physical map unlocks the structure and evolution of the complex genome of *Aegilops tauschii*, the wheat D-genome progenitor. *PNAS* **110**: 7940-7945.

Nkongolo KK, Quick JS, Limi E, Fowler DB (1991) Sources and inheritance of resistance to Russian wheat aphid in *Triticum* species amphiploids and *Triticum tauschii. Canadian Journal of Plant Science* **71**: 703-708.

Paux E, Legeai F, Guilhot N *et al.* (2008) Physical mapping in large genomes: accelerating anchoring of BAC contigs to genetic maps through in silico analysis. *Functional and Integrative Genomics* **8**: 29-32.

Rimbert H, Darrier B, Navarro J *et al.* (2017) High throughput SNP discovery and genotyping in hexaploid wheat. *PLoS One* **13**(1): e0186329.

Šimková H, Šafář J, Kubaláková M *et al.* (2011) BAC libraries from wheat chromosome 7D: efficient tool for positional cloning of aphid resistance genes. *Journal of Biomedicine and Biotechnology* 2**011**: 302543.

Staňková H, Valárik M, Lapitan NLV *et al.* (2015) Chromosomal genomics facilitates fine mapping of Russian wheat aphid resistance gene. *Theoretical and Applied Genetics* **128**: 1373-1383.

Staňková H, Hastie A, Chan S *et al.* (2016) BioNano genome mapping of individual chromosomes supports physical mapping and sequence assembling in complex plant genomes. *Plant Biotechnology Journal* **14:** 1523 – 1531.

Tiwari VK, Heesacker A, Riera-Lizarazu O *et al.* (2016) A whole-genome radiation hybrid mapping resource of hexaploid wheat. *Plant Journal* **86**(2):195-207.

Tulpová Z, Luo MC, Toegelová H *et al.* (2019a) Integrated physical map of bread wheat chromosome arm 7DS to facilitate gene cloning and comparative studies. *New Biotechnology* 48:12-19.

Tulpová Z, Toegelová H, Lapitan NLV *et al.* (2019b) Accessing a Russian wheat aphid resistance gene in bread wheat by long-read technologies. *Plant Genome*, doi: 10.3835/plantgenome2018.09.0065

Valdez VA, Byrne PF, Lapitan NLV *et al.* (2012) Inheritance and genetic mapping of Russian wheat aphid resistance in Iranian wheat landrace accession PI626580. *Crop Science* **52**: 676-682.

Voothuluru P, Meng J, Khajuria C *et al.* (2006) Categories and inheritance of resistance to Russian wheat aphid Hmoptera: *Aphididade*) biotype 2 in a selection from wheat cereal introduction 2401. *Journal of Economic Entomology* **99**: 1854-1861.

Yazdani M, Baker G, DeGraaf H *et al.* (2017) First detection of Russian wheat aphid *Diuraphis noxia* Kordjumov (Hemiptera: Aphididae) in Australia: a major threat to cereal production. *Austral Entomology*, doi: 10.1111/aen.12292

Zimin AV, Puiu D, Hall R *et al.* (2017) The first near-complete assembly of the hexaploid bread wheat genome, *Triticum aestivum*. *GigaScience* **6**: 1-7.

7. List of author's publication

7.1 Original papers

Tulpová Z, Luo MC, Toegelová H, Visendi P, Hayashi S, Vojta P, Paux E, Kilian A, Abrouk M, Bartoš J, Hajdúch M, Batley J, Edwards D, Doležel J, Šimková H (2019a) Integrated physical map of bread wheat chromosome arm 7DS to facilitate gene cloning and comparative studies. *New Biotechnology* 48:12-19.

Tulpová Z, Toegelová H, Lapitan NLV, Peairs F, Macas J, Novák P, Lukaszewski A, Kopecký D, Mazáčová M, Vrána J, Holušová K, Leroy P, Doležel J, Šimková H (2019b) Accessing a Russian wheat aphid resistance gene in bread wheat by long-read technologies. *Plant Genome*, doi: 10.3835/plantgenome2018.09.0065

IWGSC – The International Wheat Genome Sequencing Consortium (2018) Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* 361: eaar7191.

