Published in final edited form as:
Nature. 2012 November 29; 491(7426): 705–710. doi:10.1038/nature11650.

Analysis of the bread wheat genome using whole genome shotgun sequencing

Rachel Brenchley[1,§], **Manuel Spannagl**[2,§], **Matthias Pfeifer**[2,§], **Gary L.A. Barker**[3,§], **Rosalinda D'Amore**[1,§], **Alexandra M. Allen**[3], **Neil McKenzie**[4], **Melissa Kramer**[5], **Arnaud Kerhornou**[6], **Dan Bolser**[6], **Suzanne Kay**[1], **Darren Waite**[4], **Martin Trick**[4], **Ian Bancroft**[4], **Yong Gu**[7], **Naxin Huo**[7], **Ming-Cheng Luo**[8], **Sunish Sehgal**[9], **Sharyar Kianian**[9], **Bikram Gill**[9], **Olin Anderson**[7], **Paul Kersey**[6], **Jan Dvorak**[8], **Richard McCombie**[5], **Anthony Hall**[1,*], **Klaus F.X. Mayer**[2,*], **Keith J. Edwards**[3,*], **Michael W. Bevan**[4,*], and **Neil Hall**[1,*]

[1]Centre for Genome Research, University of Liverpool, Liverpool, UK

[2]MIPS/IBIS, Helmholtz-Zentrum München, Neuherberg, DE

[3]School of Biological Sciences, University of Bristol, Bristol, UK

[4]John Innes Centre, Norwich, UK

[5]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA

[6]European Bioinformatics Institute, Hinxton, UK

[7]USDA Western Regional Laboratory, Albany, CA, USA

[8]Dept Agronomy and Range Science, UC Davis, Davis, CA, USA

[9]Dept Plant Pathology, Kansas State University, Manhattan, KS, USA.

Summary

Bread wheat (*Triticum aestivum*) is a globally important crop, accounting for 20% of the calories consumed by mankind. We sequenced its large and challenging 17 Gb hexaploid genome using 454 pyrosequencing and compared this with the sequences of diploid ancestral and progenitor genomes. Between 94,000-96,000 genes were identified, and two-thirds were assigned to the A, B and D genomes. High-resolution synteny maps identified many small disruptions to conserved gene order. We show the hexaploid genome is highly dynamic, with significant loss of gene family members upon polyploidization and domestication, and an abundance of gene fragments. Several classes of genes involved in energy harvesting, metabolism and growth are among expanded gene families that could be associated with crop productivity. Our analyses, coupled

[*]senior and corresponding authors .
[§]joint first authors
Requests for materials should be addressed to Michael Bevan, and correspondence to Klaus Mayer, Michael Bevan, Neil Hall or Keith Edwards.

Author Contributions K.J.E., M.W.B., N.H. and A.H. conceived the project, R.M., M.K., M.T., I.B., J.D., M.C.L., O.A., S.K., N.Huo, B.G., S.S., provided data and advice, R.dA., N.McK. and S.K. conducted experiments, K.F.X.M., N.H. and M.W.B. planned and conducted analyses, R.B., M.S., M.P., G.B., A.A., D.B., D.W., P.K., A.H. carried out analyses. K.J.E., A.H., R.M., R.B., contributed to the text and M.W.B., N.H., and K.F.X.M. wrote the manuscript. All authors commented on the manuscript.

Supplementary Information is linked to the online version of the paper at www.nature/com/nature

Author Information All sequence data has been deposited in publicly accessible databases, described in the Supplemental Online Information. Sequence assemblies, annotated gene sequences and their relationships are available for download from EBI and viewing in a synteny-based Ensembl genome browser at www.ebi.ac.uk.

Conflict of Interest: W.R.M. has participated in Illumina sponsored meetings and received travel reimbursement and honoraria for presenting at these events.

Europe PMC Funders Author Manuscripts

with the identification of extensive genetic variation, provide a new resource for accelerating gene discovery and improving this major crop.

Keywords

wheat/polyploid/genome; analysis/food; security/next-generation; sequencing

Europe PMC Funders Author Manuscripts

Introduction

With a global output of 681million tonnes in 2011[1], bread wheat (*Triticum aestivum*) accounts for 20% of human consumption of calories[2] and is an important source of protein, vitamins and minerals. It originated from hybridization between cultivated tetraploid emmer wheat (AABB, *T. dicoccoides*) and diploid goat grass (DD, *Aegilops tauschii*) approximately 8,000 years ago[3]. Bread wheat cultivation and domestication has been directly associated with the spread of agriculture and settled societies such that it is now one of the most widely cultivated crops due to its high yields and nutritional and processing qualities. The three diploid progenitor genomes, AA from *T. urartu*, BB from an unknown, possible *Sitopsis* section species (to which *Ae. speltoides* belongs), and DD from *Ae. tauschii*, radiated from a common Triticeae ancestor between 2.5-4.5 MYA (million years ago) and AABB tetraploids arose less than 0.5 MYA[4,5]. Nucleotide diversity in the AABB and DD genomes is substantially reduced compared to ancestral populations, indicating a major diversity bottleneck upon the transition to cultivated lines[6].

