Mol Breeding (2014) 34:1531–1548
DOI 10.1007/s11032-014-0179-z

Genotyping by sequencing transcriptomes in an evolutionary pre-breeding durum wheat population

Jacques David · Yan Holtz · Vincent Ranwez · Sylvain Santoni · Gautier Sarah · Morgane Ardisson · Gérard Poux · Fréderic Choulet · Clémence Genthon · Pierre Roumet · Muriel Tavaud-Pirra

Received: 19 July 2014 / Accepted: 8 October 2014 / Published online: 6 November 2014

Abstract The genetic diversity in durum wheat, *Triticum turgidum durum*, has been strongly reduced since the domestication of the wild *Triticum turgidum dicoccoides*. Monitoring durum wheat composite crosses incorporating related tetraploid taxa, such as wild and domesticated emmer wheat, is a suitable evolutionary pre-breeding method. Transcriptome sequencing paves the way for a genomic survey of single nucleotide polymorphisms (SNPs) segregating in such populations, offering the possibility of genotyping by sequencing to use these resources in genome-wide association studies (GWAS) and genomic selection (GS) programs. Evolutionary Pre-breeding pOpulation (EPO) is an evolutionary durum wheat pre-breeding population. Sequencing the transcriptome of 179 durum wheat lines (175 from EPO) led to the detection of 103,262 SNPs on two reference transcriptomes: one from the International Wheat Genome Sequencing Consortium and one assembled de novo on durum wheat. Using strict filtering to remove dubious heterozygous SNPs, EPO genetic diversity was eventually described with 76,188 high-confidence SNPs. The percentage of missing genotyping data depended on the expression level, and 88 individuals out of 175 were genotyped per SNP on average. Using the 3B pseudo-molecule of bread

Electronic supplementary material The online version of this article (doi:10.1007/s11032-014-0179-z) contains supplementary material, which is available to authorized users.

J. David (✉) · Y. Holtz · V. Ranwez · M. Tavaud-Pirra
Montpellier SupAgro, UMR Amélioration Génétique et Adaptation des Plantes, 2 place Viala, 34060 Montpellier, France
e-mail: Jacques.David@supagro.inra.fr

Y. Holtz
e-mail: holtz@supagro.inra.fr

V. Ranwez
e-mail: vincent.ranwez@supagro.inra.fr

M. Tavaud-Pirra
e-mail: Muriel.Tavaud@supagro.inra.fr

S. Santoni · G. Sarah · M. Ardisson · G. Poux · P. Roumet
INRA, UMR Amélioration Génétique et Adaptation des Plantes, 2 place Viala, 34060 Montpellier, France
e-mail: Sylvain.Santoni@supagro.inra.fr

G. Sarah
e-mail: Gautier.sarah@supagro.inra.fr

M. Ardisson
e-mail: Morgane.Ardisson@supagro.inra.fr

G. Poux
e-mail: Gerard.poux@supagro.inra.fr

P. Roumet
e-mail: Pierre.Roumet@supagro.inra.fr

F. Choulet
INRA UMR1095 Genetics, Diversity and Ecophysiology of Cereals, 5 chemin de Beaulieu, 63039 Clermont-Ferrand, France
e-mail: Frederic.Choulet@clermont.inra.fr

wheat, the transcription and diversity levels were shown to be higher in distal regions than in proximal regions, but SNPs were available throughout the chromosomes. Assuming good synteny with Hordeum, the trend was similar on the 14 chromosomes of the durum wheat genome. EPO hosts a high level of diversity, has a number of SNPs in low linkage disequilibrium (<40 Mb) and would be well suited for GWAS and GS programs.

Keywords Durum wheat · Genotyping by sequencing · SNP · RNAseq · Diversity · Dispensable · Repeatability

Introduction

Durum wheat (*Triticum turgidum* subsp. *durum*) is the modern representative of a group of allotetraploid subspecies (*Triticum turgidum* subsp.) that were domesticated from the wild *T. turgidum* subsp. *dicoccoides* (Kilian et al. 2009; Özkan et al. 2011; Luo et al. 2007). It is closely related to bread wheat (*Triticum aestivum* L.) that arose via spontaneous interspecific hybridization between a domesticated *T. turgidum* spp. form (AB genomes, $2n = 4x = 28$) and the wild diploid *Ae. tauschii* (D genome, $2n = 14$) (Caldwell et al. 2004). Due to successive bottlenecks during its history, durum wheat underwent one of the strongest reductions in genetic diversity that has been observed among crops (Haudry et al. 2007; Thuillet et al. 2005). Interesting diversity for breeding durum wheat is thus available in other *T. turgidum* subspecies (Zaharieva et al. 2010; Nevo 2002), and the current durum diversity level could be highly reinforced through the use of wild and ancestral domesticated *T. turgidum* taxa, mostly *T. t.* subsp. *dicoccoides*, *T. t.* subsp. *dicoccum* (hulled forms), *T. t.* subsp. *polonicum* and *T. t.* subsp. *turgidum* (free-threshing forms).

It was proposed several decades ago (Suneson 1956) that composite cross-populations (CCPs) can be an effective way to remobilize diversity by crossing a large number of parents and promoting recombination. Such populations evolve across successive generations and are subject to continuous anthropic and local natural selection and should thus be promoted as evolutionary breeding pools (Enjalbert et al. 2011; Phillips and Wolfe 2005). By extension, CCP management can be seen as an evolutionary pre-breeding method for introgressing wild and exotic germplasm to create resources with a very broad genetic base. Such massive exotic introgression in the current elite background induces long-range linkage disequilibrium (LD). In selfing species such as barley or wheat, monitored male sterility can effectively promote chromatid exchanges and reduce LD (Suneson 1956). Composite crosses managed in the long term are thus very close to the Multi-parent advanced generation inter-cross (MAGIC) population concept recently introduced as an effective way to produce lines adapted to genome-wide association studies (GWAS) or genomic selection programs (Mackay and Powell 2007). With recent advances in sequencing technologies, exploiting and monitoring CCPs that include wild and/or exotic accessions have become easier and more efficient. With broad diversity, a sufficient recombination to reduce the long-distance linkage disequilibrium hence the number of spurious genotype x phenotype associations, CCPs could provide material fully tailored for GWAS and genomic selection programs. Genome-wide surveys of their polymorphism will thus help define their relative advantages compared to usual panels based on collections of varying spatiotemporal origins.

Durum wheat is a minor crop compared to bread wheat. It has benefitted from the development of knowledge on bread wheat genomes. Large SNP sets have been produced on bread wheat and to a lesser extent on durum wheat (Cavanagh et al. 2013; Paux et al. 2010; Pont et al. 2013; Wang et al. 2014), which have been used to develop high throughput genotyping arrays, based on various technologies (Cavanagh et al. 2013; Ganal et al. 2011; Wang et al. 2014). Such assays are high throughput, cost-effective and do not require long bioinformatic computations for genotype calling. Genetic maps are now available from these SNPs in bread wheat (Cavanagh et al. 2013) and durum wheat as well (see this volume).

Such arrays are very efficient, but can only reveal already detected polymorphisms suited to the chosen array technology. Although adapted for studying

C. Genthon
MGX-Montpellier GenomiX, c/o Institut de Génomique Fonctionnelle, Montpellier, France
e-mail: clemence.genthon@mgx.cnrs.fr

modern panels of elite bread and durum wheats, their use is questionable in pre-breeding durum wheat CCPs involving wild and *T. turgidum* subsp. of broad origin since assays could overlook a large part of the population diversity. Furthermore, markers from high-density arrays may be prone to ascertainment bias when developed for SNPs discovered on reference samples of limited size and focused on an elite pool (Albrechtsen et al. 2010, Moragues et al. 2010). Diversity measurements can be distorted and alter the population genetic inferences, especially with rare variants (Bhatia et al. 2013).

Sequencing is a good alternative but on wheat, a genome reduction is still needed. Genotyping by sequencing (GBS) is another potential alternative to arrays and captures for pre-breeding populations and can be used for de novo genotyping, even for the large, complex and polyploid wheat genome (Poland et al. 2012a). The principle is to sequence a reduced portion of the genome. GBS does not require a dedicated assay and can be launched with a limited amount of previous genomic information. Sequences contain much more precise information than SNPs detected by fluorescent signals on arrays and could be used to disentangle complex hybridization situations, particularly in polyploidy species. Genomic GBS is currently developing rapidly in plant breeding (Liu et al. 2014, Poland and Rife 2012) and in evolutionary ecology (Davey et al. 2011). In diploid species, satisfactory results have been obtained for the discovery and use of numerous polymorphisms, but in complex polyploidy genomes with large amounts of repetitive sequences such as wheat, identifying true contigs requires complex bioinformatic analysis steps, and the percentage of informative reads are quite low due to the high concentration of nearly identical repeated sequences (Poland et al. 2012b, Liu et al. 2014). The method generally used to reduce the complexity of target genomes is restriction site-associated DNA (RAD) sequencing (Baird et al. 2008; Poland et al. 2012b). As some mutation can appear in the recognition sequence of the digestion enzyme, RAD sequencing may also lead to non-random missing data, leading to ascertainment bias and may question their use in evolutionary studies (Arnold et al. 2013). Targeting coding sequence is also a good alternative to reduce the genome complexity. Capturing genomic sequences, using baits based on 3,497 cDNAs (3.5 Mb) of *Triticeae*, has been used on durum wheat (Saintenac et al. 2011b). Such dedicated capture assay is a powerful approach for variant SNP discovery and also gene copy number variation (CNV) within cultivated and wild wheat genotypes in a small but informative part of the genome. Sequencing transcribed portions of the genome from RNA extracted from standardized tissues (RNAseq) is also a good alternative genomic reduction method since the transcribed gene-coding regions represent only one to two percent of the whole genome (Paux et al. 2006). This approach (RNAseq GBS) has been successfully used in comparative studies involving a few individuals per species to detect several thousands of polymorphisms (Cahais et al. 2012; Gayral et al. 2013; Nabholz et al. 2014). De novo assembly and mapping of RNAseq reads are far easier than genomic counterpart GBS and can deliver several thousands of SNPs within gene exons and UTRs, which presumably represent a large fraction of the functional content of a genome. The variation of expression among genes, individuals and tissues is of primary interest for functional and evolutionary studies (Renaut et al. 2010), and the repeatability of the expression profile is suited for functional analysis (Marioni et al. 2008). Unfortunately, transcriptomic libraries are more costly than genomic libraries, and constitutionally, lower expressed genes as well as those whose expression is very dependent on micro-environmental variations are likely to generate a significant portion of missing data as compared to high throughput genotyping assays or GBS on genomic DNA. For each RNAseq GBS project, a balance therefore must be found between the acceptable level of missing data and the sequencing/financial effort needed for each individual.