7.2 Published abstracts of posters

Tulpová Z, Luo MC, Toegelová H, Visendi P, Hayashi S, Vojta P, Hastie A, Kilian A, Tiwari VK, Bartoš J, Batley J, Edwards D, Doležel J, Šimková H: Integrated physical map of bread wheat chromosome arm 7DS to support evolutionary studies and genome sequencing. Plant Biotechnology: Green for Good IV. Olomouc, Czech Republic 2017.

Tulpová Z, Staňková H, Luo MC, Visendi P, Hayashi S, Kilian A, Tiwari VK, Kumar A, Hastie A, Leroy P, Rimbert H, Abrouk M, Bartoš J, Batley J, Edwards D, Doležel J, Šimková H: Completing reference sequence of the wheat chromosome arm 7DS. In: Abstracts of the International Conference "Plant and Animal Genome XXIV". P.0826. Sherago International, Inc., San Diego, 2016.

Tulpová Z, Toegelová H, Vrána J, Hastie AR, Lukaszewski A, Kopecký D, Lapitan NLV, Batley J, Edwards D, The International Wheat Genome Consortium, Doležel J, Šimková H: Structural variation of a wheat chromosome arm revealed by optical mapping. In: Abstracts of the International Conference "Plant and Animal Genome XXV". P.0837. Sherago International, Inc., San Diego, 2017.

Staňková H, Valárik M, Lapitan NLV, Berkmann P, Edwards D, Luo MC, Tulpová Z, Kubaláková M, Stein N, Doležel J, Šimková H: Poziční klonování genu pro rezistenci ke mšici zhoubné (Diuraphis noxia): konstrukce vysokohustotní genetické mapy. In: Sborník abstrakt, Bulletin České společnosti experimentální biologie rostlin, "6. Metodické dny". Seč, Česká republika, 2014.

8. Souhrn

Název práce: Sekvence a funkční analýza krátkého ramene chromozómu 7D pšenice

Předkládaná práce byla zaměřena na studium krátkého ramene chromozómu 7D pšenice, zejména pak na získání a využití jeho referenční sekvence.

V první části práce byla za použití různých strategií ukotvena fyzická kontigová mapa ramene 7DS. Pomocí 1713 markerů se známou pozicí na genetických mapách nebo na mapě radiačních hybridů se podařilo umístit na chromozóm kontigy BAC klonů pokrývající 73% celkové délky ramene. K ukotvování fyzické mapy byla rovněž využita integrovaná mapa, vytvořená kombinací fyzických map ramene 7DS z pšenice a genomu *Ae. tauschii*. Ta společně s optickou mapou odhalila genomové přestavby v oblasti (peri)centromery mezi pšenicí a jejím předchůdcem. Zároveň byly osekvenovány BAC klony představující MTP ramene 7DS a byla sestavena sekvence skládající se z 9063 scaffoldů s parametrem N50 o délce 117 kb. Tato sekvence společně s fyzickou mapou a optickou mapou ramene 7DS významně přispěly k sestavení referenční sekvence ramene 7DS pod záštitou Mezinárodního konsorcia pro sekvenování genomu pšenice.

Druhá část práce byla zaměřena na projekt pozičního klonování genu *Dn2401*, podmiňujícího rezistenci ke mšici zhoubné. Předkládaná práce tak navazuje na předchozí studii, v níž byl vymezen interval, ve kterém se gen *Dn2401* nachází, a současně bylo identifikováno pět BAC klonů, jež tento interval překlenují. V rámci této práce pak byly BAC klony individuálně osekvenovány technologií Illumina. Pro dosažení kompletní a přesné sekvence intervalu bylo nutné osekvenování jednoho z klonů také technologií dlouhých čtení. S využitím získaných sekvencí se podařilo vyvinout nové markery, s jejichž pomocí byl interval genu *Dn2401* zúžen, což přispělo i k redukci počtu kandidátních genů. Po jejich komparativní analýze na úrovni sekvencí a optických map pro citlivý kultivar a rezistentní linii byl jako nejpravděpodobnější kandidát pro *Dn2401* vybrán gen kódující epoxid hydrolázu 2.

Ve svém souhrnu předkládaná práce představuje nový zdroj dat a informací, které mohou být využity pro šlechtění nových odrůd pšenice s vyšší odolností proti hmyzímu škůdci, ale i s dalšími agronomicky cennými vlastnostmi.