Grass genomes exhibit extensive long-range conservation of gene order[7-9]. Nevertheless, they are highly dynamic due to the activities of repeats contributing to tremendous variation in genome size[10], changes in local gene order, and to pseudogene formation, particularly in larger genomes such as maize[11] and wheat[12]. From analysis of BAC contigs on chromosome 3B, the 17 Gb genome was estimated to be composed of approximately 80% repeats, primarily retroelements, with a gene density between 1/87-1/184kb[13]. Despite both the substantial knowledge gained of the wheat genome from these studies and the central importance of the wheat crop, a comprehensive genome-wide analysis of gene content has yet to be conducted due to its large size, repeat content and polyploid complexity.

We have analysed low-coverage, long-read (454) shotgun sequence of the hexaploid wheat genome using gene sequences from diverse grasses. From this we created assemblies of wheat genes in an orthologous gene family framework, utilized diploid wheat relatives to classify homoeologous relationships, and defined a genome wide catalog of single nucleotide polymorphisms (SNPs) in the A, B and D genomes. These analyses provide a new foundation for genetic and genomic analysis of this key crop.

Sequence analysis

The wheat variety Chinese Spring (CS42) was selected for sequencing because of its wide use in genome studies[14,15]. Purified nuclear DNA was sequenced using Roche 454 pyrosequencing technology (GS FLX Titanium and GS FLX+ platforms) to generate 85 Gb of sequence (220m reads), approximately 5× coverage based on an estimated genome size of 17 Gb. Supplementary Table 1 shows 79% of the reads had matches to the TREP repeat database, and most hit retrotransposons, consistent with previous studies[13]. To identify A-, B- and D-genome derived gene assemblies in the hexaploid sequences, we utilized Illumina sequence assemblies of *T. monococcum*, related to the A genome donor, *Ae. speltoidies* cDNA assemblies, and 454 sequences from the D genome donor *Aegilops tauschii* respectively. The SOLiD platform was used to generate additional sequence of CS42 and three commercial wheat varieties to increase the accuracy of homoeologous SNP

identification. Datasets are summarized in Table 1a and Supplementary Table 2, and SNP identification methods are described in Supplementary Section 5.2.

Sequence assembly

An Orthologous group Assembly (OA) (Supplementary Table 3) was created by clustering 454 reads by sequence similarity to orthologous grass gene sequences, and separate assembly of the clusters at high stringency using Newbler (Supplementary section 2). The orthologous genes were derived from rice[16], sorghum[8], *Brachypodium*,[9] and barley full-length cDNAs, by OrthoMCL[17] clustering. This generated 20,496 Orthologous Groups (OG) (Supplementary Table 4 and Supplementary Figure 1). The gene model with highest similarity to wheat (termed the "OG Representative") was selected from each OG by stringent BLASTX comparison to a Low Copy-number Genome assembly (LCG) made by filtering out repetitive sequences and assembling the remaining low-copy sequences *de novo* (Supplementary Table 3). Nearly 90% of the metabolic genes in *Arabidopsis* matched OG Representatives. The 20,051 OG Representatives matched 92% of publicly available wheat full-length cDNAs[18] and 78.7% of harvESTs (Supplementary Figure 2), indicating they represent nearly all wheat genes.

We optimized parameters for wheat gene assembly using MetaSim[19] to generate simulated 5× 454 reads from the allotetraploid maize genome and from a triplicated rice gene set, with the introduction of sequence variation (Supplementary Section 2.7). Similar degrees of coverage over the OG Representatives were seen for the simulated datasets and wheat 454 reads (Figures 1A). Rice reads followed the same depth distribution as the wheat reads (Figure 1B), suggesting they are a reasonable representation of hexaploid sequences. Maize reads covered their OG Representatives to a median depth of approximately 5, consistent with 5× coverage.

Simulated maize and triplicated rice 454 reads were used to optimize assembly parameters. Assembly at 99% minimum sequence identity (mi) using 40bp overlap length predicted gene family sizes most accurately (Supplementary Figures 3-6). Wheat 454 reads were pre-processed (Supplementary Table 5) and assembled using 99% mi (Supplementary Tables 6 and 7) to create the OA. Figure 1B shows that the depth of coverage of the OA assembly followed a similar pattern to maize, consistent with multiple gene copies. In contrast, the low depth coverage by the LCG assembly suggested gene family numbers were collapsed. The number of wheat assemblies for each OG Representative was calculated to determine gene copy numbers (Supplementary Table 7). Figure 1C shows that most OG Representatives had between 1-5 distinctive wheat gene assemblies, with a peak of 2 genes.

The A, B and *Ae. tauschii* (D) genome gene sets[13,20,21] have been estimated to be 28,000, 38,000 and 36,000 respectively. We estimated the number of genes in the hexaploid wheat genome to range between 94,000-96,000 (Supplementary Section 2.10). This is reasonably consistent with estimates based on wheat chromosome sequences[13]. Comparing our transcriptome assembly (Supplementary Sections 2.8 and 2.9) and wheat harvEST to the wheat OG Representatives showed that 76% and 65% respectively were expressed under the conditions used for RNA isolation. Similar results were found in barley[22], rice[16] and maize[23], indicating the assemblies are *bona fide* wheat genes.