Substantial methodological development is still required regarding the use of GBS on RNAseq, especially for genes with multiple copies or alternative splicing forms (Krasileva et al. 2013). In polyploid species, homeology can also lead to erroneous assemblies and mapping of reads from two homeologous genes on a single reference, hence leading to an excess of heterozygous genotypes (Trebbi et al. 2011, Krasileva et al. 2013). A high level of gene paralogy has a similar effect. The choice of a reference transcriptome to map the RNAseq reads is thus crucial. A de novo assembled transcriptome of a highly covered individual from the studied durum wheat sample would provide a reference tailored to the durum genomic data but would be challenged to accurately attribute contigs

to the A or B genomes or account for possible pan-transcriptomes (Hirsch et al. 2014) or pan-genomes (Morgante et al. 2007; Tettelin et al. 2005). The question of assembling homeolog genomes of tetraploid species has been addressed in wheat. Several tools based on the polymorphism phasing (Krasileva et al. 2013) or expression bias between homeologous copies (Ranwez et al. 2013) have been proposed. An alternative reference could be the coding sequences predicted from contigs assembled in the chromosome sequencing survey (CSS) launched by the International Wheat Genome Sequencing Consortium (IWGSC, http://www.wheatgenome.org/Projects/IWGSC-Bread-Wheat-Projects/Sequencing/Whole-Chromosome-Survey-Sequencing) and released recently (International Wheat Genome Sequencing 2014). The CSS was built upon genomic sequences from sorted chromosome arms providing a priori correctly separated assemblies of the different homeologous genomes. Mapping on the CSS could ensure better discrimination between A and B homeologous copies than mapping on a de novo durum reference and possibly between paralogs whose UTRs have accumulated sufficient divergence between copies. The complete pseudo-molecule of chromosome 3B, i.e., the first wheat chromosome assembled as a pseudo-molecule (Choulet et al. 2014), is also an important reference for precise localization of contigs. This could pave the way for in-depth studies of the impact of recombination on gene density and expression, polymorphism distribution and linkage disequilibrium decay.

In wheat and related species such as barley, gene density and polymorphism patterns are correlated with local recombination rates (International Barley Genome Sequencing et al. 2012; Choulet et al. 2010). For instance, in wheat, proximal chromosome regions have much lower recombination rates than distal regions (Saintenac et al. 2011a; Dvorak and Chen 1984). Gene density is higher in telomeric regions than in centromeric regions (Choulet et al. 2010). Polymorphism also tends to be reduced in non-recombining regions (International Barley Genome Sequencing et al. 2012; Dvorak et al. 1998), which suggests an impact of the recombination rate on the selection efficiency (Begun and Aquadro 1991; Comeron et al. 2008; Williford and Comeron 2010). RNAseq GBS could thus be biased toward telomeric vs centromeric regions.

Here, we report on an attempt to use RNAseq GBS on a sample of 189 accessions of *T. turgidum*, most of which are from an Evolutionary Pre-breeding pOpulation (EPO). EPO genetic diversity was initially broadened with wide crosses, and the population was then continuously grown over 17 generations under a theoretical 10 % outbreeding rate (Tavaud et al. in prep). We mapped reads on the IWGSC CSS, the 3B reference and a complementary de novo assembly on durum wheat. Repeatability, i.e., the robustness of genotype calling, is also a key determinant for choosing a genotyping method. We focused specifically on documenting the repeatability of SNP calling.

In this paper, we report 103,262 highly filtered SNPs and investigate the capacity of transcriptomic GBS to call repeatable genotypes. We describe and comment on the impact of gene expression and sequencing efforts on the quantity of missing data. Finally, we compare our detected SNPs to RNA-seq SNPs recently published on bread and durum wheat (Wang et al. 2014) in order to assess shared, and specific, polymorphism between those two panels.

The chromosome polymorphism distribution, minor allele frequencies, genetic structuring and linkage disequilibrium decay in EPO were checked to estimate the potential of EPO as a broaden germplasm resource in GWAS or genomic prediction studies. We hope that these resources will be helpful for the development of durum wheat genomic resources and in further GWAS in pre-bred sets of durum lines.

Materials

Panels

EPO: evolutionary pre-breeding population

The Evolutionary Pre-breeding pOpulation (EPO) was founded using a composite cross with a broadened genetic base in which a male sterility recessive nuclear gene segregates. In 1997, diversity from wild and primitive collection accessions of *T. turgidum* subsp. was introduced in this population: Blooming spikes of collection accessions were cut and used as bulked pollinators of male sterile plants previously identified in situ in the population. The seed bulk yielded on these female plants was sown to constitute the founding EPO generation.

EPO has since been grown every year, with a targeted 10 % outcrossing. To ensure this outcrossing, male sterile plants (females) are tagged at blooming time. The next generation consists of 20 % seeds from harvesting whole male sterile plants, 70 % from hermaphrodite plants, selected visually for agronomic performance, and 10 % from harvesting some highly valuable F5-F6 lines derived from a base broadening breeding program aimed at tapping the high diversity from wild and primitive crosses. All of these seeds are bulked and sown using a pneumatic sower to obtain 4,000 individualized plant plots. During the vegetative phase, the tallest and weakest (yellowish) plants are eliminated.

In 2009, 175 fertile lines were extracted from the 17th EPO generations and underwent two successive generations of selfing by single seed descent.

Other panels

To roughly estimate the level of diversity available in EPO compared to other sources (wild, exotic and extreme elite), 2 accessions of *Triticum turgidum* ssp *dicoccoides* (DD428084, DD46310), 2 *T.t.* ssp *dicoccum* (*Dic2,* DC45399), 7 elite durum varieties from the French catalog (Néodur, Ixos, Lloyd, Primadur, Pescadou, Soldur, Silur) and 3 advanced breeding lines (05ETTE3IN1, TT04DD79_27, DD79_37) from the base broadening program mentioned earlier were added to the study. These broadening lines come from crosses with *T. t.* sp. *polonicum* and *T. t.* sp. *dicoccoides* followed by one backcross on durum elite genitors (Annex Genealogy). Their agronomical level was sufficient to start the process of variety registration in the French catalog (decision pending). They have shown a particularly high level of resistance to leaf rust (unpublished data).

Methods

Molecular protocol, sequencing and read cleaning

For each accession, seeds were harvested and germinated in petri dishes with 4 ml of purified water at a constant temperature of 30 °C in the dark in a growth chamber. Coleoptiles and primary leaves were sampled from 7-day-old seedlings and crushed in liquid nitrogen. We obtained sequence data following three main steps: (i) mRNA extraction and purification, (ii) library construction and (iii) sequencing using the Illumina mRNA-Seq, paired-end indexed protocol (see molecular protocol in annexes).

Each cDNA library was sequenced on a HiSeq2000 or a HiSeq2500 by multiplex of 24 per lane (except for the first 48, which were multiplexed in a single lane). Forty-eight samples (EPO_049 to EPO_096) were sequenced three times as a problem was encountered in the first two runs, resulting in single-end sequencing instead of paired-end sequencing. Thirty-eight libraries were resequenced as their initial read quantity was too low. Reads from these problematic runs were, however, merged with those of the final correct sequencing run.

Reads were preprocessed with cutadapt (Martin 2011) to remove adaptor sequences and trim the end of reads with low-quality scores (parameter -q 20) while keeping reads with a minimum length of 35 bp. We then filtered the reads on the basis of their mean quality score, keeping those with a mean quality higher than 30. Subsequently, we removed orphan reads (i.e., those for which the mate was discarded in the previous filtering steps) using a homemade script.

Bread wheat references and durum de novo assembly

We mapped our reads on two different references. The first one was derived from the bread wheat chromosome survey sequence for the cv. Chinese Spring (http://plants.ensembl.org/triticum_aestivum) generated by the International Wheat Genome Sequencing Consortium (IWGSC). We relied on the biomart facilities of Ensembl to collect the unique reference transcript (CDS + UTRs) and the physical genomic location of each of the 66,307 genes predicted from the IWGSC on genome A and B (Ensembl release 22, http://plants.ensembl.org/biomart/martview/, http://plants.ensembl.org/Triticum_aestivum/Info/Annotation/#genebuild). As these sequences were obtained by separately sequencing each bread wheat chromosome arm, this bread wheat reference (BWr) helped us to distinguish homeologous durum wheat copies and to assign a physical position to our predicted durum wheat SNPs.

