Wheat Panache - a pangenome graph database representing presence/absence variation across 16 bread wheat genomes

Philipp E. Bayer[1], Jakob Petereit[1], Éloi Durant[2,3,4,5], Cécile Monat[3], Mathieu Rouard[4,5], Haifei Hu[6], Brett Chapman[6], Chengdao Li[6], Shifeng Cheng[7], Jacqueline Batley[1], David Edwards[1*]

[1] School of Biological Sciences, The University of Western Australia, Perth 6009, Australia

[2] DIADE, Univ Montpellier, CIRAD, IRD, Montpellier 34830, France

[3] Syngenta Seeds S.A.S., 12 chemin de l'Hobit, 31790 Saint-Sauveur, France

[4] Bioversity International, Parc Scientifique Agropolis II, Montpellier 34397, France

[5] French Institute of Bioinformatics (IFB)—South Green Bioinformatics Platform, Bioversity, CIRAD, INRAE, IRD, Montpellier 34398, France

[6] Western Crop Genetics Alliance, Murdoch University, 90 South Street, WA6150 Murdoch, Australia

[7] Shenzhen Branch, Guangdong Laboratory of Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture and Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen 518120, China

* To whom correspondence should be addressed: dave.edwards@uwa.edu.au

Abstract

Bread wheat is one of humanity's most important staple crops, characterized by a large and complex genome with a high level of gene presence/absence variation between cultivars, hampering genomic approaches for crop improvement. With the growing global population and the increasing impact of climate change on crop yield, there is an urgent need to apply genomic approaches to accelerate wheat breeding. With recent advances in DNA sequencing technology, a growing number of high-quality reference genomes are becoming available, reflecting the genetic content of a diverse range of cultivars. However, information on the presence or absence of genomic regions has been hard to visualize and interrogate due to the size of these genomes and the lack of suitable bioinformatics tools. To address this limitation, we have produced a wheat pangenome graph maintained within an online database to facilitate interrogation and comparison of wheat cultivar genomes. The database allows users to visualize regions of the pangenome to assess presence/absence variation between bread wheat genomes.

Database URL: http://www.appliedbioinformatics.com.au/wheat_panache

Introduction

Bread wheat (*Triticum aestivum*) is one of the most widely grown crops, yet there is a significant challenge to increase yield to meet the projected demands of a growing world population. With predictions of climate change-related yield losses ranging from 17% to 31% by the middle of the 21st century (1), improved genomics-based breeding approaches are required to produce climate change-ready wheat cultivars.

Wheat genomics has made rapid advances in recent years with the first draft genome assembly produced in 2014 (2) based on the shotgun sequencing of isolated chromosome arms (3-5). A first near-complete assembly of the variety Chinese Spring was produced in 2017 (6) with a final reference genome assembly available in 2018 (7). This reference assembly was rapidly followed by assemblies of fifteen additional cultivars from global breeding programs (8).

The increasing availability of reference genome assemblies made it clear that there is significant presence/absence variation (PAV) between individuals (9-12). This insight has led to the production of pangenomes that reflect the gene content of a species rather than an individual (10,13-20). Pangenomes are now available for several plant species; the first bread what pangenome representing the gene content of 16 bread wheat cultivars was published in 2017 (18). This wheat pangenome was assembled using an iterative mapping approach, which efficiently identified new gene space and called gene presence/absence between individuals. This kind of pangenome is however limited in that the physical location of the new gene space can be difficult to determine with accuracy. With the availability of multiple whole-genome references, this limitation may be addressed through the production of a graph-based pangenome. Graph-based pangenomes have recently become popular thanks to the graph data structure which can accurately represent the physical locations of genomic and structural variants with minimal reference bias, with tools such as vg (21), seqwish (22), minigraph (23), and PHG (24) being successfully applied to build variation, sequence, or haplotype graphs.

A major limitation of pangenome graphs is that few tools are available to visualize these complex graph structures. Genome visualization tools such as GBrowse (25), JBrowse2 (26), or Circos (27) are designed to display information relative to a linear reference genome, not a graph of several genomes, while graph viewers such as Bandage (28) or pangenome viewers such as ODGI (29) focus on visualizing the graph itself, but display little other information such as genome annotations.