We defined the overall extent of gene conservation between wheat and the most closely related sequenced pooid grass *Brachypodium distachyon*[9,24]. Figure 2 track 1 shows that there is a high degree of overlap between the gene sets of *Brachypodium* and wheat, but with regions of lower conservation, for example on *Brachypodium* chromosomes 1 and 4. Syntenic maps of the *Brachypodium* genome and the A, B and D chromosome groups were created by integrating high-density wheat EST-based markers[25] with *Brachypodium* genes

(Figure 2 tracks 5, 6 and 7 respectively). Supplementary Figure 7 shows the A, B and D genome markers separately. Syntenic alignments were readily identifiable and conformed to the predicted major patterns[9,26]. We identified many insertions and/or translocations of blocks of genes within the overall conserved patterns of gene order, including the major rearrangement on chromosome 4A as shown on *Brachypodium* chromosome 1[20]. Lower marker density on the D genome is evident in track 7. The higher resolution genetic map identified a new syntenic alignment of Triticeae group 5 to *Brachypodium* chromosome 3 genes.

Genome change in polyploid wheat

We determined the influence of polyploidy on gene content in hexaploid wheat by defining the sizes of gene families in hexaploid wheat and the diploid progenitor *Ae.tauschii* from the copy number of genes for each OG Representative, which were then paired with the gene family size of the OG Representative in sequenced diploid grasses (Supplementary Section 2.6). The mean family size was 1.4 members. Supplementary Figure 8 shows relationships between wheat and diploid orthologous gene family across the full scale of orthologous gene family sizes. This approach accurately reconstructed gene family sizes in simulated maize and "hexaploid" rice genomes (Figures 3A and 3B), although larger gene family sizes tended to be under-estimated. Figures 3C and 3D show the relationships between *Ae. tauschii* and wheat genes. Single member gene families in hexaploid wheat and *Ae. tauschii* were maintained to a similar extent as those seen in sequenced diploid grasses, consistent with Southern blot analyses of single copy genes[27]. Using the D genome as a diploid reference, we calculated the ratio of gene family sizes in hexaploid:diploid Triticeae as between 2.5-2.7:1, derived from the geometric mean and the slopes of the blue and red lines in Figure 3E respectively. Comparing the expected ratio of 3:1 for the hexaploid indicated the loss of between 10,000-16,000 genes in hexaploid wheat compared to the three diploid progenitors (Supplementary section 2.10). This is consistent with earlier studies of gene loss in newly synthesized wheat polyploids [28] and the erosion of genetic diversity during wheat domestication[6].

Despite this overall trend of gene family size reduction, gene families with fewer or greater than expected members were identified in *Ae. tauschii* and hexaploid wheat, shown as green dots (expanded) and brown dots (contracted) in Figures 3C and 3D, respectively. Supplementary Tables 10, 11 and 12 show the over- and under-represented functional categories of proteins. Most of the over-represented categories in expanded gene families are common to wheat and *Ae.tauschii*: these include ribosome proteins, components of photosystem II, storage proteins, transposon-related proteins, cytochrome P450s, NB-ARC domain proteins involved in defense responses, pollen allergen-related proteins, and F-box proteins. Five of the 11 families encoding hydrogen ion transmembrane transporters were significantly more numerous in *Ae. tauschii* compared to wheat. Analysis of gene families (Supplementary Figure 9) showed they encode different subunits of ATPases. We speculate that they may provide proton gradients to support Na^+ exclusion in *Ae. tauschii*[29] and the accumulation of minerals in other *Aegilops* species[30].

Pseudogene analysis

Several classes of plant DNA transposons[31,32] and retroelements[33] create and amplify gene fragments, disrupt genes and create pseudogenes, which can influence gene expression through epigenetic mechanisms[34]. We identified a set of almost 233,000 gene fragments that mapped to the same regions of their OG Representatives, forming "stacks" that were sufficiently divergent not to assemble into their cognate gene assemblies (Figure 4A). Two classes were identified; those containing Pfam domains, and those aligning with non-Pfam domains of OG Representatives. Nearly 30% of the OG Representatives had associated gene

fragments (Supplementary Table 13) that most frequently covered between 10-12% of their length (Figure 4B). Figure 4C shows that the alignment identities of gene fragments against their OG Representatives were substantially lower than the identities for cognate regions within wheat gene assemblies. Supplementary Figure 10 shows the distribution of "stacks" along genes and Ka/Ks analyses. Pfam domains found in "stacks" were enriched for Zn-finger motifs in Mutator transposons (Supplementary Table 14), consistent with their role in pseudogene formation[31]. F-box, protein kinase and NB-ARC domains, found in the most rapidly evolving gene families in plants[9,35], are also over-represented.

