



**HAL**  
open science

## The complete genome sequence of elite bread wheat cultivar, “Sonmez”

Bala Ani Akpinar, Philippe Leroy, Nathan Watson-Haigh, Ute Baumann, Valérie Barbe, Hikmet Budak

► **To cite this version:**

Bala Ani Akpinar, Philippe Leroy, Nathan Watson-Haigh, Ute Baumann, Valérie Barbe, et al.. The complete genome sequence of elite bread wheat cultivar, “Sonmez”. F1000Research, 2022, 11, pp.614. 10.12688/f1000research.121637.1 . hal-03710409

**HAL Id: hal-03710409**

**<https://hal.inrae.fr/hal-03710409v1>**

Submitted on 30 Jun 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.





Distributed under a Creative Commons Attribution 4.0 International License



## GENOME NOTE

# The complete genome sequence of elite bread wheat cultivar, “Sonmez” [version 1; peer review: 3 approved]

Bala Ani Akpinar<sup>1</sup>, Philippe Leroy<sup>2</sup>, Nathan Watson-Haigh<sup>3-5</sup>, Ute Baumann<sup>3</sup>, Valerie Barbe <sup>6</sup>, Hikmet Budak <sup>1</sup>

<sup>1</sup>Genomics and Genome Editing, Montana BioAg. Inc, Missoula, Montana, 59802, USA

<sup>2</sup>Diversité et Écophysologie des Céréales, INRAE-UCA, UMR 1095, Génétique,, Clermont-Ferrand, 10951095, France

<sup>3</sup>School of Agriculture, Food and Wine, University of Adelaide, Plant Genomics Centre, Waite Campus, Hartley Grove, Urrbrae SA 5064, Australia, Adelaide, Australia

<sup>4</sup>South Australian Genomics Centre, SAHMRI, North Terrace, Adelaide SA 5000, Australia (NSWH), Adelaide, Australia

<sup>5</sup>Australian Genome Research Facility, Victorian Comprehensive Cancer Centre, Melbourne, VIC 3000, Australia

<sup>6</sup>Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Evry, France

**V1** First published: 06 Jun 2022, 11:614  
<https://doi.org/10.12688/f1000research.121637.1>

Latest published: 06 Jun 2022, 11:614  
<https://doi.org/10.12688/f1000research.121637.1>

## Abstract

High-yielding crop varieties will become critical in meeting the future food demand in the face of worsening weather extremes and threatening biotic stressors. The bread wheat cultivar Sonmez-2001 is a registered variety that is notable for its performance under low-irrigation conditions, which further improves upon irrigation. Additionally, Sonmez-2001 is resilient against certain biotic stressors, particularly soil-borne pathogens.

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 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. Within this *de novo* assembly, we identified 35 gene models, of which 11 were high-confidence, that may contribute to the favorable traits of this high-performing variety. 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.

## Keywords




Wheat, genome sequencing, Triticum aestivum, yield, Sonmez




This article is included in the [Agriculture, Food and Nutrition](#) gateway.

## Open Peer Review

Approval Status 

	1	2	3
<b>version 1</b>			
06 Jun 2022	<a href="#">view</a>	<a href="#">view</a>	<a href="#">view</a>

1. **Søren K. Rasmussen** , University of Copenhagen, Frederiksberg C, Denmark

2. **Zahide Neslihan Öztürk Gökçe**, Niğde Ömer Halisdemir University, Niğde, Turkey

3. **Gabriel Doredo Perez**, Universidad de Córdoba, Córdoba, Spain

Any reports and responses or comments on the article can be found at the end of the article.



This article is included in the **Genome Sequencing gateway**.

**Corresponding author:** Hikmet Budak ([hikmet.budak@icloud.com](mailto:hikmet.budak@icloud.com))

**Author roles:** **Akpinar BA:** Data Curation, Formal Analysis, Writing – Original Draft Preparation, Writing – Review & Editing; **Leroy P:** Data Curation, Writing – Review & Editing; **Watson-Haigh N:** Data Curation, Formal Analysis; **Baumann U:** Data Curation, Formal Analysis; **Barbe V:** Formal Analysis; **Budak H:** Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Methodology, Project Administration, Resources, Supervision, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

**Competing interests:** H. Budak and A. Akpinar are employed by Montana BioAgriculture Inc. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Grant information:** This research was supported by the Budak Family Foundation (BFF) in 2014.