Panache is a recent pangenome visualization tool that can process linearized assembly graphs and display shared regions as a web-based dynamic heatmap (30). Panache has so far only been applied to visualize presence/absence variation in the banana pangenome (14), but has the potential to be expanded to other species, even for crop genomes as large as wheat. Here, we present a graph pangenome representing 16 bread wheat cultivars hosted within a public Wheat Panache database, with a new web-based browser for visualizing genomic regions across the wheat pangenome, along

with the graph formatted for minimap2 (31) and Giraffe (32). This tool offers researchers and breeders the ability to assess genome variation between these varieties, mining the diversity present in this large and complex genome.

Results and discussion

A wheat graph pangenome

We constructed a graph pangenome using 16 high-quality wheat genome assemblies Representing the global variation of modern bread wheat cultivars. The assembled graph had a total size of 15.8 Gbp in comparison with the founder genome assembly sizes of 13.9 to 14.2 Gbp (33). After aligning all genomes back to the graph, these 15.8 Gbp were split up into 2,791,482 segments present in at least one individual. The segments had an average size of 5.6 Mbp (median: 498 bp), ranging from 2 bp to 37.6 Mbp (Figure 1A). Realignment of the 16 genome assemblies to the graph revealed that out of the 2.7 million segments, 542,711 (19%) segments were present in all individuals (total size: 10.2 Gbp ranging from 2 bp to 4.9 Mbp) with the remaining 2,248,771 segments (total size: 5.6 Gbp) being present in a median of 8 individuals (Supplementary Figure 1). 10,437 segments (0.4% of all segments) with a total length of 19.9 Mbp (average length: 1.9 Kbp) were not covered by any genome assembly during the realignment step, probably due to these segments being too small and/or too repetitive.

Interestingly, the cultivar with the most unique segments was the reference cultivar Chinese Spring, with 158,503 (7%) of segments with a total size of 140.5 Mbp being only present in Chinese Spring (Figure 1B). This may be due to the genomic distance between Chinese Spring and the other cultivars, consistent with previous observations (18), and reflecting Chinese Spring's age (collected around 1900) and its lack of agronomic characters that were selected for in modern cultivars (34). The distance between the Chinese Spring assembly and the 15 other assemblies is also supported by 1.2 Gbp of the graph in 901,475 segments not being present in Chinese Spring but in at least one other cultivar, reflecting the complex history of introgressions in modern bread wheat (8,35). We aligned

the IWGSC v1 gene annotation for Chinese Spring (7) back to the graph by intersecting the linearized graph with gene positions. We found a position in the graph for 110,790 (100%) genes confirming that the graph assembly contains all gene models of the IWGSC assembly.

The Wheat Panache web portal

Using this graph, we built a web-based Panache instance (30), allowing users to visualize regions or genes of interest for presence/absence across the chosen wheat cultivars. The webserver is available at http://www.appliedbioinformatics.com.au/wheat_panache.

Wheat Panache displays a linear version of the pangenome graph subdivided into blocks based on presence/absence of the selected individuals. A block is defined to have no internal presence/absence variation and to contain at least one gene. Blocks are named based on the pseudomolecule they originated in, and as we started the assembly with the IWGSC assembly, most blocks (1,890,035 out of 2,791,483 blocks, 67%) are named after their position in the IWGSC assembly.

The interface displays the linearized pangenome as a chain of such graph segments, with one horizontal track per cultivar (Figure 2). Coordinates are based on the pangenome graph assembly. Genes are represented as black dots above blocks and hovering over a gene reveals its coordinates within the assembly and exon structure. Three summary tracks below the cultivar tracks show which blocks are core or variable based on a user-definable threshold, how long the block is, and how often the block is repeated within Panache. Users can zoom into blocks, or search for 'hollow areas' (areas of consecutive absence based on a user-defined threshold) using the Hollow Area Finder, which is a convenient way to automatically focus on large PAV areas. Users can sort the cultivars alphanumerically, by gene presence/absence status, or by a phylogeny based on Mash v2.3 (36). The graph assembly displayed in Wheat Panache, including a version pre-indexed for vg v1.37.0's Giraffe (32) is available at https://doi.org/10.5281/zenodo.6085239 (37) allowing for downstream analyses of the population graph.