Determining homoeologous relationships of gene assemblies

We classified gene assemblies as A-, B- or D-derived according to sequence similarity to Illumina sequence assemblies from *T. monococcum*, cDNA assemblies from *Ae. Speltoides*, and 454 reads from *Ae. tauschii* respectively by applying a Support Vector Machine-learning approach (SVM, Supplementary Section 5, Supplementary Figures 11 and 12, Supplementary Tables 15-18). Supplementary Figure 13 shows 66% of the gene assemblies were classified with high overall precision (>70%) and recall into 28.3% A genome, 29.2% B genome and 33.8% D genome. The remaining 34% with low classification probabilities are likely to be very similar homoeologs. Comparison to a subset of A, B and D genome SNPs confirmed 72% of A genome classifications and 85% of D genome classifications (Figure 2 and Supplementary Table 19). Discrimination of putative B genome genes was only ≈60%, possibly due to the use of cDNA sequences for classification when most of the informative sequence polymorphisms are intronic, and due to uncertainty about the ancestry of the B genome[5]. The set of 132,552 SNPs allocated to the A, B and D genomes is displayed using *Brachypodium* as a template in Figure 2, tracks 2-4.

There were no significant differences in the distribution of GOSlim molecular function categories in the A, B and D genes (Supplementary Figure 14), indicating that at this level of functional categorization there is no biased gene loss[36] in one of the genomes. Nevertheless, analysis of GOSlim terms associated with stop codons in A, B and D gene assemblies showed a strong tendency to retain functional copies of genes encoding transcription factors in all three genomes (Supplementary Figure 15), similar to the preferential retention of these genes in Arabidopsis genome duplications[37]. This indicates that genome-specific transcriptional regulatory networks tend to be maintained in wheat.

Conclusions

Using whole genome 454 sequencing, we assembled gene sequences representing an essentially complete gene set, and a significant number were assigned to the A, B or D genomes. Although the assemblies are fragmentary, they form a powerful framework for identifying genes, accelerating further genome sequencing, and facilitating genome-scale analyses. The identification of over 132,000 SNP in A, B and D genes facilitates QTL analysis and association studies of traits. Comparison with the sequences of diploid progenitors and relatives showed pronounced reductions in the size of large gene families in wheat despite the relatively recent formation of the hexaploid (Figure 3E), consistent with smaller scale analyses[28,38]. The scale of gene loss in hexaploid wheat compared to maize[36] and *Brassica rapa*[39] is significantly less, possibly due to its relatively recent origin and the absence of inter-genome recombination[40]. Nevertheless gene loss in wheat could be rapid, as shown in the newly created allopolyploid *Tragopogon miscellus*[41]. Most functional classes show equal gene loss in the three genomes, but families of transcription factors showed a clear tendency to be retained as functional genes in all three genomes. These may maintain transcriptional networks in each genome and contribute to non-additive gene expression[42] and genome plasticity. In contrast to the overall loss of gene family members, several classes of gene families with predicted roles in defence, nutritional content, energy

metabolism and growth have increased sizes in the Triticeae lineage, possibly due to selection during domestication. Major efforts are underway to improve wheat productivity by increasing genetic diversity in breeding material and through genetic analysis of traits[43]. The genomic resources we have created promise to accelerate progress by facilitating the identification of useful variation in genes of wheat landraces and progenitor species, and by providing genomic landmarks to guide the progeny selection. Analysis of complex polygenic traits such as yield and nutrient use efficiency will also be accelerated, contributing to sustainable increases in wheat crop production.

Methods Summary

A single seed descent line of *Triticum aestivum* landrace "Chinese Spring" was sequenced as it is widely used for cytogenetic analysis[44] and physical mapping[15]. *Triticum monococcum* accession 4342-96 is a community standard line for TILLING, physical mapping and genetic analysis, and *Aegilops tauschii* ssp *strangulata* accession AL8/78, used for physical and genetic mapping, was sequenced using 454 technology[48].

Sequence for *T. aestivum* wheat gene assembly was generated using Roche 454 Pyrosequencing on the GS FLX Titanium and GS FLX+ platforms. Additional sequence read datasets for *T. aestivum*, *T. monococcum* and *A. tauschii* were generated using 3 platforms: Illumina, 454 and SOLiD, to analyse homoeologous sequences and SNPs (a list of all datasets is in Supplementary Table 2). Orthologous Groups were created from rice, sorghum, and *Brachypodium distachyon* genome sequences and barley full-length cDNA sequences. Wheat gene assemblies were named according to their Orthologous Group Representative, and identified by a seven-digit identifier and their predicted genome *viz*: Traes_Bradi1g12345_0000001_D and Traes_Sb3g33333_6543210_A. Gene and cDNA assemblies can be searched at http://mips.helmholtz-muenchen.de/plant/wheat/index.jspp. Sequence assemblies are available for download from EBI accession PRJEB217 (OA: CALO01000001-CALO01945079; LCG: CALP010000001-CALP-15321847; cDNA) and annotated gene sequences and their relationships can be viewed in a *Brachypodium* synteny-based Ensembl genome browser at: http://plants.ensembl.org/Brachypodium_distachyon.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