We also used a durum wheat de novo reference (DWr), which was assembled using the strategy of (Ranwez et al. 2013), which is briefly summarized hereafter. First, reads of 106 sequenced EPO individuals

were assembled using abyss + CAP3 (Huang 1999, Simpson et al. 2009), as advised in (Cahais et al. 2012). The resulting 106 de novo individual assemblies were then assembled using CAP3 (>125 bp overlap with ≥95 % similarity) to limit chimeric contig creations. Reads were mapped on this draft durum wheat transcriptome and Homeosplitter software (Ranwez et al. 2013) was used to unravel the homeologous copies erroneously merged in a single chimeric contig. The resulting durum wheat de novo reference (DWr) allowed us to identify SNPs on contigs specific to the durum wheat transcriptome.

Mapping and SNP discovery

For each library, reads were mapped once on BWr and DWr transcriptomes using BWA (Li and Durbin 2009) while allowing 3 errors (-n 3 in the aln step). We then used Picard tools (http://picard.sourceforge.net) to remove PCR and optical duplicates. Reads per kilobase per million (RPKM (Mortazavi et al. 2008)) were computed at all sites for each individual using a custom Perl script to parse the mapping files. RPKM was used as a gene expression proxy.

Genotype calling was carried out using read2SNP (Gayral et al. 2013). Read2SNP considers the number of reads displaying each nucleotide at a given position and returns the most probable genotype given these nucleotide counts and a *Fis* parameter characterizing the population mating system. *Fis* measures the heterozygous deficit compared to theoretical panmixia (Wright 1950). Under complete selfing, *Fis* has to be set at 1, while under panmixia the *Fis* value is 0. In our case, *Fis* was set at 0.8 based on previous empirical estimates (data not shown). This *Fis* estimation is in line with what is expected for our durum EPO lines extracted from a mixed mating durum population and two successive selfings. Genotypes inferred with a coverage of less than 10 reads, a read2SNP probability of below 99 % or having an observed Fis value below 0.8, were discarded and considered as missing data. Furthermore, a polymorphism was considered reliable only if it had exactly two alleles and each of them was found to be homozygous in at least one genotype. Genotype and SNP callings were done separately for each population: EPO, elite, broadened line (BR), *dicoccum* (DC), *dicoccoides* (DD) and their union (all) on the two possible references (BWr and DWr), leading to 10 overlapping SNP sets, denoted REF-POP-SNP (e.g., BWr-EPO-SNP, DWr-DD-SNP). BWr and DWr are two overlapping sets of contigs, and for each POP, DWr-POP-spSNP denotes the set of SNPs called on DWr-specific contigs. The union of all our SNPs for a given POP was obtained by taking the union of BWr-POP-SNP and DWr-POP-spSNP, which was denoted as POP-SNP.

Genotype-calling repeatability

We tested the repeatability of genotype calling with respect to (i) plant mRNA extraction, (ii) library construction and (iii) sequencing. These tests are described below, along with the number (n) of repetitions per test. First, we tested the genotype-calling repeatability when using RNA extractions from two plants of the same accession in the same selfing generation ($n = 2$) and also between two plants from two successive selfing generations with the same accession ($n = 1$). Second, we tested the genotype-calling repeatability when using (three) different libraries for the same RNA samples ($n = 4$). Third, we tested the genotype-calling repeatability when using (three) different sequencing runs on the same library ($n = 1$).

Read mapping, SNP detection and genotype calling were done separately in these 24 Illumina runs. The percentage of genotype mismatch (excluding missing data) was calculated to assess the genotype-calling repeatability.

Contigs and SNP annotation

The IWGSC provides a pseudo-molecule of the 3B chromosome (Choulet et al. 2014). We used *Hordeum* (Ensembl database) for the other chromosomes. This provided us with the physical localization of the 3B contigs of the BWr, and with the *Hordeum* syntenic position of other BWr contigs. The homology search was done using BLAST, requiring an overlap of at least 150 bp with a similarity higher than 95 %. Blasting our de novo DWr contigs against the bread wheat genome allowed us to identify and locate those coming from the 3B chromosome and to estimate the relevance of using DWr instead of BWr for identifying specific durum SNPs. The BWr open reading frames were downloaded from Ensembl, except for those of the 3B chromosome that where positioned on the 3B pseudo-molecule (Choulet et al. 2014). DWr contigs

were annotated using prot4EST (Wasmuth and Blaxter 2004) and blasted against the Hordeum genome (release IBSC_1.0 of the International Barley genome Sequencing Consortium available through Ensembl, release 22).

SNP validation

In the absence of a mapping population to validate the SNPs, the flanking sequences of the 90 K SNPs recently published in the bread and durum wheat coding sequences (Wang et al. 2014) were compared to our EPO_SNPs. We characterized each SNP by a 101-bp sequence centered on the polymorphic sites (encoded with the IUPAC code) and including 50 flanking nucleotides on both sides. We excluded SNPs having a SNP neighbor in this 101-nucleotide vicinity as well as those located on contig extremities for which the flanking regions were not available to focus on those unambiguously characterized by a single 101-bp sequence. Exact matches between the characterizing sequences of SNPs of Wang et al. and ours led to identification of shared polymorphisms. We also sought the exact flanking regions of these 90 K SNPs in our whole DWr in order to differentiate cases in which our DWr did not include the corresponding locus from cases where our DWr included this locus but no SNP.

Diversity and genomic context

Genomic environment impact was investigated by computing the density of expressed genes, the expression level and nucleotide diversities π (Tajima 1983) on 40-Mb physical windows along chromosomes. Synonymous (π_S) and non-synonymous (π_N) nucleotide values were computed with custom Perl scripts (Nabholz et al. 2014).

Intra- and inter-population structures and linkage disequilibrium

Pairwise genetic distances among accessions were computed as the percentage of different genotypes observed at polymorphic sites in All-SNP (i.e., using Sokal–Michener simple matching). This was done using a custom Java program that ignores missing data. The resulting distance matrix was a 2D plot using multiple dimensional scaling (MDS) projection implemented by the *cmdscale* function of R (v 3.1.0).

Linkage disequilibrium (LD) decay analysis was conducted using the subset of 3B chromosome SNPs of the EPO population genotyped for at least 70 accessions. As the EPO population may be structured, r^2 values (measuring LD among two SNPs) were corrected using the LDcorSV R package (Mangin et al. 2012).

Results

Individual read coverage

An approximate total of 3 billion read pairs were produced with a number of reads per accession ranging from 4.5 to 51 million. As reads have an average length of 100 bp, this study hence relies on more than 600 Gb of sequenced data.

We obtained a mean of 6,301,223 clusters for each sample sequenced in a 24 multiplex lane, with a minimum, maximum and standard deviation of 4.1 million, 11.8 million and 1.3 million, respectively. The cleaning steps resulted in 5.5 million of read pairs per sample on average ($\min = 3.4 \times 10^6$, $\max = 10.1 \times 10^6$ and $\text{sd} = 1.1 \times 10^6$), with an average of 392,913 orphan reads ($\min = 179{,}601$, $\max = 846{,}175$ and $\text{sd} = 124{,}223$). This resulted in an average of 11.5 million usable reads per accession ($\min = 7.2 \times 10^6$, $\max = 20.5 \times 10^6$, $\text{sd} = 2.2 \times 10^6$).

De novo durum wheat assembly as compared to bread wheat transcripts

The final de novo DWr assembly of our reads consisted of 80,691 contigs, whereas BWr consisted of only 66,307 reference transcripts. Apart from the durum wheat genomic specificity, this could also be due to the fragmentation of some transcripts into several contigs (due to partial read coverage of some transcripts) and to alternative splicing (as BWr contains a single transcript per gene). About 1/3 of BWr transcripts (22,746/66,307) did not have homologs among DWr. They could correspond to specific bread wheat transcripts or to genes whose expression level in young plantlets was insufficient to be correctly assembled in DWr. Conversely, approximately 1/3 of

Table 1 Number of SNPs, with or without Fis filtering, detected by mapping on DWr and BWr on each panel. Specific SNPs correspond to SNPs filtered (Fis > 0.8) not shared with any other population

		EPO ($n = 175$)	Dicoccoides ($n = 2$)	Dicoccum ($n = 2$)	Durum ($n = 7$)	Broaden ($n = 3$)	All ($n = 189$)
DWR	SNP called	124,817	27,232	13,912	19,479	9,513	178,721
	with *Fis* > 0.8	95,036	27,232	13,912	19,479	9,513	141,642
BWr	SNP called	84,710	20,938	12,075	19,939	9,440	117,570
	with *Fis* > 0.8	57,659	20,938	12,075	19,939	9,440	84,410
BWr + *DWr_sp*	SNP called	108,777	23,359	13,224	21,715	10,307	151,064
	with *Fis* > 0.8	76,188	23,359	13,224	19,004	10,042	103,262
	specific SNPs	53,265	14,791	5,087	2,670	1,671	

DWr contigs (25,650/80,691) did not have a homolog in BWr, so SNPs found on these contigs were thus specific to durum wheat (DWr_sp-SNP). Only 1/3 of those DWr-specific contigs (8,201/25,650) were found by prot4EST to have a significant homology with a gene present in the three multi-species databases (Swissprot, Trembl and Genbank NR). Out of the 25,650 specific DWr contigs, only 1,247 (3,922, respectively) had a significant blast of 200-bp alignment with a 90 % similarity (100 bp and 80 %, respectively) on the MIPS repeat database for *Poaceae* (Nussbaumer et al. 2013).

SNP discovery and filtering

SNP calling was done based on different durum read subsets (corresponding to durum populations) mapped on either BWr or DWr and with or without applying post-filtering based on *Fis* in addition to our other strict filtering (see Methods section for details). Table 1 summarizes all of these results. Before (resp. after) *Fis* filtering, mapping EPO reads (175 lines) on BWr led to 84,710 (resp. 57659) SNPs on 15,317 distinct transcripts. Before (resp. after) Fis filtering, mapping EPO reads on DWr led to 124,817 (resp. 95,036) SNPs. The loss of SNP induced by *Fis* filtering was hence significantly greater (X^2 test, p value $<2.2e^{-16}$) with BWr $\sim$ 1/3 (1−57,659/84,710) than with DWr $\sim$ 1/4 (1−95,036/124,817).

