Reference Genome Sequencing Of The Elite Bread Wheat Cultivar, “Sonmez”

B. Ani Akpinar
Montana BioAg. Inc.
Philippe Leroy
INRAE USC1338: Centre d'Ecologie Fonctionnelle et Evolutive
Nathan S. Watson-Haigh
Australian Genome Research Facility
Ute Baumann
The University of Adelaide School of Agriculture Food and Wine
Valerie Barbe
Genescope
Hikmet Budak (hikmet.budak@icloud.com)
Montana BioAg. Inc. https://orcid.org/0000-0002-2556-2478

Research Article

Keywords: Wheat, genome sequencing, Triticum aestivum, yield, Sonmez

Posted Date: February 25th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1095548/v1

Abstract

Objectives

High-yielding crop varieties will become critical in meeting the future food demand in the face of worsening weather extremes and threatening biotic stressors. Bread wheat cultivar Sonmez-2001 is a registered variety that is notable for its performance under low irrigation conditions, which further improves upon irrigation. In order to explore and exploit the genomic potential of this elite variety, in 2015, we sequenced the modern elite cultivar Sonmez genome which represents a diverse genetic background.

Data description

Here, we provide a reference-guided whole genome sequence of Sonmez-2001, assembled into 21 chromosomes of the A, B and D genomes and totaling to 13.3 gigabase-pairs in length. Additionally, a *de novo* assembly, of an additional 1.05 gigabase-pairs, was generated that represents either Sonmez-specific sequences or sequences that considerably diverged between Sonmez and Chinese Spring. We identified up to 24 million sequence variants, of which up to 2.4% reside in coding sequences, that can be used to develop molecular markers that should be of immediate use to the cereal community.

Objective

Triticum aestivum cv. Sonmez-2001 *(Sonmez, hereafter)* is a registered, elite bread wheat variety that has been bred particularly for drylands. Accordingly, Sonmez exhibits remarkable tolerance against drought and performs considerably better than its ancestor, Bezostaya-1, in terms of yield, stress tolerance and disease resistance. Sonmez variety is notable for high yield and grain quality of ~15% protein content under rain-fed conditions, both of which further improve with supplemental irrigation. Sonmez is also highly resistant against causal agents of devastating diseases, in particular, against cereal cyst nematode and yellow rust. Sonmez has superior resistance against soil-borne pathogens, and exhibit good tolerance against diseases affecting leaves and inflorescence. Due to these attributes, Sonmez is the cultivar of choice for most of the Central Anatolian Plateau. Facing a fast-growing world population, estimated to reach 9 billion people in the next three decades, and changing climate trends with harmful effects on agriculture, securing the food demand of upcoming generations will require extensive improvements in crop yields. With cereals being the staple food for the developing world, Sonmez is a promising candidate that can contribute to meeting this demand. Here, we report a reference sequence of the Sonmez genome, and its comparative analysis with *Triticum aestivum* genotype Chinese Spring, for which extensive data, including a high quality reference genome sequence, is available.

Data Description

A paired-end (PE) library with an insert size of 350 base-pair was produced and sequenced on Illumina HiSeq 4000 platform at Genoscope - National Center of Sequencing, (Évry-Courcouronnes, France). The 970.6 gigabase-pair (Gbp) of PE reads passing quality filters were mapped against the *Triticum aestivum* Chinese Spring (CS) RefSeq v1.0 genome [1] in a 2-step approach. In the first step, an ungapped alignment was performed using BioKanga v3.4.5 using default parameters but allowing for 2 mismatches per 100 bp. In the second step, the unmapped reads were mapped with Bowtie2 v2.3.0 [2], allowing a single indel of length ≤ 9 bp with zero mismatches. Read alignments from both mapping steps were merged using Sambamba v0.6.5 [3]. Regions containing read alignments with indels were identified and re-aligned using GATK v3.7 using default parameters with minor modifications (Data file 1) [4].