DNA sequence was generated by: The University of Liverpool Centre for Genomic Research, Liverpool, UK; 454 Life Sciences (Banford, CT, USA); The Cold Spring Harbor Woodbury Genome Centre (Woodbury, NY, USA); and The Genome Analysis Centre (Norwich, UK). This work was supported by the UK Biological and Biotechnological Sciences Research Council (BBSRC) grants BB/G012865, BB/G013985/1 and BB/G013004/1 to KJE, MWB and NH, a Wolfson Merit Award from the Royal Society to NH, BBSRC Strategic Programme Grant B/J004588/1 (GRO) to MWB, EC TriticeaeGenome grant #212019 to KFXM and MWB, The TRITEX Project of the Plant20130 Initiative of the German Ministry of Education and Research grant #0315954C to KFXM, EC Transplant Grant 283496 to KFXM and PK, A BBSRC Career Development Fellowship BB/H022333/1 to AH, NSF grants IOS-1032105 and DBI-0923128 to WRM, USDA-NIFA grant 2008-35300-04588 to BSG, and NSF Grants DBI-0701916 to JD and DBI-0822100 to SFK.

References

1. WASDE. World Agricultural Supply and Demand. 2012.

2. FAOSTAT. Food and Agriculture Organisation of the United Nations; Rome, Italy: 2011.
3. Nesbitt, M.; Samuel, D. From staple crops to extinction? The archaeology and history of hulled wheats. International Plant Genetic Resources Institute; 1996.
4. Dvorak J, Akhunov ED, Akhunov AR, Deal KR, Luo MC. Molecular characterization of a diagnostic DNA marker for domesticated tetraploid wheat provides evidence for gene flow from wild tetraploid wheat to hexaploid wheat. Molecular Biology and Evolution. 2006; 23:1386–1396. [PubMed: 16675504]
5. Salse J, et al. New insights into the origin of the B genome of hexaploid wheat: evolutionary relationships at the SPA genomic region with the S genome of the diploid relative Aegilops speltoides. BMC Genomics. 2008; 9:555. [PubMed: 19032732]
6. Haudry A, et al. Grinding up wheat: a massive loss of nucleotide diversity since domestication. Molecular Biology and Evolution. 2007; 24:1506–1517. [PubMed: 17443011]
7. Moore G, Devos KM, Wang Z, Gale MD. Cereal genome evolution. Grasses, line up and form a circle. Curr Biol. 1995; 5:737–739. [PubMed: 7583118]
8. Paterson AH, et al. The Sorghum bicolor genome and the diversification of grasses. Nature. 2009; 457:551–556. [PubMed: 19189423]
9. Internation Brachypodium Initiative. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. Nature. 2010; 463:763–768. [PubMed: 20148030]
10. Smith DB, Flavell RB. Characterisation of the wheat genome by association genetics. Chromosoma. 1975; 50:223–242.
11. Baucom RS, et al. Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. PLoS Genetics. 2009; 5:e1000732. [PubMed: 19936065]
12. Wicker T, et al. Frequent gene movement and pseudogene evolution is common to the large and complex genomes of wheat, barley, and their relatives. The Plant Cell. 2011; 23:1706–1718. [PubMed: 21622801]
13. Choulet F, et al. Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. The Plant Cell. 2010; 22:1686–1701. [PubMed: 20581307]
14. Gill BS, et al. A workshop report on wheat genome sequencing: International Genome Research on Wheat Consortium. Genetics. 2004; 168:1087–1096. [PubMed: 15514080]
15. Paux E, et al. A physical map of the 1-gigabase bread wheat chromosome 3B. Science. 2008; 322:101–104. [PubMed: 18832645]
16. International Rice Genome Sequencing Project. The map-based sequence of the rice genome. Nature. 2005; 436:793–800. [PubMed: 16100779]
17. Li L, Stoeckert CJ Jr. Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Research. 2003; 13:2178–2189. [PubMed: 12952885]
18. Mochida K, Yoshida T, Sakurai T, Ogihara Y, Shinozaki K. TriFLDB: a database of clustered full-length coding sequences from Triticeae with applications to comparative grass genomics. Plant Physiology. 2009; 150:1135–1146. [PubMed: 19448038]
19. Richter DC, Ott F, Auch AF, Schmid R, Huson DH. MetaSim: a sequencing simulator for genomics and metagenomics. PLoS One. 2008; 3:e3373. [PubMed: 18841204]
20. Hernandez P, et al. Next-generation sequencing and syntenic integration of flow-sorted arms of wheat chromosome 4A exposes the chromosome structure and gene content. The PlantJjournal. 2012; 69:377–386.
21. Massa AN, et al. Gene space dynamics during the evolution of Aegilops tauschii, Brachypodium distachyon, Oryza sativa, and Sorghum bicolor genomes. Molecular Biology and Evolution. 2011; 28:2537–2547. [PubMed: 21470968]
22. Internation Barley Sequencing Consortium. Sequence and analysis of the barley genome. 2012. submitted
23. Schnable PS, et al. The B73 maize genome: complexity, diversity, and dynamics. Science. 2009; 326:1112–1115. [PubMed: 19965430]
24. Lee EK, et al. A functional phylogenomic view of the seed plants. PLoS Genetics. 2011; 7