About 1/3 (32,589/108,777) of SNP predicted on BWr + DWr_sp had *Fis* < 0.8 (Table 1 and Figure S1) and were discarded. The final EPO-SNP set contains 76,188 high-confidence SNPs. Using the same BWr + DWr_sp reference and *Fis* filtering, 23,359 SNPs were found in the two *dicoccoides*, 13,224 in the two *dicoccum*, 19,004 in the 7 durum lines, and 10,042 in the three broadened lines. The total number of high-confidence durum SNPs unraveled using reads from the entire population (All-SNP) was 103,262.

Genotyping repeatability

The repeatability was extremely good on the dedicated 24 library experiments. The genotypes were called slightly differently in the repeated libraries: 0.49 % of differences within the same library, 0.48 % between libraries of a single RNA extraction and 0.36 % between RNA extractions of two independent seeds of the same accession at the same selfing generation. The differences were somewhat sensitive to *Fis* SNP filtering: Using a less stringent threshold decreased the repeatability, e.g., for within library comparisons, we observed 1.47 % of genotype differences when using a threshold Fis of 0.6 versus 0.49 % with a Fis of 0.8.

SNP comparison and validation

Having sequenced a 2010 plant (using destructive RNA extraction) and the offspring of a sister seed of this plant after an additional selfing, we were able to confirm that the observed heterozygosity decay was consistent with the expected value of 1/2. Using *Fis* $\geq$ 0.8 filtering led to $\sim$0.6 % of observed heterozygosity in the 2010 plant (59 heterozygous genotypes among 8,765 SNPs for

about 2 M reads) versus 0.2 % (28/13,685 for about 9 M reads) in the 2011 plant. This heterozygosity decay (from 0.6 to 0.2 %) was consistent with the expected reduction of 1/2 in the observed heterozygosity after an additional selfing.

Out of our 76,188 high-confidence EPO-SNPs, 41,476 SNPs were unambiguously characterized by an associated 101-nucleotide sequence, from one of the 87,964 SNPs included on the recently released 90 K array (Wang et al. 2014). A fraction of 9.1 % (3,797/41,476) of our characterized SNPs was present in the 90 K array built on discovery panels. Hence, about 90 % of our SNPs were absent from the 90 K microarray for which hexaploid accessions were much more represented that tetraploid ones in the panel. Although we cannot exclude that the exact match was too stringent and that this percentage of novel SNPs was probably slightly overestimated, our EPO-SNPs seem to be a highly valuable resource that could be used to better track durum diversity in future genomic studies.

Genomic distribution of diversity

As summarized in Table S1, EPO and *T. dicoccoides* had the highest nucleotide diversity in coding sequences (resp. $\pi = 3.36 \times 10^{-3}$ per base and $\pi = 3.31 \times 10^{-3}$), while the French elite durum lines had the lowest ($\pi = 1.95 \times 10^{-3}$). EPO not only have high nucleotide diversity ($\pi = 3.36 \times 10^{-3}$) but also have numerous non-synonymous polymorphisms since the non-synonymous nucleotide diversity value, $\pi_{N,}$ equals 1.28×10^{-3}. Although π values are much more robust to sample sizes than raw SNP numbers, they should be cautiously considered for all populations except EPO since other population sample sizes are very small (only 2 for *dicoccoides* and *dicoccum*, which is the strict minimum to compute a diversity index).

Table S2 summarizes the distribution of observed diversity along the chromosomes of bread wheat genomes A and B. Considering EPO-SNPs, no real trend appeared in a global comparison of diversity in the A and B genomes. As no chromosome had less than 2,500 SNPs, a more in-depth analysis at the chromosome level would be possible. The findings of this analysis highlighted major differences among the nucleotide diversities noted for the different chromosomes. Indeed, chromosomes 6A, 3B and 5B had twofold π values ($\pi > 4.4 \times 10^{-3}$) compared to chromosomes 3A and 5A ($\pi \leq 2.17 \times 10^{-3}$).

Going a step further, intra-chromosomal variations along the pseudo-molecule of the 3B chromosome could also be studied. As illustrated in Fig. 1, the density of BWr transcripts varied along chromosome 3B, with much more transcripts in the telomeric region than in the centromeric region. The SNP density per kb of transcribed sequence followed the same trend (Fig. 1). These two effects reinforced each other, providing many more SNPs in telomeric regions.

Figure 1 plots the variation in π along chromosome 3B for the various populations. As shown in this figure, the difference between two populations tended to be almost constant along 3B chromosomes. We split the 3B chromosome in three regions according to the recombination rate proposed by (Choulet et al. 2014): two highly recombining distal regions R1 (0–68 Mb, short arm) and R3 (715–774 Mb, long arm) and a low recombining proximal region R2 (68–715 Mb). Among SNPs detected on BWr and genotypes for at least 70 accessions, R1, R2 and R3 had 561, 836 and 289 SNPs, respectively (Table S3). This means that the lowly recombining region R2, much less rich in SNPs, might still be covered for the identification of genotype x phenotype associations.

As illustrated in Figure S2, the same three trends (more transcripts and SNPs in telomeric regions, but a similar gap in population diversity along the 3B chromosome) were also observed on other chromosomes (studied through barley synteny). This could be considered with caution for the 4A, 5A and 7B chromosomes which are implied in ancestral translocations (Devos et al. 1995, Hernandez et al. 2012) that are not shared with barley. Incorrect physical locations might have occurred on these chromosomes.

Thanks to these dense durum-specific EPO-SNPs, it should thus be possible to find SNPs to map durum QTLs, even in low recombining centromeric regions using appropriate material (GWAS or highly recombining populations).

Genotyping by sequencing transcriptomes on the EPO

On an average, 88.6 genotypes out of 175 were called per SNP. The initial numbers of reads per accession ranged from 4.5 to 51 million reads, and the number of called genotypes per individual (on EPO polymorphic

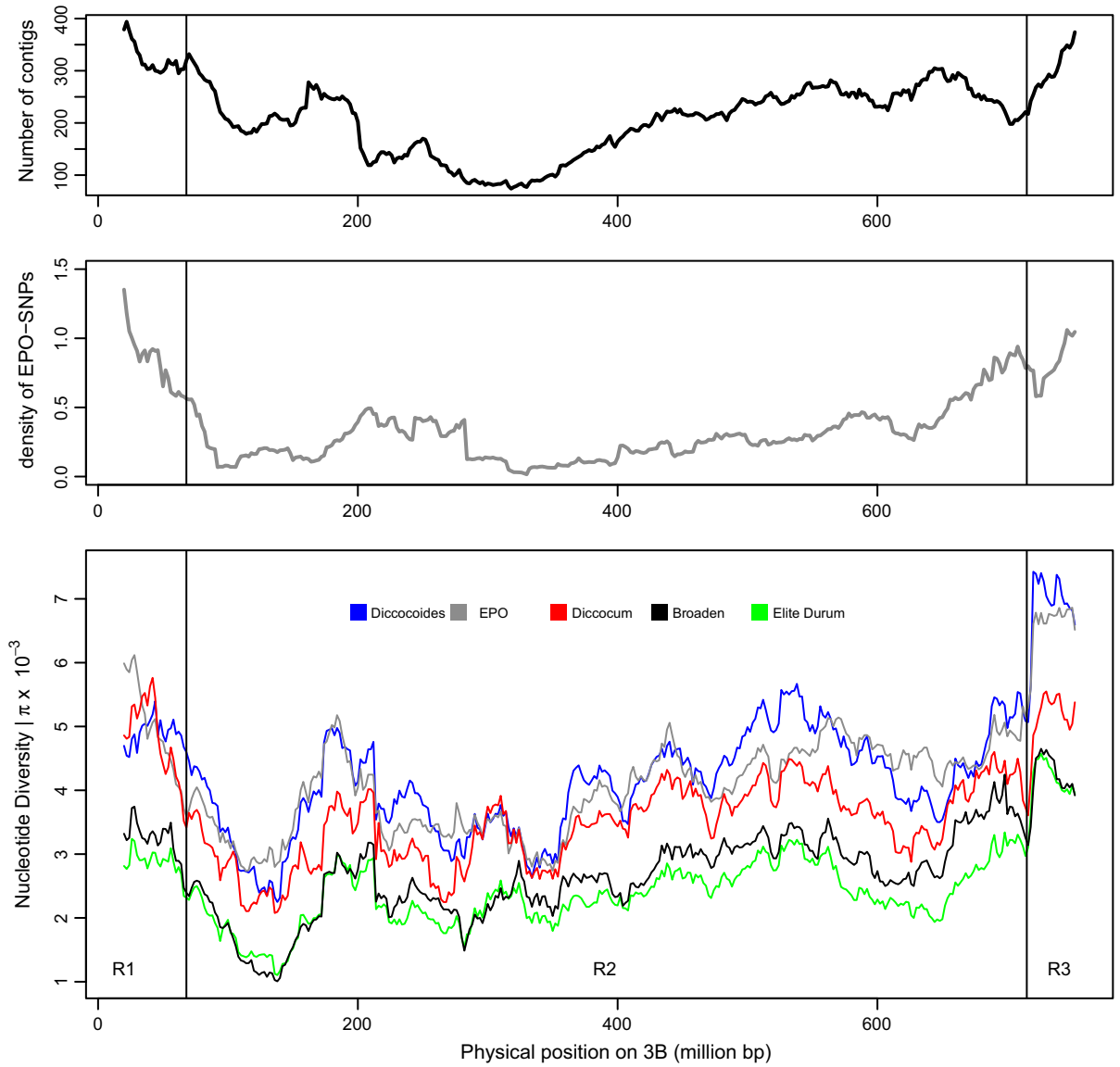

Fig. 1 SNP density and nucleotide diversity along the 3B chromosome. Those three plots represent the distribution of contigs (at the *top*), EPO-SNP (*middle*) and nucleotide diversity (*bottom*) along the physical map of the 3B chromosome using sliding windows of 40 Mb with a step of 2 Mb. The *vertical bars* represent the limits of telomeric/centromeric regions proposed by (Choulet et al. 2014). The nucleotide diversity of *each panel* is represented on the bottom plot using *different colors*. (Color figure online)

sites) ranged from 13,543 to 73,264. As illustrated in Fig. 2, there was a strong correlation between those two values and no real plateaus were noted, even for the most covered accessions.

