A searchable BLAST database of improved wheat genome sequence assemblies

A new Whole Genome Shotgun (WGS) assembly of the Chinese Spring reference wheat genome is now available for analysis on the Grassroots Genomics BLAST server at The Genome Analysis Centre (TGAC) in Norwich, UK (http://www.tgac.ac.uk/grassroots-genomics/). The new assembly captures over 75% of the 17Gb genome in very large sequence scaffolds (Table 1).

Arm	Total bp	N20	N50	N80	N%	Count
1AL	355,144,189	159,693	80,107	30,798	5.57%	19,140
1AS	200,141,416	176,516	85,799	32,413	5.48%	11,382
1BL	427,850,462	212,050	105,411	41,787	5.43%	19,349
1BS	224,120,373	204,783	99,660	39,287	5.36%	11,813
1DL	292,316,462	127,480	65,923	23,018	6.59%	19,204
1DS	155,677,507	123,950	62,097	19,441	6.74%	12,849
2AL	408,449,610	164,629	84,674	33,270	5.49%	19,410
2AS	318,533,889	183,072	90,023	33,061	5.40%	17,435
2BL	423,469,708	227,122	117,486	45,691	5.14%	16,714
2BS	317,593,121	215,046	108,705	45,716	5.19%	12,136
2DL	335,204,207	133,166	70,105	26,700	6.67%	19,424
2DS	245,159,861	140,704	72,904	24,794	6.56%	16,533
3AL	381,464,830	165,249	84,656	33,372	5.64%	17,063
3AS	277,280,281	188,759	93,882	40,580	5.27%	10,234
3B	789,970,040	223,860	116,546	47,041	5.13%	29,090
3DL	340,636,885	136,140	68,689	24,264	6.53%	22,646
3DS	228,916,862	145,224	72,644	23,143	6.42%	16,817
4AL	363,230,010	179,374	89,157	33,873	5.46%	18,295
4AS	276,247,067	181,019	91,272	35,335	4.98%	14,167
4BL	272,849,020	240,935	127,687	58,815	4.99%	7,632
4BS	310,515,948	224,543	110,746	45,899	4.90%	14,697
4DL	306,806,261	171,404	80,284	28,140	6.31%	18,791
4DS	171,621,745	137,248	68,499	21,787	6.30%	13,021
5AL	413,139,451	161,674	81,944	33,128	5.90%	18,826
5AS	231,190,161	180,634	89,316	35,125	5.14%	11,705
5BL	466,173,773	207,503	107,733	43,825	5.21%	19,325
5BS	182,789,732	209,845	107,461	40,181	5.16%	9,793
5DL	345,449,775	130,074	65,820	23,183	7.02%	23,851
5DS	173,821,965	133,804	64,345	18,898	6.58%	14,481
6AL	302,563,130	168,100	85,773	33,526	5.53%	14,457
6AS	264,274,034	160,498	81,455	30,863	5.68%	14,315
6BL	362,924,849	203,268	110,331	45,402	5.22%	13,913
6BS	299,250,616	185,879	100,360	38,835	5.51%	13,349
6DL	236,649,310	143,791	71,511	24,364	6.34%	16,246
6DS	178,741,401	146,601	65,202	21,073	6.62%	13,586
7AL	334,861,391	184,024	92,381	37,818	5.49%	13,158

7AS	259,954,140	187,229	99,434	47,521	5.56%	7,777
7BL	406,571,657	203,402	107,841	45,705	5.17%	15,233
7BS	287,930,109	222,106	119,366	48,224	4.95%	10,813
7DL	273,279,341	135,861	69,599	23,246	6.84%	18,964
7DS	303,641,845	133,599	68,218	24,284	6.63%	19,510
U	680,947,588	192,507	78,842	6,368	6.58%	88,799
TOTAL	**13,427,354,022**	**180,094**	**88,778**	**32,825**	**5.73%**	**735,943**

Table 1. Summary Assembly Statistics and Chromosome Classifications

Contig assembly

Contigs were assembled from 250bp paired-end reads generated using a PCR-free protocol. TGAC's Algorithms Development Team modified DISCOVAR de novo [1] specifically to cope with large data volumes and to enable it to perform efficient cleaning of the complex wheat genome assembly graph. We used KAT [2] spectra-cn plots to QC motif representation, and tailored our data generation to generate maximum complexity, precisely sized, low bias sampling.

Scaffolding

Multiple Nextera Long Mate Pair libraries were constructed, QC'd, and chosen for sequencing as described in TGAC's new protocol [3], and pre-processed with a pipeline based on Nextclip [4]. Contigs were scaffolded using SOAPdenovo2 [5]. SOAPdenovo2 replaces N-stretches (gaps) in contigs with Cs and Gs during scaffolding, so to correct this contigs were mapped back to the scaffolds and the gaps converted back to Ns.