In summary, we present the first wheat graph pangenome assembly, based on 16 cultivars with an online visual representation of the graph within the Panache visualization tool. The graph assembly will be a valuable tool for wheat genomics researchers looking for a more accurate reference assembly. The web platform Panache allows users to interrogate this graph and search for structural variants around regions of interest.

Materials and Methods

We used publicly available genome assemblies, including fifteen high-quality *Triticum aestivum* genome assemblies (8) and the IWGSC v1 *T. aestivum* cv. Chinese Spring assembly (7), to assemble a graph using minigraph v0.14 (23). To optimize assembly, we used k-mers that appear fewer than 100 times (-f.1) for the graph assembly and assembled the graph genome by genome, starting with IWGSC v1 followed by alphabetical order of cultivar names, ending with the *T. aestivum* ssp. *spelta* PI190962 assembly.

All assemblies were aligned with the final graph using minimap2 v2.18 (31) and alignments were converted to BED format. The main graph was linearized using gfatools gfa2bed v0.4 with default parameters (https://github.com/lh3/gfatools/releases) and merged with all minimap2 alignments using bedtools v2.30.0 multiinter (38). The resulting blocks were intersected with the IWGSC gene annotation using bedtools v2.30.0 intersect.

The data was converted to Panache JSON format and a Panache instance was set up to serve the data (30). To make the display feasible on a regular workstation, we retained only blocks overlapping with Chinese Spring genes and then merged adjacent blocks if they showed identical PAV behaviour across all individuals.

Acknowledgments

This work is funded by the Australia Research Council (Projects DP210100296, DP200100762, and DE210100398). This work was supported by resources provided by the Pawsey Supercomputing Centre with funding from the Australian Government and the Government of Western Australia.

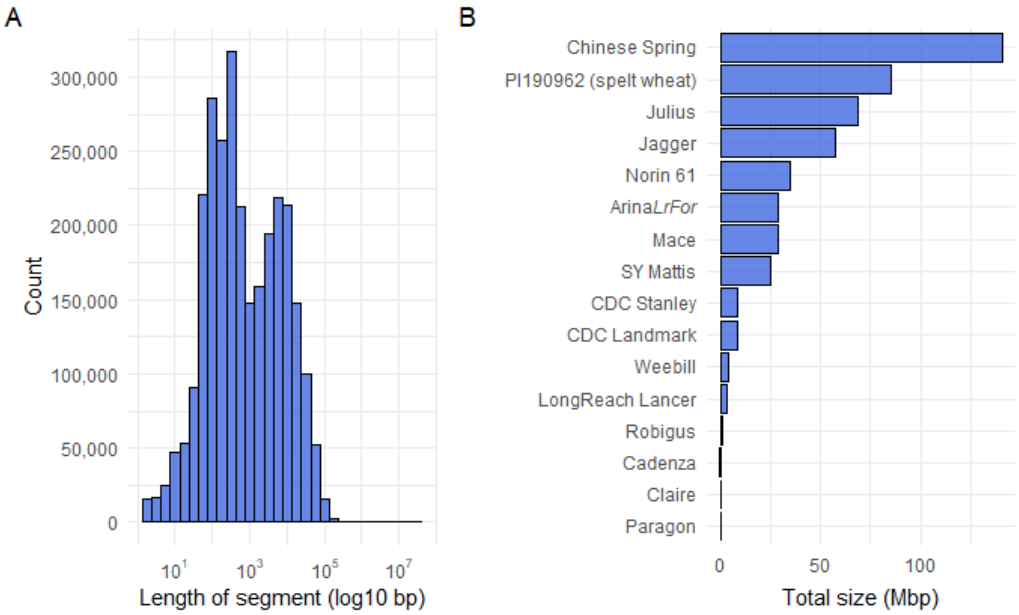
References

1. Obembe, O.S., Hendricks, N.P., Tack, J. (2021) Decreased wheat production in the USA from climate change driven by yield losses rather than crop abandonment. *Plos one*, **16**, e0252067.
2. International Wheat Genome Sequencing, C. (2014) A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. *Science*, **345**.