As illustrated in Fig. 3, the contig expression was a good predictor of the percentage of missing SNP data. Indeed, 99 % of the SNPs observed on the 13,330 contigs with RPKM < 10 had more than 20 % missing data for both BWR-EPO-SNPs and DWR-sp-EPO-SNPs. This percentage fell to 57 % (resp. 68 %) for BWR-EPO-SNPs (resp. DWR-sp-EPO-SNPs) found in contigs with RPKM > 10. Interestingly, the vast majority of DWr-sp contigs (5,730 among 6,334) had a low expression level, with RPKM < 10 (Fig. 3). By contrast, less than half of the BWr contigs had RPKM < 10 (7,600 among 15,317).

The usefulness of EPO-SNPs in GBS for GWAS depends on the distribution of missing data per SNP

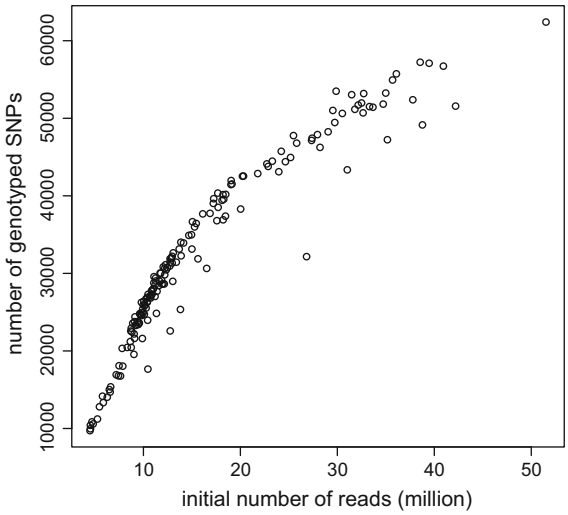

Fig. 2 Influence of sequencing depth on genotype calling. *Each dot* represents an accession positioned according to its number of sequenced reads (*X* axis) and its number of called genotyped (*Y* axis)

and on the minor allele frequencies (MAF) spectrum of SNPs. Figure S3 displays the number of EPO-SNPs available for at least n accessions (i.e., those suitable for GWAS requiring less than a given percentage of missing data per SNP). Among the 76,188 EPO-SNPs, 26,542 (35 %) were genotyped for at least 100 accessions, 29,055 (38 %) for at least 90 accessions and 41,834 (56 %) for 50 accessions. Due to the difference in expression between DWr-sp contigs and BWr contigs, very few DWr-sp were genotyped simultaneously for a large number of individuals. BWr-EPO-SNPs and DWr-sp-EPO-SNPs had a different MAF spectrum (Figure S4), but both sets had a large fraction (59.57 and 59.21 %, respectively) of SNPs with MAF > 0.2 that would be well suited for a balanced analysis of allelic effects on any phenotype. This percentage of $\sim$60 % of SNPs with MAF > 0.2 is lower than the 72 % found in previously studied durum wheat SNP discoveries (Trebbi et al. 2011). This highlights the richness of rare uncommon alleles in EPOs that could be interesting for breeding.

Population structure

The MDS plot representing EPO accessions according to their genetic distances showed that the accessions were mainly regularly dispersed on the first coordinate (X axis revealing the genetic proximity between these accessions), while some outliers are scattered on the second coordinate (Y axis, Figure S5). Small groups of highly related accessions were also revealed by this 2D representation. These groups probably corresponded to sister lines coming from the same ancestral line (either some ancestral successful lines or some immigrating lines from the base broadening program).

While durum wheat lines shared 78 % of their SNPs with EPO, *dicoccoides* accessions shared only 29 % of their polymorphic sites with EPO (*dicoccum* 47 % and the broadened lines 75 %) and had the highest proportion of specific alleles as compared to the other panels (Table 1). This specificity of *dicoccoides* and *dicoccum* accessions clearly appeared in the multiple dimensional scaling analysis that included all our durum populations (Fig. 4) in which the seven durum elite lines are grouped with the 175 EPO accessions spread on the right. Broadened lines are close to modern durum lines and have little of EPO diversity.

Linkage disequilibrium analysis

A subset of 721 SNPs of the BWr-EPO-SNP set fulfilled the conditions we required for computing kinship corrected LD (r^2) along the 3B chromosome. The observed LD was generally consistent with the predicted physical distance on the 3B pseudo-molecule (Figure S6). On average, LD rapidly decreased after 10–20 Mb and reached the plateau of long-distance LD equilibrium after 40 Mb. Our dataset was not sufficient to test whether this decay was faster in the distal region than in the proximal region. However, a few contigs were involved in medium to long-distance LD. Although the possibility of a hidden structure or strong epistatic selection cannot be completely excluded, this most likely reflected improper mapping of these contigs in the telomeric region of the 3B long arm (Figure S7).

Discussion

Transcriptomic GBS is useful for detecting valuable SNPs

Under our standardized seedling growth and tissue sampling conditions, an Illumina lane of 160 million read pairs used for sequencing 24 accessions obtained

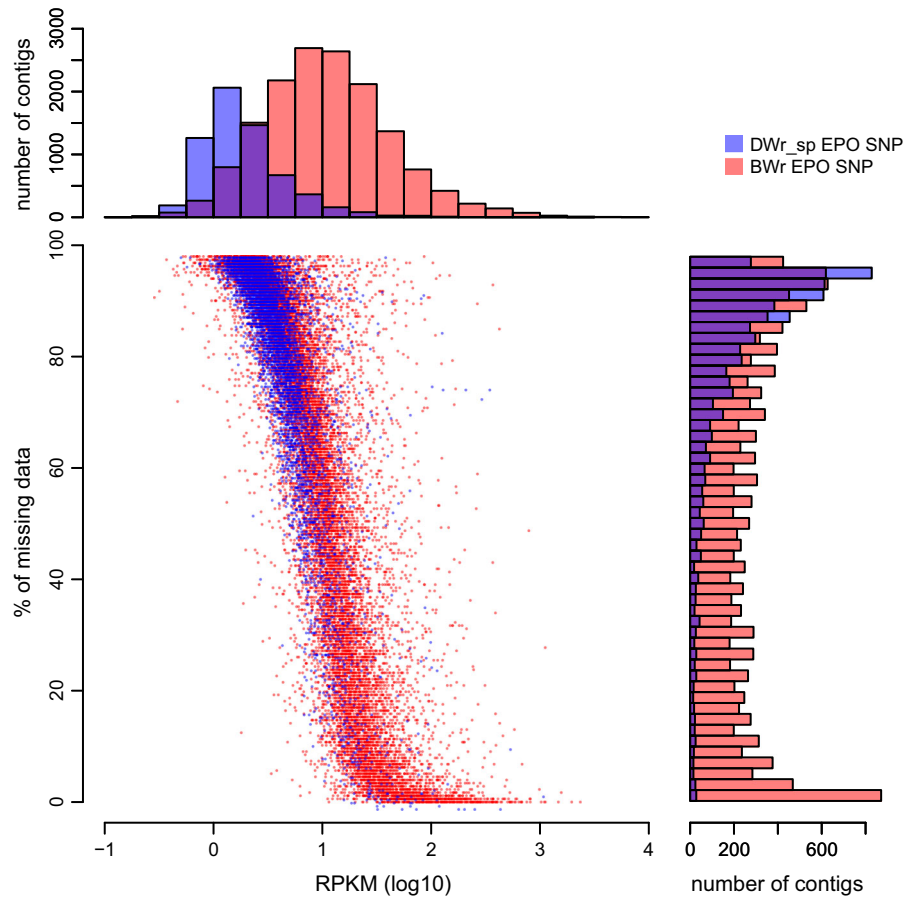

Fig. 3 Impact of sequencing depth on missing data. *Each dot*, representing a contig, is positioned according to its RPKM (*X* axis) and percentage of the 175 EPO accessions not genotyped for the SNPs of these contigs (*Y* axis). Distribution of contigs number according to missing data percentage and RPKM is also displayed on the top and left part of the figure. *Blue* (*resp red*) color represents the fact that the data refer to DWr (resp. BWr) mapping. (Color figure online)

an average 11.5×10^6 clean reads per accession. Mapping these reads on the durum wheat de novo assembly (DWr) gave significantly better results (better mapping with less heterozygous excess) but did not enable us to allocate a contig to the A or B genome as can be achieved by mapping on the BWr. This highlights the need for a reference durum genomic sequence.