*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

**Copyright:** © 2022 Akpinar BA *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**How to cite this article:** Akpinar BA, Leroy P, Watson-Haigh N *et al.* **The complete genome sequence of elite bread wheat cultivar, “Sonmez” [version 1; peer review: 3 approved]** F1000Research 2022, **11**:614 <https://doi.org/10.12688/f1000research.121637.1>

**First published:** 06 Jun 2022, **11**:614 <https://doi.org/10.12688/f1000research.121637.1>

## Introduction

*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, building up to  $\approx 15\%$  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, 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 over 9 billion people in the next three decades, and changing climate trends with destructive 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-guided sequence of the Sonmez genome, and its comparative analysis with the reference species, *Triticum aestivum* genotype Chinese Spring, for which extensive data, including a high-quality genome sequence, is available.

## Methods

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), generating almost 3.3 billion  $2 \times 150$  bp reads. The 970.6 gigabase-pair (Gbp) of PE reads passing quality filters were mapped against the *T. aestivum* Chinese Spring (CS) RefSeq v1.0 genome<sup>1</sup> in a two-step approach. In the first step, an ungapped alignment was performed using BioKanga v3.4.5 using default parameters but allowing for two mismatches per 100 bp (--substitutions=2). In the second step, the unmapped reads were mapped with Bowtie2 v2.3.0,<sup>2</sup> allowing a single insertion/deletion of length  $\leq 9$  bp with zero mismatches (--very-sensitive --ignorequals --mp 999,999 --np 999 --rdg 10,1 --rfg 10,1 --score-min L,-19,0 --n-ceil L,0,0). Read alignments from both mapping steps were merged using Sambamba v0.6.5.<sup>3</sup> Regions containing read alignments with insertions/deletions were identified and re-aligned using GATK v3.7 using default parameters with minor modifications (LODThresholdForCleaning=0.4 defaultBaseQualities=30).

Sequence variations, including single nucleotide polymorphisms (SNPs) and insertion/deletion polymorphisms (indels) were called by BCFtools v1.3.1 on pileups generated by SAMtools v1.3.1.<sup>4</sup> Homozygous SNP and indel variants were identified using GATK's SelectVariants to retain only variants with no support for the CS reference allele at a series of read depth thresholds (1, 5, 10, 20, 30 and 40). BEDTools v2.26.0 intersect tool was used to identify intersects between gene annotation coordinate ranges and the identified variants. Homozygous variants were analysed by SNPeff v4.3i<sup>5</sup> to estimate their impact in the context of the CS RefSeq v1.0 High Confidence gene annotations, excluding intergenic regions (-no-intergenic). Using all identified homozygous variants, we recalled the reference to generate a "Sonmez genome 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 two-step alignment approach were assembled *de novo* to uncover Sonmez-specific genomic contigs. k-mers of length 71 bp and occurring  $\geq 9$  times in the unmapped reads were extracted using KMC v3.0.1.<sup>6</sup> These extracted k-mers were assembled into contigs using merutils v0.7.15 kxend command; contigs  $< 250$  bp in length were filtered out. This assembly approach ensures that contig extension only occurs if there is an unambiguous 1 bp extension possible in the input k-mer data set. *Methylobacterium* are well documented, common contaminants of reagents used in Illumina sequencing. As such, contigs showing high sequence identity to one of several *Methylobacterium* genomes (NZ\_CP006992.1, NC\_010511.1, NZ\_CP017640.1, CP001029.1, AP014813.1, AP014810.1) or phiX (NC\_001422.1) were also filtered out. These *de novo* assembled sequences are referred as "Sonmez-specific contigs" hereafter.

## Results

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  $\approx 50\times$ , enabling an almost complete, first construction of the Sonmez genome. Additionally, sequences that are either unique to Sonmez (*e.g.* introgressions) or significantly divergent compared to CS were used to build up a *de novo* assembly. This assembly totaled 1.05 Gbp in length, with the longest contig being 15,887 bp (N50=427 bp, N90=269 bp). An updated version (v5.3p01) of the TriAnnot pipeline<sup>7</sup> 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. While the *de novo* assembly was highly fragmented, compared to the recalled Sonmez genome, we were still able to pick up 35 gene models, of which 11 were high-confidence (*Extended data*<sup>8</sup>).