Contamination screening and filtering

The scaffolds were checked for contamination against the NCBI nucleotide database using BLAST+ and the results were joined to NCBI's taxonomy database. Results were filtered to show hits of >98% identity over >90% of their length. From this list, scaffolds identified with a taxonomy containing "BEP" (the grass BEP clade), "Poales" (the order encompassing grasses) or "eudicotyledons" (the dicot group of angiosperms) were kept and the remaining scaffolds were considered to be contamination. These were mainly short contigs containing PhiX.

Chromosome arm binning

Scaffolds were classified into chromosome-arm bins using arm-specific CSS reads [6]. Scaffolds from 3B were not separated into short/long arm bins as individual arm datasets were not generated for this chromosome in the CSS project. The 'sect' method of KAT was used to compute kmer coverage over each scaffold using each CSS read set. Each non-repetitive kmer in a scaffold was scored proportionally to coverage on each CSS arm and scaffolds were classified using the following set of rules:

1. Scaffolds with less than 10% of the kmers producing a vote were left as *unclassified* (marked as Chromosome arm "U"). These are mostly small and/or repetitive sequences.
2. Scaffolds with a top score towards a CSS set *at least double* the second top score were classified to the highest scoring chromosome arm.
3. Scaffolds with a top score towards a CSS set *less than double* the second top score were left as *unclassified* (marked as Chromosome arm "U", but with the two top scores and CSS sets included in the sequence name). This category contains scaffolds that are classified as combinations of the two arms from the same chromosome, probably due to imprecise identification during flow-sorting. It also contains scaffolds from regions of the genome with specific flow-sorting biases, and assembly chimeras, which will all be investigated further.

Sequence length and content filter

Rather than using a simple length cutoff to include scaffolds in the final assembly, a content filter was applied to the scaffolds classified into each chromosome-arm bin in order to ensure short scaffolds containing unique content were not excluded from the assembly. Scaffolds were sorted by length, longest first. Scaffolds longer than 5Kbp were automatically added to the assembly. Scaffolds between 5Kbp and 500bp were added from longest to smallest if 20% of the kmers in the scaffold were not already present in the assembly. Scaffolds shorter than 500bp were excluded.

Sequence naming

For assigned scaffolds, the arm assignment is included in the FASTA identifier. For unassigned scaffolds with more than 10% voting kmers, the highest and second highest vote is included in the FASTA identifier to indicate possible arms.

Data release policy

This assembly has been deposited at the EMBL-EBI in accordance with assembly versioning, mapping and release policy (http://ensemblgenomes.org/info/data/assemblies). Following assessment and processing as part of this procedure, the assembly will be available in the public domain to aid the wheat community in their work. We estimate that this will be completed by the end of December 2015. TGAC and their collaborators plan to publish a whole-genome analysis paper so please contact us before conducting large-scale analysis. These data are released under the Toronto agreement http://www.nature.com/nature/journal/v461/n7261/full/461168a.html

References

1) http://www.broadinstitute.org/software/discovar/blog/ and Weisenfeld, N. et al. "Comprehensive variation discovery in single human genomes. Nature Genetics **46**, 1350-1355, 2014.

2) http://www.tgac.ac.uk/KAT/

3) D. Heavens, G. G. Accinelli, B. Clavijo, and M. D. Clark, "A method to simultaneously construct up to 12 differently sized Illumina Nextera long mate pair libraries with reduced DNA input, time, and cost.," *BioTechniques*, **59**, 42–45, 2015.

4) R. M. Leggett, B. J. Clavijo, L. Clissold, M. D. Clark, and M. Caccamo, "NextClip: an analysis and read preparation tool for Nextera long mate pair libraries," *Bioinformatics*, p. btt702, 2013.

5) R. Luo, et al "SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler.," *Gigascience*, **1**,18, 2012.

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

Current Work

Researchers at TGAC, John Innes Centre, The Sainsbury Laboratory, the European Bioinformatics Institute and Rothamsted Research are continuing to increase the long-range scaffolding of the WGS assembly and test its accuracy. Several other wheat varieties are also being sequenced. We anticipate releasing improved and newly annotated assemblies early in 2016 for public use.

This work is supported by the UK Biological and Biotechnological Sciences Research Council (BBSRC).

Contact about Grassroots Genomics BLAST server:

If you have any questions regarding access and usage of the server or suggestions for improvements please contact grasshelpdesk@tgac.ac.uk. As academic researchers we especially welcome suggestions from the wheat breeding community.

Contacts about the assembled sequences, methods, data and current work:

Bernardo Clavijo, Algorithms Team Leader, TGAC. Bernardo.Clavijo@tgac.ac.uk
Matt Clark, Co-Principal Investigator, TGAC. Matt.Clark@tgac.ac.uk
Mike Bevan, Co-Principal Investigator, JIC. Michael.Bevan@jic.ac.uk