Our final set of SNPs included 103,262 filtered SNPs, and 76,188 were found in EPO. We genotyped a mean of 42,394 SNPs per accession out of these 76,188. Only 3,417 of these 76,188 SNPs had an exact match in the 90 K chip developed for bread and durum panels (Wang et al. 2014), indicating that most of our SNPs are original and tailored for durum wheat studies. Using genomic GBS in a set of 254 advanced breeding lines from the CIMMYT wheat breeding program, 41,371 single nucleotide polymorphisms were discovered, many of which are dominant markers (Poland et al. 2012a). The proportion of useful reads in transcriptomic GBS was higher than that in genomic GBS and, although the transcriptome was less polymorphic than other portions of the genome, the results in terms of valuable SNPs appeared roughly equivalent with our situation. In a recent barley mapping study using genomic GBS, 1,332 high-quality SNPs with less than 20 % missing data were incorporated into the map (Liu et al. 2014). The authors also mentioned that genomic GBS data were challenging to manage as compared to their current multiplex SNP assay technology. From this standpoint, bioinformatics pipelines are much easier to implement with RNAseq data than with genomic data since the reference transcriptome is much less complex between individuals than the reference genome which contains a lot of redundant, repetitive sequences (Choulet et al. 2014, Paux et al. 2006). Furthermore, developing a high throughput marker for multiplex assay from a RNAseq SNP is straightforward.

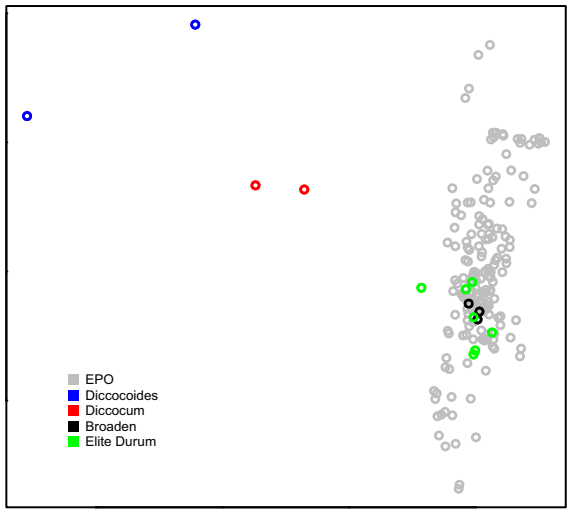

Fig. 4 Genetic diversity of the various panels. This multiple dimensional scaling (MDS) projection of the 191 accessions (colored according to their originating panel) reflects their genetic proximity. (Color figure online)

Another advantage of using GBS on a large sample is that *Fis* values may be used to check for significantly more heterozygosity than expected according to the inbreeding rate in the studied population. In these latter cases, mapping on homeologs or recent paralog genes could be suspected and the concerned SNPs are discarded. The presence of such excessive heterozygosity in our BWr_SNP could suggest that, even in the BWr established on sequences properly sorted by chromosome arms, the divergence between the A and B homeologous copies could be so slight that reads may still inappropriately map on one of the copies and create an excess of local heterozygosity. On the 3B chromosome, a high proportion of genes are subjected to high inter- and intra-chromosomal duplication (Choulet et al. 2014), and the divergence between copies may not be sufficient for a good split of reads during mapping. In a complex genome such as durum wheat, our heterozygosity filter leads to discarding almost 30 % of correctly covered and detected SNPs.

SNPs were distributed all across the chromosomes, and they could be useful in GWAS or for QTL detection. In the low recombining regions (R2), EPO had 1.3 SNPs per Mb on average considering only SNPs genotyped on at least 70 out of 175 individuals. With a rate of recombination of 0.05 cM/Mb (Choulet et al. 2014), this translates in 26 SNP/cM in R2. In the distal regions (R1 and R3), the density of SNPs genotyped on 70 individuals at least is 6.6 SNPs per Mb. As the recombination rate is estimated to 0.6–1 cM/Mb in the distal R1 and R3 region of bread wheat, EPO is covered by ca. 6 SNPs per cM on average in this area. In recombination hotspots (Saintenac et al. 2011a), this density using RNAseq GBS may be not sufficient.

The proportion of missing data depends on the gene expression level. The most expressed genes provided the lowest number of missing data for the whole sample. For the lowest expressed genes, limiting the number of missing genotypes could be obtained by adjusting the sequencing effort. This dependence between gene expression and SNP calling could be an advantage of RNAseq GBS as it allows targeting specific candidate genes whose expression depends strongly on plant organs or environmental conditions. Missing data could also be explained by the percentage of dispensable genes in the sample (see Evolution section).

In brief, RNAseq GBS revealed polymorphic sites specific to the studied panel, providing access to the expression and polymorphism in coding and regulatory regions using relatively simple standardized bioinformatics procedures. Moreover, a reasonable subset of molecular markers with few missing data could be obtained even when studying numerous accessions. The main shortcoming of RNASeq GBS is cost since the libraries are more costly than genomic libraries. If the genotyping density has to be increased, a high throughput genotyping array adapted for durum wheat could benefit from this work.

EPO is a useful panel for GWAS

The evolutionary pre-breeding population (EPO) presented here revealed a good level of genetic diversity. With a comparable sequencing effort (666 million reads) on 18 durum wheat varieties representative of the worldwide diversity (Maccaferri et al. 2011), 52,646 variants were recently discovered (Wang et al. 2014) by mapping durum reads on the reference transcriptome of the Svevo cultivar and the bread wheat CSS. As the filtering rules were not the same, it was hard to compare the raw numbers between EPO and the durum panel of Wang et al., but it could be assumed that, with 76,188 segregating highly filtered SNPs, EPO spans a high level of

diversity. Hence, it could represent an interesting source for GWAS and genomic selection programs as compared to elite durum.

In this work, durum wheat was mainly represented by seven recent Western European varieties registered in the French catalog, and their estimated π values should not be considered as a representative estimate for the whole durum taxon. It is likely that π could be higher if the panel were to be composed of some traditional landraces and lines from several countries as reported elsewhere (Maccaferri et al. 2011). Nonetheless, there has been a marked reduction in diversity through the different phases of the durum wheat history (Haudry et al. 2007, Thuillet et al. 2005), and there has also been a sharp reduction in diversity in *T.t.* ssp *durum* in recent decades (David et al.; submitted). Surprisingly, the three most promising lines from the base broadening breeding program did not show any particular specificity even if they had π values slightly larger than the elite French varieties (Fig. 1). It is possible that a strong selection for a quick return to a valuable agronomic value might have swept out many of the introgressed segments from *dicoccum* and *dicoccoides* used as parents. EPO thus represents a good opportunity to valorize a large diversity, recombined during 17 generations, in modern breeding programs.

The outcrossing rate is maintained at 10 % throughout 17 generations in EPO (Tavaud et al., in prep), thus effectively breaking long-range linkage disequilibrium. EPO appeared to be weakly structured as compared to panels usually used for GWAS, which are composed of lines of large spatiotemporal origins (old landraces and breeding lines, large geographical area of cultivation). Nevertheless, some groups of highly related lines were detected in EPO. New family lines are continuously created, submitted to human and natural selection and fixed due to the 90 % of selfing in the population. This situation is clearly expected in a preferentially selfing population (Allard 1975), but the number of groups of related lines is surprisingly high for a population evolving with a demographic size of 4,000 individuals, i.e., the probability of sampling more than two related lines from a random sample of 175 individuals could be expected to be quite low in such large populations. However, the real genetic effective sizes are often much smaller than the demographic sizes in populations (Frankham 2007). The CCPS of bread wheat managed dynamically also revealed repeatedly large differences between the apparent demographic and genetic effective sizes (Enjalbert et al. 2011, Raquin et al. 2008). In our case, the 10 % migration from the breeding program performing lines may also have an impact on the EPO structure. In GWAS based on EPO, the presence of groups of related sister lines would thus be in favor of the use of a kinship matrix in the association model (Yu et al. 2006) and the correction of linkage disequilibrium (Mangin et al. 2012). The LD decay on the 3B chromosome was very steep and not significant after 40 Mb apart from some long-distance LD on the long arm of the 3B chromosome, probably due to a structural difference in distal regions of the 3B chromosome between durum and Chinese Spring or more likely to an error in the positioning DWR-sp contigs using the best BLAST hit.

EPO thus appears to be a highly valuable resource for GWAS and for initiating a genomic selection program, i.e., it is polymorphic, weakly structured and the linkage disequilibrium decay is steep.

GBS provides information on durum wheat evolution

Genotyping via transcriptome sequencing durum wheat accessions revealed its capacity to deliver a high number of valuable SNPs with a highly repeatable genotype calling on the BWr assembled from the Chinese Spring reference. The IWGSC reference was complemented with de novo assembled durum contigs for which no significant matches on the BWr were found. If a relative small fraction of these specific DWr contigs were likely to be the transcription of repeated elements (less than 4,000 out of 25650), the majority seems to belong to another category of DNA.

Such DWr-specific contigs (DWr_spe) could be explained by two hypotheses. First, some of them correspond to long intergenic RNAs which are not annotated as genes in the reference Chinese Spring sequence (Liu et al. 2012). Second, EPO bears a significant fraction of dispensable genes (1/3 of specific DWr contigs had a blast with a known gene) that are present in some accessions and absent in others. Core and pan-genomes have been documented in bacteria and maize (Tettelin et al. 2005, Morgante et al. 2007). Common contigs between BWr and DWr may correspond to common genes (core genome) between Chinese Spring and EPO, while the specific

contigs BWr_spe and DWr_spe could correspond to dispensable genes present in Chinese Spring and EPO, respectively. In the case of dispensable genes, missing data could be the sign of the absence of the transcribed sequence, and in this case, the presence or absence of this sequence is a dominant polymorphism. This is an important issue since this source of dispensable transcripts may be a source of interesting variation for adaptation and breeding. Unfortunately, in the absence of genomic data, we cannot conclude on this point, but some of the 25,650 DWr_sp have a very low level of missing data. They are either sequences without a homolog in Chinese Spring or genes not yet sequenced or annotated by the IWGSC. Little is known about the pan and core transcriptomes of wheat compared to other species such as maize (Hirsch et al. 2014), and our data suggest that this phenomenon should be investigated in further detail. The use of a broad genetic base in EPO and in the sample panels could have amplified the phenomenon. Genetic drift, through successive bottlenecks during the *T. turgidum* sp. history (Thuillet et al. 2005), might have reduced the number of dispensable genes in the elite compartment, as is the case with the number of polymorphic sites. EPO with its large genetic basis may hold a higher number of dispensable transcripts than the elite compartment. Investigating this phenomenon in wheat will require a dedicated effort to carry out genomic sequencing of a targeted set of material, since the non-detection of a given gene in the transcriptome of an individual could be due to several causes, and does not necessarily mean that the gene does not exist in the genome of this individual. This question has to be properly addressed since the presence and the absence of dispensable polymorphisms will create a non-random distribution of missing data in the sample genealogy and may lead to ascertainment bias (Arnold et al. 2013). Similarly, for lowly expressed genes generating a number of missing data, a mutation modifying gene regulation may be responsible of a dominant variation. This claims to restrict the analysis to contigs with low levels of missing data in evolutionary studies. For GWAS in EPO, this is not really an issue.