We identified between 3.15 – 23.96 million variants, depending on the coverage threshold used, of which between 0.03 – 3.23% were indel variants (*Extended data*<sup>9,10</sup>). We found that 1.47 – 2.39% of all variants fell within the RefSeq v1.0 High Confidence gene annotations (*Extended data*<sup>9</sup>). Of these, approx. 40% fell within coding regions. Of the homozygous variants supported by  $\geq 5$  reads, we observed approximately one variant per 500 bp in the A and B genomes and approximately one variant per 4,000 bp in the D genome.

Here, we present the complete genome of the elite wheat variety Sonmez, notable for its performance under low-irrigation conditions. In the face of climatic extremes and other factors that challenge the food safety of upcoming generations, genome sequences of multiple genotypes, varieties and close relatives will not only help us understand complex traits, such as yield and stress responses, but also enable us to efficiently explore the genetic diversity within germplasms for favorable genotypes and/or traits for crop improvement through the use of molecular tools.

## Data availability

### Underlying data

Sonmez complete genome sequence v1.0 and *de novo* assembly are available from the dedicated [URGI database](#).

### Extended data

Figshare: Sonmez\_Extended\_Data1, <https://doi.org/10.6084/m9.figshare.16992337>.<sup>8</sup>

This project contains the following extended data:

- Extended\_data1\_Sonmez\_TriAnnotAnalysis\_v1.xlsx (Gene models and repeat annotations of Sonmez-specific contigs)

Figshare: Sonmez\_Extended\_Data2, <https://doi.org/10.6084/m9.figshare.16992322.v3>.<sup>9</sup>

This project contains the following extended data:

- Extended\_data2\_Sonmez\_vs\_CS\_variantssummary\_v1.pdf (Summary information of sequence variants between Sonmez and CS)

Figshare: Sonmez\_Extended\_Data3, <https://doi.org/10.6084/m9.figshare.16992388.v2>.<sup>10</sup>

This project contains the following extended data:

- Sonmez.alt\_fasta.vcf. (Homozygous SNP/indel variants identified between Sonmez and CS)

Data are available under the terms of the [Creative Commons Attribution 4.0 International license](#) (CC-BY 4.0).

## Acknowledgements

We acknowledge BFF for supporting science for 20 years. Their 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). [PubMed Abstract](#) | [Publisher Full Text](#)
2. Langmead B, Salzberg S: **Fast gapped-read alignment with Bowtie 2**. *Nat. Methods* 2012; **9**: 357–359. [PubMed Abstract](#) | [Publisher Full Text](#)
3. Tarasov A, Vilella AJ, Cuppen E, *et al.*: **Sambamba: fast processing of NGS alignment formats**. *Bioinformatics* 2015; **31**(12): 2032–2034. [PubMed Abstract](#) | [Publisher Full Text](#)
4. Danecek P, Bonfield JK, Liddle J, *et al.*: **Twelve years of SAMtools and BCFtools**. *GigaScience*. 2021; **10**(2): gjab008. [PubMed Abstract](#) | [Publisher Full Text](#)
5. Cingolani P, Platts A, Wang le L, *et al.*: **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. [PubMed Abstract](#) | [Publisher Full Text](#)
6. Kokot M, Dlugosz M, Deorowicz S: **KMC 3: counting and manipulating k-mer statistics**. *Bioinformatics* 2017; **33**(17): 2759–2761. [PubMed Abstract](#) | [Publisher Full Text](#)
7. Leroy P, Guilhot N, Sakai H, *et al.*: **TriAnnot: A Versatile and High Performance Pipeline for the Automated Annotation of Plant Genomes**. *Front. Plant Sci.* 2012; **3**: 5. [PubMed Abstract](#) | [Publisher Full Text](#)

8. **Extended Data 1: Gene models and repeat annotations of Sonmez-specific contigs.**  
[Publisher Full Text](#)
9. **Extended Data 2: Summary information of sequence variants between Sonmez and CS. Figshare.**  
[Publisher Full Text](#)
10. **Extended Data 3: Homozygous SNP/indel variants identified between Sonmez and CS. Figshare.**  
[Publisher Full Text](#)

# Open Peer Review

Current Peer Review Status:   