Sequence variations, including Single Nucleotide Polymorphisms (SNPs) and indels, were called by BCFtools v1.3.1 on pileups generated by SAMtools v1.3.1 [5]. Strict homozygous SNP and indel variants were identified using GATK's SelectVariants to retain only variants with no support for the CS reference allele. These homozygous variants were analysed by SNPeff v4.3i [6] to estimate their impact in the context of the CS RefSeq v1.0 High Confidence gene annotations, excluding intergenic regions. Using all identified homozygous variants, we recalled the reference to generate a "Sonmez reference sequence v1.0". Where there was no coverage of the CS reference, we softmasked the Sonmez genome sequence. It should be noted that these softmasked bases could represent regions which are either deletions in Sonmez or insertions in CS. Finally, the read pairs that remained unmapped following the 2-step alignment approach were *de novo* assembled to uncover Sonmez-specific genomic contigs. Sequences that may arise from contaminants were filtered out (Data file 1) [4]. These *de novo* assembled sequences are referred as "Sonmez-specific contigs" hereafter.

In total, 13.3 Gbp (91.51%) of the 14.5 Gbp CS reference genome assembly were covered by Sonmez reads, with a mean depth of coverage of ~50x. We identified between 3.15 – 23.96 million variants, depending on coverage threshold used, of which between 0.03 – 3.23% were indel variants (Data file 2, 3) [7, 8]. We found that 1.47 – 2.39% of all variants fell within the RefSeq v1.0 High Confidence gene annotations (Data file 2) [7]. Of these, approx. 40% fell within coding regions. Of the homozygous variants supported by ≥ 5 reads, we observed approximately 1 variant per 500 bp in the A and B genomes and approximately 1 variant per 4,000 bp in the D genome.

The Sonmez-specific contigs totaled 1.05 Gbp in length, the longest being 15,887 bp and having N50 and N90 of 427 bp and 269 bp, respectively. These contigs represent regions of the Sonmez genome which are either unique to Sonmez (e.g. introgressions) or significantly divergent compared to CS. An updated version (v5.3p01) of the TriAnnot pipeline [9] optimized for wheat was used to generate similarity-based and *ab initio* gene models and annotate repetitive elements on contigs that are longer than 10 kilobases (Data file 4) [10]. Overall, 28 Sonmez-specific contigs contained 35 gene models, of which 11 were high-confidence.

The reference genome of Sonmez provides the first, in-depth look at the genome of an elite bread wheat variety that is particularly notable for its performance under low-irrigation conditions. With this genome sequence, the long-awaited Chinese Spring whole genome sequence will finally be available for immediate use in breeding, whereby the two genomes will not only unravel the ultimate content and organization of the bread wheat genome, but also generate a pool of sequence variations that can be utilized to design and implement molecular markers.

Table 1: Overview of data files/data sets

Label	Name of data file/data set	File types (file extension)	Data repository and identifier (DOI or accession number)
Data file 1	*Assembly and variant identification details for Sonmez*	*Portable Document Format file (.pdf)*	https://doi.org/10.6084/m9.figshare.16992229 [4]
Data file 2	*Summary information of sequence variants between Sonmez and CS*	*Portable Document Format file (.pdf)*	https://doi.org/10.6084/m9.figshare.16992322 [7]
Data file 3	*Homozygous SNP/indel variants identified between Sonmez and CS*	*Variant Call Format file (.vcf.gz)*	https://doi.org/10.6084/m9.figshare.16992388 [8]
Data file 4	*Gene models and repeat annotations of Sonmez-specific contigs*	*XML Spreadsheet file (.xlsx)*	https://doi.org/10.6084/m9.figshare.16992337 [10]
Data set 1	*Sonmez reference genome sequence v1.0*	*Fasta file (.fasta.gz)*	https://urgi.versailles.inra.fr/download/iwgsc/Turkish_elite_cultivar/ [11]
Data set 2	*Sonmez-specific contigs*	*Fasta file (.fasta.gz)*	https://urgi.versailles.inra.fr/download/iwgsc/Turkish_elite_cultivar/ [12]

Limitations

Since the generation of the Sonmez-2001 v1.0 genome assembly, updates on the genome and genome annotation for the reference genotype Chinese Spring was released (Latest version v2.1; https://wheat-urgi.versailles.inra.fr/Seq-Repository/Assemblies). An updated genome sequence for Sonmez-2001 based on the latest release of the reference genome, in the future, may contain small differences to the v1.0 genome published here.