3. Berkman, P.J., Skarshewski, A., Lorenc, M.T., *et al.* (2011) Sequencing and assembly of low copy and genic regions of isolated Triticum aestivum chromosome arm 7DS. *Plant Biotechnology Journal*, **9**, 768-775.
4. Berkman, P.J., Skarshewski, A., Manoli, S., *et al.* (2012) Sequencing wheat chromosome arm 7BS delimits the 7BS/4AL translocation and reveals homoeologous gene conservation. *Theoretical and applied genetics*, **124**, 423-432.
5. Lai, K., Lorenc, M.T., Lee, H.C., *et al.* (2015) Identification and characterization of more than 4 million intervarietal SNP s across the group 7 chromosomes of bread wheat. *Plant biotechnology journal*, **13**, 97-104.
6. Zimin, A.V., Puiu, D., Hall, R., *et al.* (2017) The first near-complete assembly of the hexaploid bread wheat genome, Triticum aestivum. *Gigascience*, **6**, gix097.
7. Appels, R., Eversole, K., Feuillet, C., *et al.* (2018) Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science*, **361**, eaar7191.
8. Walkowiak, S., Gao, L., Monat, C., *et al.* (2020) Multiple wheat genomes reveal global variation in modern breeding. *Nature*, **588**, 277-283.
9. Hurgobin, B., Edwards, D. (2017) SNP discovery using a pangenome: has the single reference approach become obsolete? *Biology*, **6**, 21.
10. Golicz, A.A., Bayer, P.E., Barker, G.C., *et al.* (2016) The pangenome of an agronomically important crop plant Brassica oleracea. *Nature Communications*, **7**, 13390.
11. Golicz, A.A., Bayer, P.E., Bhalla, P.L., *et al.* (2020) Pangenomics comes of age: From bacteria to plant and animal applications. *Trends in Genetics*, **36**, 132-145.
12. Bayer, P.E., Golicz, A.A., Scheben, A., *et al.* (2020) Plant pan-genomes are the new reference. *Nat. Plants*, **6**, 914-920.
13. Bayer, P.E., Scheben, A., Golicz, A.A., *et al.* (2021) Modelling of gene loss propensity in the pangenomes of three Brassica species suggests different mechanisms between polyploids and diploids. *Plant Biotechnology Journal*, **n/a**.
14. Rijzaani, H., Bayer, P.E., Rouard, M., *et al.* (2021) The pangenome of banana highlights differences between genera and genomes. *The Plant Genome*, **n/a**, e20100.
15. Zhao, J., Bayer, P.E., Ruperao, P., *et al.* (2020) Trait associations in the pangenome of pigeon pea (*Cajanus cajan*). *Plant Biotechnol Journal*.
16. Jensen, S.E., Charles, J.R., Muleta, K., *et al.* (2020) A sorghum practical haplotype graph facilitates genome-wide imputation and cost-effective genomic prediction. *The Plant Genome*, **13**, e20009.
17. Franco, J.A.V., Gage, J.L., Bradbury, P.J., *et al.* (2020) A Maize Practical Haplotype Graph Leverages Diverse NAM Assemblies. *bioRxiv*, 2020.2008.2031.268425.
18. Montenegro, J.D., Golicz, A.A., Bayer, P.E., *et al.* (2017) The pangenome of hexaploid bread wheat. *The Plant Journal*, **90**, 1007-1013.
19. Ruperao, P., Thirunavukkarasu, N., Gandham, P., *et al.* (2021) Sorghum Pan-Genome Explores the Functional Utility for Genomic-Assisted Breeding to Accelerate the Genetic Gain. *Frontiers in plant science*, **12**, 963.
20. Song, J.-M., Guan, Z., Hu, J., *et al.* (2020) Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of *Brassica napus*. *Nature Plants*, 1-12.
21. Hickey, G., Heller, D., Monlong, J., *et al.* (2020) Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome biology*, **21**, 1-17.
22. Garrison, E., Guarracino, A. (2022) Unbiased pangenome graphs. *bioRxiv*, 2022.2002.2014.480413.
23. Li, H., Feng, X., Chu, C. (2020) The design and construction of reference pangenome graphs with minigraph. *Genome Biology*, **21**, 265.
24. Jensen, S.E., Charles, J.R., Muleta, K., *et al.* (2020) A sorghum practical haplotype graph facilitates genome-wide imputation and cost-effective genomic prediction. *The Plant Genome*, **n/a**, e20009.