---

Version 1

Reviewer Report 24 June 2022

<https://doi.org/10.5256/f1000research.133524.r140015>

© 2022 Perez G. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



## Gabriel Doredo Perez

Dep. Bioquímica y Biología Molecular, Campus Rabanales C6-1-E17, Campus de Excelencia Internacional Agroalimentario (ceiA3), Universidad de Córdoba, Córdoba, Spain

The manuscript "The complete genome sequence of elite bread wheat cultivar, "Sonmez", published on F1000Research by Akpinar et al, is an interesting work on structural genomics, covering both re-sequencing of 21 chromosomes of the A, B and D genomes (13.3 Gbp), as well as de novo sequencing (1.05 Gbp). Such latter interesting result represents either Sonmez-specific sequences, or sequences that considerably diverged between Sonmez and Chinese Spring (used as reference genome). The de novo assembly identified 35 gene models, of which 11 were annotated with high-confidence, that may contribute to the favorable traits of this variety. A total of 24 million sequence variants were identified, of which up to 2.4% reside in coding sequences.

Interestingly, the Sonmez cultivar is resilient against certain biotic stressors, particularly soil-borne pathogens. Besides, it is notable for its performance under low-irrigation conditions. Therefore, this work is particularly relevant in the current trend of global warming and climate change, including worldwide drought. Comparison of wheat genomes will allow to decipher complex traits, like abiotic and biotic stresses. That should allow the development of molecular markers for wheat breeding. These developments will help to address future food demand for a growing worldwide population.

**Are the rationale for sequencing the genome and the species significance clearly described?**

Yes

**Are the protocols appropriate and is the work technically sound?**

Yes

**Are sufficient details of the sequencing and extraction, software used, and materials provided to allow replication by others?**

Yes

**Are the datasets clearly presented in a usable and accessible format, and the assembly and annotation available in an appropriate subject-specific repository?**

Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Molecular biology.

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Reviewer Report 13 June 2022

<https://doi.org/10.5256/f1000research.133524.r140013>

© 2022 Öztürk Gökçe Z. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Zahide Neslihan Öztürk Gökçe**

Ayhan Şahenk Faculty of Agricultural Sciences and Technologies, Department of Agricultural Genetic Engineering, Niğde Ömer Halisdemir University, Niğde, Turkey

In the article, the authors performed reference-guided whole genome sequencing of an elite wheat cultivar Sonmez having high yield under stress conditions and specified with resistance against some biotic factors. The data are well presented and the information provided will be of great use for scientific community in the development of abiotic stress resilient wheat cultivars. Therefore my recommendation is to be accepted for indexing as it is.

Best regards

**Are the rationale for sequencing the genome and the species significance clearly described?**

Yes

**Are the protocols appropriate and is the work technically sound?**

Yes

**Are sufficient details of the sequencing and extraction, software used, and materials provided to allow replication by others?**

Yes

**Are the datasets clearly presented in a usable and accessible format, and the assembly and annotation available in an appropriate subject-specific repository?**

Yes

**Competing Interests:** No competing interests were disclosed.



**Reviewer Expertise:** Transcriptomics of abiotic stress tolerance

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Reviewer Report 13 June 2022

<https://doi.org/10.5256/f1000research.133524.r140014>

© 2022 Rasmussen S. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Søren K. Rasmussen** 

Department of Plant and Environmental Sciences, University of Copenhagen, Frederiksberg C, Denmark

A draft genome sequence of the hexaploidy wheat 'Sonmez', a Turkish wheat bread cultivar, is presented. A large number of sequence variants are identified, and as expected the highest density of these putative SNP markers are located on the A- and B-genome and much lower number on the D-genome. This can facilitate efficient markers-assisted selection taking advantage of Sonmez drought tolerance as emphasized.

**Are the rationale for sequencing the genome and the species significance clearly described?**

Yes

**Are the protocols appropriate and is the work technically sound?**

Yes

**Are sufficient details of the sequencing and extraction, software used, and materials provided to allow replication by others?**

Yes

**Are the datasets clearly presented in a usable and accessible format, and the assembly and annotation available in an appropriate subject-specific repository?**

Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Molecular plant breeding, quality traits, plant genetic resources, grain cereals and legumes,

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

---

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact [research@f1000.com](mailto:research@f1000.com)

**F1000Research**