Abbreviations

CS: Chinese Spring; Gbp: Gigabase-pair; PE: Paired-End; SNP: Single Nucleotide Polymorphism

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

Data files 1–4 described in this data note are publicly available on Figshare (https://figshare.com/) [4, 7, 8, 10]. The Sonmez reference genome sequence v1.0 (Data set 1) and Sonmez-specific contigs (Data set 2) are deposited at URGI (https://urgi.versailles.inra.fr/download/iwgsc/Turkish_elite_cultivar/) [11, 12]. Both data sets are freely available upon the publication of this manuscript. Table 1 lists the details and information on the access to these data files/sets.

Competing interests

The authors declare no competing interests.

Funding

This research was supported by Budak Family Foundation (BFF) in 2014.

Authors ’contributions

HB conceived the idea of the study and supervised all experiments and computational analyses. BAA combined all analyses and drafted the manuscript. VB carried out sequencing experiments. NSWH and UB generated the Sonmez genome assemblies and identified sequence variants. PL performed TriAnnot analyses optimized for wheat.

Acknowledgements

We acknowledge BFF for supporting science for 20 years. Its advocacy for unwavering belief, has been invaluable in integrating and transferring data to knowledge.

References

[1] IWGSC, Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science. 2018; 361(6403). https://doi.org/10.1126/science.aar7191

[2] Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012; 9: 357-359. https://doi.org/10.1038/nmeth.1923.

[3] Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: fast processing of NGS alignment formats. Bioinformatics. 2015; 31(12): 2032–2034. https://doi.org/10.1093/bioinformatics/btv098

[4] Data file 1: Assembly and variant identification details for Sonmez. Figshare: https://doi.org/10.6084/m9.figshare.16992229 (2021)

[5] Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H. Twelve years of SAMtools and BCFtools. GigaScience. 2021; 10(2): giab008. https://doi.org/10.1093/gigascience/giab008

[6] Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly. 2012; 6(2): 80-92. https://doi.org/10.4161/fly.19695

[7] Data file 2: Summary information of sequence variants between Sonmez and CS. Figshare: https://doi.org/10.6084/m9.figshare.16992322 (2021)

[8] Data file 3: Homozygous SNP/indel variants identified between Sonmez and CS. Figshare: https://doi.org/10.6084/m9.figshare.16992388 (2021)

[9] Leroy P, Guilhot N, Sakai H, Bernard A, Choulet F, Theil S, Reboux S, Amano N, Flutre T, Pelegrin C, Ohyanagi H, Seidel M, Giacomoni F, Reichstadt M, Alaux M, Gicquello E, Legeai F, Cerutti L, Numa H, Tanaka T, Mayer K, Itoh T, Quesneville H, Feuillet C. TriAnnot: A Versatile and High Performance Pipeline for the Automated Annotation of Plant Genomes. Frontiers in Plant Science. 2012; 3:5. https://doi.org/10.3389/fpls.2012.00005.

[10] Data file 4: Gene models and repeat annotations of Sonmez-specific contigs. https://doi.org/10.6084/m9.figshare.16992337 (2021)

[11] Data set 1: Sonmez reference genome sequence v1.0. URGI. https://urgi.versailles.inra.fr/download/iwgsc/Turkish_elite_cultivar/ (2021)

[12] Data set 2: Sonmez-specific contigs. URGI. https://urgi.versailles.inra.fr/download/iwgsc/Turkish_elite_cultivar/ (2021)

Datafile

Datafile3 is not available with this version

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SonmezDatafile1.pdf
- SonmezDatafile2.pdf
- SonmezDatafile4.xlsx