25. Allen AM, et al. Transcript-specific, single-nucleotide polymorphism discovery and linkage analysis in hexaploid bread wheat (Triticum aestivum L.). Plant Biotechnology Journal. 2011; 9:1086–1099. [PubMed: 21627760]
26. Salse J, et al. Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. The Plant Cell. 2008; 20:11–24. [PubMed: 18178768]
27. Qi LL, et al. A chromosome bin map of 16,000 expressed sequence tag loci and distribution of genes among the three genomes of polyploid wheat. Genetics. 2004; 168:701–712. [PubMed: 15514046]
28. Ozkan H, Levy AA, Feldman M. Allopolyploidy-induced rapid genome evolution in the wheat (Aegilops-Triticum) group. The Plant Cell. 2001; 13:1735–1747. [PubMed: 11487689]
29. Shavrulov Y, Langridge P, Tester M. Salinity tolerance and sodium exclusion in genus Triticum. Breeding Science. 2009; 59:671–678.
30. Wang S, Yin L, Tanaka K, Tanaka H, Tsujimoto H. Wheat-Aegilops chromosome addition lines showing high iron and zinc contents in grains. Breeding Science. 2011; 61:189–195.
31. Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR. Pack-MULE transposable elements mediate gene evolution in plants. Nature. 2004; 431:569–573. [PubMed: 15457261]
32. Morgante M, et al. Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. Nature Genetics. 2005; 37:997–1002. [PubMed: 16056225]
33. Jin YK, Bennetzen JL. Integration and nonrandom mutation of a plasma membrane proton ATPase gene fragment within the Bs1 retroelement of maize. The Plant Cell. 1994; 6:1177–1186. [PubMed: 7919987]
34. Lippman Z, et al. Role of transposable elements in heterochromatin and epigenetic control. Nature. 2004; 430:471–476. [PubMed: 15269773]
35. Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature. 2000; 408:796–815. [PubMed: 11130711]
36. Schnable JC, Springer NM, Freeling M. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proceedings of the National Academy of Sciences of the United States of America. 2011; 108:4069–4074. [PubMed: 21368132]
37. Maere S, et al. Modeling gene and genome duplications in eukaryotes. Proceedings of the National Academy of Sciences of the United States of America. 2005; 102:5454–5459. [PubMed: 15800040]
38. Gu YQ, Coleman-Derr D, Kong X, Anderson OD. Rapid genome evolution revealed by comparative sequence analysis of orthologous regions from four triticeae genomes. Plant Physiology. 2004; 135:459–470. [PubMed: 15122014]
39. Mun JH, et al. Genome-wide comparative analysis of the Brassica rapa gene space reveals genome shrinkage and differential loss of duplicated genes after whole genome triplication. Genome Biology. 2009; 10:R111. [PubMed: 19821981]
40. Riley R. Genetic control of cytologically diploid behaviour of hexaploid wheat. Nature. 1958; 182:713–715.
41. Buggs RJ, et al. Rapid, repeated, and clustered loss of duplicate genes in allopolyploid plant populations of independent origin. Current Biology. 2012; 22:248–252. [PubMed: 22264605]
42. Pumphrey M, Bai J, Laudencia-Chingcuanco D, Anderson O, Gill BS. Nonadditive expression of homoeologous genes is established upon polyploidization in hexaploid wheat. Genetics. 2009; 181:1147–1157. [PubMed: 19104075]
43. Tester M, Langridge P. Breeding technologies to increase crop production in a changing world. Science. 2010; 327:818–822. [PubMed: 20150489]
44. Sears, ER. Nullisomic-tetrasomic combinations in hexaploid wheat. Oliver and Boyd; 1966. p. 22-45.

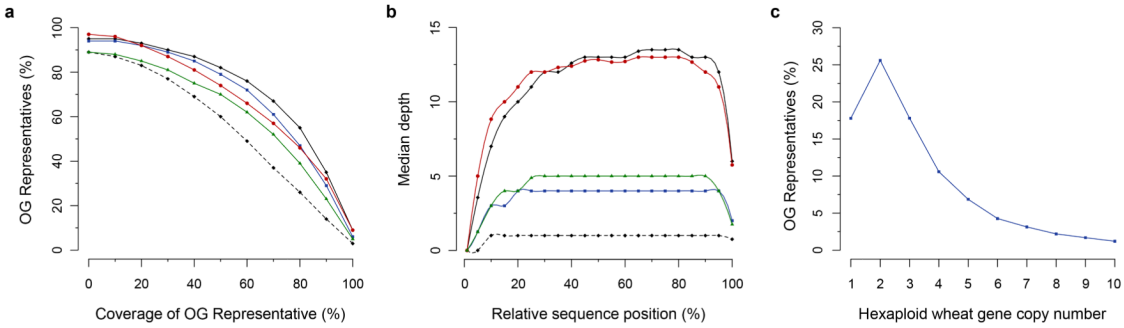

Figure 1. Coverage of OG Representatives by wheat 454 sequence reads and simulated 454 reads from rice and maize

a. Coverage of OG Representatives by repeat-masked wheat 454 sequence reads (black line), wheat LCG (black dashed line), the OA (blue line), together with rice genes (red line) and maize simulated reads (green line).