25. Donlin, M.J. (2009) Using the generic genome browser (GBrowse). *Current protocols in bioinformatics*, **28**, 9.9. 1-9.9. 25.
26. Buels, R., Yao, E., Diesh, C.M., *et al.* (2016) JBrowse: a dynamic web platform for genome visualization and analysis. *Genome biology*, **17**, 66.
27. Krzywinski, M., Schein, J., Birol, I., *et al.* (2009) Circos: an information aesthetic for comparative genomics. *Genome research*, **19**, 1639-1645.
28. Wick, R.R., Schultz, M.B., Zobel, J., *et al.* (2015) Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics*, **31**, 3350-3352.
29. Guarracino, A., Heumos, S., Nahnsen, S., *et al.* (2021) ODGI: understanding pangenome graphs. *bioRxiv*.
30. Durant, É., Sabot, F., Conte, M., *et al.* (2021) Panache: a web browser-based viewer for linearized pangenomes. *Bioinformatics*, **37**, 4556-4558.
31. Li, H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**, 3094-3100.
32. Jouni, S., Jean, M., Xian, C., *et al.* (2021) Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science*, **374**, abg8871.
33. Walkowiak, S., Gao, L., Monat, C., *et al.* (2020) Multiple wheat genomes reveal global variation in modern breeding. *Nature*.
34. Sears, E., Miller, T. (1985) The history of Chinese Spring wheat. *Cereal Research Communication*, 261-263.
35. Keilwagen, J., Lehnert, H., Berner, T., *et al.* (2021) Detecting Major Introgressions in Wheat and their Putative Origins Using Coverage Analysis.
36. Ondov, B.D., Treangen, T.J., Melsted, P., *et al.* (2016) Mash: fast genome and metagenome distance estimation using MinHash. *Genome biology*, **17**, 1-14.
37. Bayer, P.E., Petereit, J., Durant, E., *et al.* (2022) Bread wheat genomes graph pangenome. Zenodo.
38. Quinlan, A.R., Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841-842.
39. Gao, L., Koo, D.-H., Juliana, P., *et al.* (2021) The Aegilops ventricosa 2NvS segment in bread wheat: cytology, genomics and breeding. *Theoretical and Applied Genetics*, **134**, 529-542.
40. Keilwagen, J., Lehnert, H., Berner, T., *et al.* (2022) Detecting major introgressions in wheat and their putative origins using coverage analysis. *Scientific Reports*, **12**, 1908.

234 **Figures**



235

236 Figure 1: A) Bar chart showing the distribution of the size of all assembly graph segments (log-scale).

237 B) Total size of unique segments per cultivar in Mbp. PI190962 is a line of species *Triticum spelta,*

238 Chinese Spring is the reference cultivar of *T. aestivum*.

239

240

241

242 Figure 2 – Wheat Panache screenshot showing a *Aegilops ventricosa* introgression at the beginning
243 of chromosome 2 in cultivars Stanley, Jagger, Mace, and SY Mattis (39,40). Black boxes were added
244 to show the region missing in cultivars where the introgression replaced parts of 2A. The graph
245 assembly started with the IWGSC v1 assembly leading to linearized regions following the same
246 naming scheme as the IWGSC v1.0 assembly (chr1A_part1, chr1A_part2, chr2A_part1, …). CS stands
247 for Chinese Spring. Shown here is the beginning of the first part of chr2A. Black blocks are gene
248 models. White regions correspond to regions that are present in the graph but contain no genes.

249