On chromosome 3B, expressed genes and SNPs were concentrated in telomeric regions, thus confirming that chromosomes are more gene rich in telomeric region contigs than in centromeric region contigs (Choulet et al. 2010; Rustenholz et al. 2011; Rustenholz et al. 2010; Choulet et al. 2014). The proportion of polymorphic sites per Mb was also higher in the distal portion of chromosomes. This trend was observed in the different panels and surely reflects the complex interactions between recombination and selection efficiency at different loci. This reduction in the nucleotide diversity in centromeric regions argues in favor of a high level of background selection in low recombining regions of wheat chromosomes (Comeron et al. 2008; Dvorak et al. 1998).

The 175 fixed lines and their data are available on request. Specific DWr contigs sequences and SNP information are in table S4 and S5, respectively.

Acknowledgments YH had a grant from the flagship project Agropolis Resource Center for Crop Conservation, Adaptation and Diversity (ARCAD projet No 0900-001) funded by Agropolis Fondation. INRA funded the data production through its Actions Incitatives program (EPO project) and the CropDL project of the MetaSelgen Metaprogram. Sequencing was performed on the HiSeq 2000 sequencer at the Montpellier Genomix http://www.mgx.cnrs.fr sequencing facility. We thank David Manley for English editing and Oscar Defrain and Julien Bader for script developments. We also warmly thank the two anonymous referees whose comments and remarks helped to improve the manuscript.

Conflict of interest The authors declare there is no conflict of interest.

Annex 1: Genealogy of broaden lines

05ETTEU3IN1:

Pedigree: PO 45 274/837.1//S.307-3.///DD74031/131.12.//.RABD.9464/00D1102

Parents

PO 45 274: *Triticum turgidum polonicum* accession (ICARDA)

DD74031 *dicoccoïdes* accession

131.12, 837.1 = INRA durum breeding line

RABD9464, 00D1102, S307.3: durum breeding line of the GIE Blé dur

TT04DD79_27 and 79.37:

Pedigree DD74032/0.988.12//GA.5.B.38/Orlu

Parents

DD74032: dicoccoïdes accession

988.12: l INRA durum breeding line

GA.5.B.38: breeding line from GAE

ORLU: recorded durum variety, unknown pedigree

Annex 2: Molecular protocol

About 50 mg of tissue was used for each individual library. RNA was extracted using RNeasy Plant Mini Kit (Qiagen) with DNAse treatment, yielding from 30 to 60 μg of total RNAs. RNA quality was determined by the RIN (RNA integrity number) using the Agilent RNA 6000 Nano chip. RNA was quantified using the Quant-iT^{TM} RiboGreen® RNA Assay Kit and normalized leading to a RNA quantity of 2 μg per sample. Individual libraries were prepared using the TruSeq RNA sample Preparation v2 kit (Illumina Inc, CA), composed of different steps: selection, purification and fragmentation of mRNA (4 min to 94 °C), reverse transcription, synthesis of DNA double strand, and ligation of individual adaptors included index sequences, in order to allow multiplexing. To increase sequences with ligated adaptors, enrichment was made by 15 cycles of PCR using PE1.0 and PE2.0 Illumina primers and with Phusion DNA polymerase (NEB, MA).

Each indexed cDNA library was verified and quantified using a DNA 100 Chip on a Bioanalyzer 2100 and then equally mixed in pools of 24. The final library was then quantified by real-time PCR with the KAPA Library Quantification Kit for Illumina Sequencing Platforms (Kapa Biosystems Ltd, SA) adjusted to 10 nM in water and provided to the Montpellier Genomix platform for sequencing (http://www.mgx.cnrs.fr/). Final pooled cDNA library was sequenced using the Illumina mRNA-Seq, paired-end indexed protocol on a HiSeq2000 sequencer, for 2 × 100 cycles.

References

Albrechtsen A, Nielsen FC, Nielsen R (2010) Ascertainment biases in SNP chips affect measures of population divergence. Mol Biol Evol 27:2534–2547

Allard RW (1975) The mating system and microevolution. Genetics 79 Suppl:115–126

Arnold B, Corbett-Detig RB, Hartl D, Bomblies K (2013) RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. Mol Ecol 22:3179–3190

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS ONE 3:e3376

Begun DJ, Aquadro CF (1991) Molecular population genetics of the distal portion of the X chromosome in Drosophila: evidence for genetic hitchhiking of the yellow-achaete region. Genetics 129:1147–1158

Bhatia G, Patterson N, Sankararaman S, Price AL (2013) Estimating and interpreting FST: the impact of rare variants. Genome Res 23:1514–1521

Cahais V, Gayral P, Tsagkogeorga G, Melo-Ferreira J, Ballenghien M, Weinert L, Chiari Y, Belkhir K, Ranwez V, Galtier N (2012) Reference-free transcriptome assembly in non-model animals from next-generation sequencing data. Mol Ecol Resour 12:834–845

Caldwell KS, Dvorak J, Lagudah ES, Akhunov E, Luo MC, Wolters P, Powell W (2004) Sequence polymorphism in polyploid wheat and their d-genome diploid ancestor. Genetics 167:941–947

Cavanagh CR, Chao S, Wang S, Huang BE, Stephen S, Kiani S, Forrest K, Saintenac C, Brown-Guedira GL, Akhunova A, See D, Bai G, Pumphrey M, Tomar L, Wong D, Kong S, Reynolds M, da Silva ML, Bockelman H, Talbert L, Anderson JA, Dreisigacker S, Baenziger S, Carter A, Korzun V, Morrell PL, Dubcovsky J, Morell MK, Sorrells ME, Hayden MJ, Akhunov E (2013) Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. Proc Natl Acad Sci USA 110:8057–8062

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

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 F, Alaux M, Jamilloux V, Poulain J, Durand C, Bellec A, Gaspin C, Safar J, Dolezel J, Rogers J, Vandepoele K, Aury J-M, Mayer K, Berges H, Quesneville H, Wincker P, Feuillet C (2014) Structural and functional partitioning of bread wheat chromosome 3B. Science 345(6194):1249721

Comeron JM, Williford A, Kliman RM (2008) The Hill-Robertson effect: evolutionary consequences of weak selection and linkage in finite populations. Heredity (Edinb) 100:19–31

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet 12:499–510

Devos KM, Dubcovsky J, Dvorak J, Chinoy CN, Gale MD (1995) Structural evolution of wheat chromosomes 4A, 5A, and 7B and its impact on recombination. Theor Appl Genet 91:282–288

Dvorak J, Chen KC (1984) Distribution of nonstructural variation between wheat cultivars along chromosome arm 6Bp: evidence from the linkage map and physical map of the arm. Genetics 106:325–333

Dvorak J, Luo MC, Yang ZL (1998) Restriction fragment length polymorphism and divergence in the genomic regions of

high and low recombination in self-fertilizing and cross-fertilizing aegilops species. Genetics 148:423–434

Enjalbert J, Dawson JC, Paillard S, Rhone B, Rousselle Y, Thomas M, Goldringer I (2011) Dynamic management of crop diversity: from an experimental approach to on-farm conservation. C R Biol 334:458–468

Frankham R (2007) Effective population size/adult population size ratios in wildlife: a review. Genet Res 89:491–503

Ganal MW, Durstewitz G, Polley A, Berard A, Buckler ES, Charcosset A, Clarke JD, Graner EM, Hansen M, Joets J, Le Paslier MC, McMullen MD, Montalent P, Rose M, Schon CC, Sun Q, Walter H, Martin OC, Falque M (2011) A large maize (Zea mays L.) SNP genotyping array: development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. PLoS ONE 6:e28334

Gayral P, Melo-Ferreira J, Glemin S, Bierne N, Carneiro M, Nabholz B, Lourenco JM, Alves PC, Ballenghien M, Faivre N, Belkhir K, Cahais V, Loire E, Bernard A, Galtier N (2013) Reference-free population genomics from next-generation transcriptome data and the vertebrate-invertebrate gap. PLoS Genet 9:e1003457

Haudry A, Cenci A, Ravel C, Bataillon T, Brunel D, Poncet C, Hochu I, Poirier S, Santoni S, Glémin S, David J (2007) Grinding up wheat: a massive loss of nucleotide diversity since domestication. Mol Biol Evol 24:1506–1517

Hernandez P, Martis M, Dorado G, Pfeifer M, Galvez S, Schaaf S, Jouve N, Simkova H, Valarik M, Dolezel J, Mayer KF (2012) Next-generation sequencing and syntenic integration of flow-sorted arms of wheat chromosome 4A exposes the chromosome structure and gene content. Plant J 69:377–386

Hirsch CN, Foerster JM, Johnson JM, Sekhon RS, Muttoni G, Vaillancourt B, Penagaricano F, Lindquist E, Pedraza MA, Barry K, de Leon N, Kaeppler SM, Buell CR (2014) Insights into the maize pan-genome and pan-transcriptome. Plant Cell 26:121–135