b. Median coverage depth over protein coding regions of OG Representatives (N terminus = 0; C = 100). The colour coding is the same as in panel 1A, except simulated hexaploid reads from rice (red line) were used.

c. The distribution of wheat gene copy numbers from the OA.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

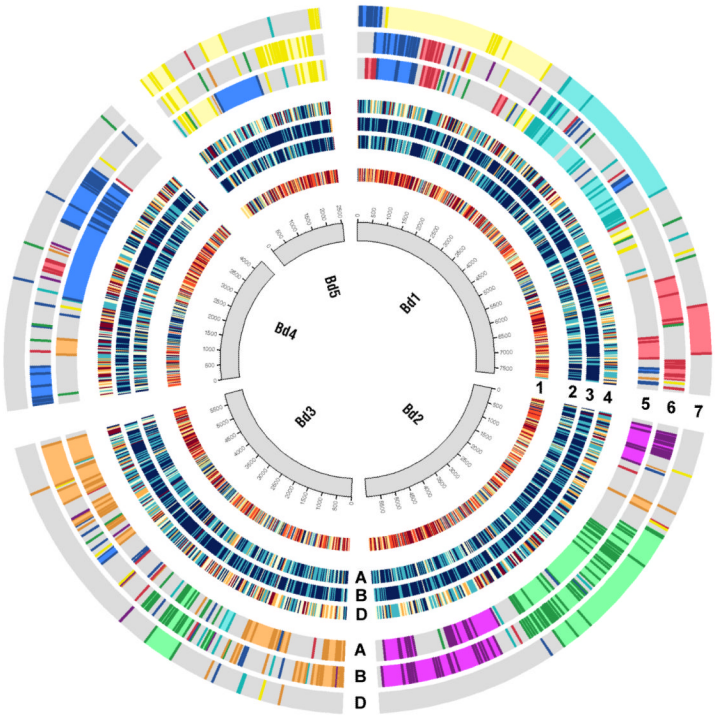

Figure 2. Alignment of wheat 454 reads, SNPs and genetic maps to the *Brachypodium distachyon* genome

The inner circle represent gene order on the 5 *Brachypodium* chromosomes. Track 1 illustrates conservation between wheat 454 reads and *Brachypodium* genes, shown as a window of genes present in wheat. Tracks 2-4 show SNP density (the mean number of SNPs per gene in a window of 20 genes) in the A (track 2), B (track 3) and D (track 4) genomes of wheat. Tracks 5-7 display wheat synteny with *Brachypodium* for the A (track 5) B (track 6) and D (track 7) genomes. Genetic markers[25](shown in darker colours) were colour-coded by wheat chromosome. Gaps between markers were filled in to show synteny (lighter colours).

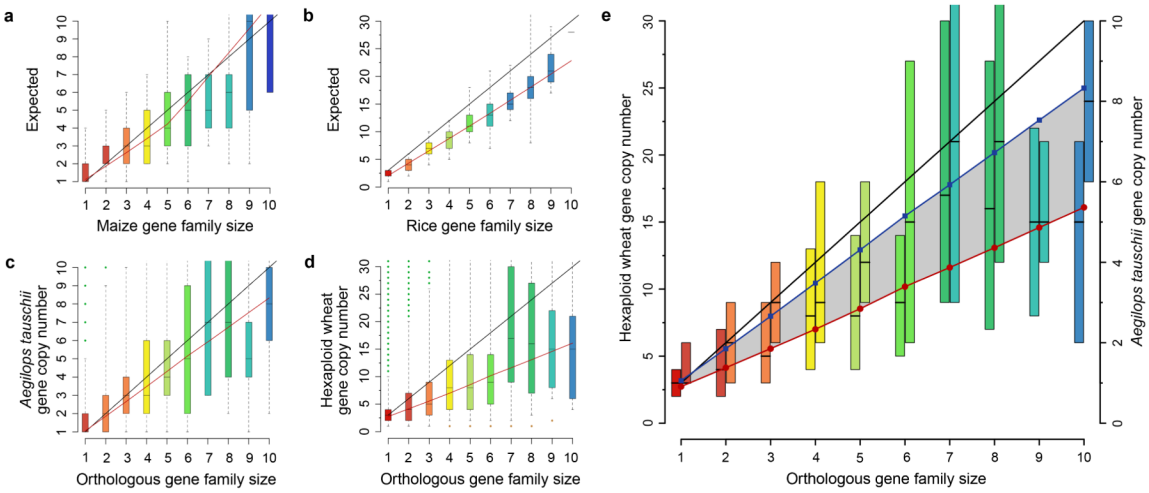

Figure 3. Gene family sizes in orthologous assemblies of hexaploid wheat, *Ae. tauschii*, simulated maize and hexaploid rice

The boxes and whiskers contain 50% and 90% of the OA genes respectively. The box colours indicate the number of genes in diploid gene families of different sizes (x axis). The black lines represent expected gene family sizes; the red line is the gene family size determined from the OA, derived by polynomial regression fit. Only gene families with up to ten members are shown.

a. Maize gene family sizes predicted from orthologous assembly of simulated 454 reads.

b. Rice gene family sizes predicted from orthologous assembly of simulated 454 reads derived from triplicated rice genes.

c. *Ae. tauschii* gene family sizes obtained from orthologous assembly of repeat-masked 454 reads. Expanded gene families are shown as green dots.

d. Wheat gene family sizes in the OA.

e. Amalgamation of wheat and *Ae.tauschii* gene copy numbers. The black line shows the expected gene copy numbers for wheat and *Ae. tauschii* respectively. The red line shows the regression fit for wheat, and the blue line for *Ae.tauschii*. The grey zone between these lines estimates the extent of gene loss in hexaploid wheat.

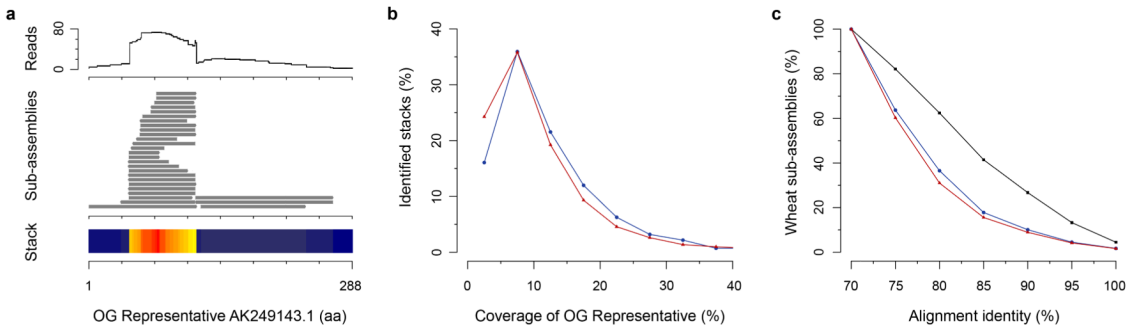

Figure 4. Pseudogene identification and analysis

a. Visualization of an OG Representative and associated wheat sequences. The top track shows the hit count profile of mapped 454 reads. The lower tracks shows sub-assemblies of three wheat genes and a stacked region of gene fragments. Read depth is represented by the heat map.

b. The distribution shows coverage of the OG Representative by Pfam-containing gene fragments and pseudogenes. The blue and red lines represent stacks with and without protein domains, respectively.

c. The distribution shows protein identity between sub-assemblies forming stacks of gene fragments. The blue and red lines represent stacks with and without protein domains, respectively, and the black line represents sub-assemblies forming genes.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Table 1a

Summary of sequence sources used for analysis.

Genome	Platform	Size of dataset	Reference
T. aestivum (CS42) genomic DNA	454 GS FLX Titanium/454 GS FLX+	85Gb	EBI Study: ERP000319
T. aestivum (CS42) genomic DNA from sorted chr 1A, 1B, 1D	454 GS FLX Titanium	1A: 287Mb 1B: 392Mb 1D: 375Mb	Wicker *et al.*, 2011[20]
T. aestivum (CS42, Avalon, Rialto, Savannah) genomic DNA	SOLiD 3/SOLiD 4	15.2 billion reads	EBI Study: ERP001493
T. aestivum (CS42) cDNA	454 GS FLX Titanium/454 GS FLX+	1.6Gb	EBI Study: ERP001415
T. monococcum genomic DNA	Illumina GAIIx/HiSeq	A/B/D sequences: 3.7Gb A/B/D SNPs: 401Gb	NCBI archive: SRP004490.3
Ae. speltoides cDNA	[Pre-assembled data]	151Mb	(M. Trick & I. Bancroft, unpublished)
Ae. tauschii genomic DNA	454 GS FLX Titanium	12.8Gb	Luo *et al.*, 2012, submitted
Ae. tauschii genomic DNA	SOLiD 4	80-100x	(J. Dvorak, unpublished)

Europe PMC Funders Author Manuscripts

Table 1b

Summary of assembly statistics of the Orthologous Assembly (OA), the LowCopy Genome (LCG) and cDNA assemblies.

	OA 99% mi[1]	LCG[2]	cDNA assembly[2]
Number of sequences	949,279	5,321,847	97,481
Total sequence (bp)	437,512,281	3,800,325,216	93,340,842
Minimum / maximum length (bp)	79 / 7,312	100 / 21,721	100 / 10,382
N10 / N50 / N90 (bp)	766 / 481 / 331	2,234 / 884 / 420	2,707 / 1,325 / 509
Mean length (bp)	460.89	714.10	957.53
GC content (%)	48.25	47.69	47.74

[1] combined set of 454 sequences that cluster and form contigs and 454 sequences remaining singletons

[2] only set of 454 sequences that cluster and form contigs