Huang X (1999) CAP3: a DNA Sequence Assembly Program. Genome Res 9:868–877

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

Kilian B, Özkan H, Pozzi C, Salamini F (2009) Domestication of the Triticeae in the fertile crescent. In: Genetics and genomics of the triticeae. Springer, New York, pp 81–119

Krasileva KV, Buffalo V, Bailey P, Pearce S, Ayling S, Tabbita F, Soria M, Wang S, Consortium I, Akhunov E, Uauy C, Dubcovsky J (2013) Separating homeologs by phasing in the tetraploid wheat transcriptome. Genome Biol, 14:R66

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760

Liu J, Jung C, Xu J, Wang H, Deng S, Bernad L, Arenas-Huertero C, Chua NH (2012) Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in Arabidopsis. Plant Cell 24:4333–4345

Liu H, Bayer M, Druka A, Russell JR, Hackett CA, Poland J, Ramsay L, Hedley PE, Waugh R (2014) An evaluation of genotyping by sequencing (GBS) to map the Breviaristatum-e (ari-e) locus in cultivated barley. BMC Genom 15:104

Luo MC, Yang ZL, You FM, Kawahara T, Waines JG, Dvorak J (2007) The structure of wild and domesticated emmer wheat populations, gene flow between them, and the site of emmer domestication. Theor Appl Genet 114:947–959

Maccaferri M, Sanguineti MC, Demontis A, El-Ahmed A, Garcia del Moral L, Maalouf F, Nachit M, Nserallah N, Ouabbou H, Rhouma S, Royo C, Villegas D, Tuberosa R (2011) Association mapping in durum wheat grown across a broad range of water regimes. J Exp Bot 62:409–438

Mackay I, Powell W (2007) Methods for linkage disequilibrium mapping in crops. Trends Plant Sci 12:57–63

Mangin B, Siberchicot A, Nicolas S, Doligez A, This P, Cierco-Ayrolles C (2012) Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. Heredity (Edinb) 108:285–291

Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res 18:1509–1517

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMB Net 17:10–12

Moragues M, Comadran J, Waugh R, Milne I, Flavell AJ, Russell JR (2010) Effects of ascertainment bias and marker number on estimations of barley diversity from high-throughput SNP genotype data. Theor Appl Genet 120:1525–1534

Morgante M, De Paoli E, Radovic S (2007) Transposable elements and the plant pan-genomes. Curr Opin Plant Biol 10:149–155

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5:621–628

Nabholz B, Sarah G, Sabot F, Ruiz M, Adam H, Nidelet S, Ghesquiere A, Santoni S, David J, Glemin S (2014) Transcriptome population genomics reveals severe bottleneck and domestication cost in the African rice (Oryza glaberrima). Mol Ecol 23:2210–2227

Nevo E (2002) Evolution of wild emmer and wheat improvement: population genetics, genetic resources, and genome organization of wheat's progenitor, Triticum dicoccoides. Springer, New York

Nussbaumer T, Martis MM, Roessner SK, Pfeifer M, Bader KC, Sharma S, Gundlach H, Spannagl M (2013) MIPS PlantsDB: a database framework for comparative plant genome research. Nucleic Acids Res 41:D1144–D1151

Özkan H, Willcox G, Graner A, Salamini F, Kilian B (2011) Geographic distribution and domestication of wild emmer wheat (Triticum dicoccoides). Genet Resour Crop Evol 58:11–53

Paux E, Roger D, Badaeva E, Gay G, Bernard M, Sourdille P, Feuillet C (2006) Characterizing the composition and evolution of homoeologous genomes in hexaploid wheat through BAC-end sequencing on chromosome 3B. Plant J 48:463–474

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

Phillips S, Wolfe M (2005) Evolutionary plant breeding for low input systems. J Agric Sci 143:245–254

Poland JA, Rife TW (2012) Genotyping-by-sequencing for plant breeding and genetics. Plant Genome 5:92–102

Poland J, Endelman J, Dawson J, Rutkoski J, Wu S, Manes Y, Dreisigacker S, Crossa J, Sánchez-Villeda H, Sorrells M (2012a) Genomic selection in wheat breeding using genotyping-by-sequencing. The Plant Genome 5:103–113

Poland JA, Brown PJ, Sorrells ME, Jannink JL (2012b) Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS ONE 7:e32253

Pont C, Murat F, Guizard S, Flores R, Foucrier S, Bidet Y, Quraishi UM, Alaux M, Dolezel J, Fahima T, Budak H, Keller B, Salvi S, Maccaferri M, Steinbach D, Feuillet C, Quesneville H, Salse J (2013) Wheat syntenome unveils new evidences of contrasted evolutionary plasticity between paleo- and neoduplicated subgenomes. Plant J 76:1030–1044

Ranwez V, Holtz Y, Sarah G, Ardisson M, Santoni S, Glemin S, Tavaud-Pirra M, David J (2013) Disentangling homeologous contigs in allo-tetraploid assembly: application to durum wheat. BMC Bioinformatics 14(Suppl 15):S15

Raquin AL, Brabant P, Rhone B, Balfourier F, Leroy P, Goldringer I (2008) Soft selective sweep near a gene that increases plant height in wheat. Mol Ecol 17:741–756

Renaut S, Nolte AW, Bernatchez L (2010) Mining transcriptome sequences towards identifying adaptive single nucleotide polymorphisms in lake whitefish species pairs (Coregonus spp. Salmonidae). Mol Ecol 19(Suppl 1):115–131

Rustenholz C, Hedley PE, Morris J, Choulet F, Feuillet C, Waugh R, Paux E (2010) Specific patterns of gene space organisation revealed in wheat by using the combination of barley and wheat genomic resources. BMC Genom 11:714

Rustenholz C, Choulet F, Laugier C, Safar J, Simkova H, Dolezel J, Magni F, Scalabrin S, Cattonaro F, Vautrin S, Bellec A, Berges H, Feuillet C, Paux E (2011) A 3,000-loci transcription map of chromosome 3B unravels the structural and functional features of gene islands in hexaploid wheat. Plant Physiol 157:1596–1608

Saintenac C, Faure S, Remay A, Choulet F, Ravel C, Paux E, Balfourier F, Feuillet C, Sourdille P (2011a) Variation in crossover rates across a 3-Mb contig of bread wheat (Triticum aestivum) reveals the presence of a meiotic recombination hotspot. Chromosoma 120:185–198

Saintenac C, Jiang D, Akhunov ED (2011b) Targeted analysis of nucleotide and copy number variation by exon capture in allotetraploid wheat genome. Genome Biol 12:R88

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I (2009) ABySS: a parallel assembler for short read sequence data. Genome Res 19:1117–1123

Suneson CA (1956) An evolutionary plant breeding method. Agron J 48:188–191

Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. Genetics 105:437–460

Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, Deboy RT, Davidsen TM, Mora M, Scarselli M, Margarit y Ros I, Peterson JD, Hauser CR, Sundaram JP, Nelson WC, Madupu R, Brinkac LM, Dodson RJ, Rosovitz MJ, Sullivan SA, Daugherty SC, Haft DH, Selengut J, Gwinn ML, Zhou L, Zafar N, Khouri H, Radune D, Dimitrov G, Watkins K, O'Connor KJ, Smith S, Utterback TR, White O, Rubens CE, Grandi G, Madoff LC, Kasper DL, Telford JL, Wessels MR, Rappuoli R, Fraser CM (2005) Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". Proc Natl Acad Sci USA 102:13950–13955

Thuillet AC, Bataillon T, Poirier S, Santoni S, David JL (2005) Estimation of long-term effective population sizes through the history of durum wheat using microsatellite data. Genetics 169:1589–1599

Trebbi D, Maccaferri M, de Heer P, Sørensen A, Giuliani S, Salvi S, Sanguineti MC, Massi A, van der Vossen EAG, Tuberosa R (2011) High-throughput SNP discovery and genotyping in durum wheat (Triticum durum Desf.). Theor Appl Genet 123:555–569

Wang S, Wong D, Forrest K, Allen A, Chao S, Huang BE, Maccaferri M, Salvi S, Milner SG, Cattivelli L, Mastrangelo AM, Whan A, Stephen S, Barker G, Wieseke R, Plieske J, International Wheat Genome Sequencing C, Lillemo M, Mather D, Appels R, Dolferus R, Brown-Guedira G, Korol A, Akhunova AR, Feuillet C, Salse J, Morgante M, Pozniak C, Luo MC, Dvorak J, Morell M, Dubcovsky J, Ganal M, Tuberosa R, Lawley C, Mikoulitch I, Cavanagh C, Edwards KJ, Hayden M, Akhunov E (2014) Characterization of polyploid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array. Plant Biotechnol J 12(6):787–796

Wasmuth JD, Blaxter ML (2004) prot4EST: translating expressed sequence tags from neglected genomes. BMC Bioinform 5:187

Williford A, Comeron JM (2010) Local effects of limited recombination: historical perspective and consequences for population estimates of adaptive evolution. J Hered 101(Suppl 1):S127–S134

Wright S (1950) Genetical structure of populations. Nature 166:247–258

International Barley Genome Sequencing C, Mayer KF, Waugh R, Brown JW, Schulman A, Langridge P, Platzer M, Fincher GB, Muehlbauer GJ, Sato K, Close TJ, Wise RP, Stein N (2012) A physical, genetic and functional sequence assembly of the barley genome. Nature 491:711–716

Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38:203–208

Zaharieva M, Ayana NG, Al Hakimi A, Misra SC, Monneveux P (2010) Cultivated emmer wheat (Triticum dicoccon Schrank), an old crop with promising future: a review. Genet Resour Crop Evol 57:937–962