**Credits:**
Hikmet Budak, Ani Akpinar, Sezgi Bıyıklıoğlu, Valérie Barbe
Montana BioAg.Inc (USA) and Genoscope (France)

### *Library generation and sequencing*

A paired-end (PE) library with an insert size of 350bp was generated from *Triticum aestivum* cv. Sonmez-2001. This library was sequenced on Illumina HiSeq 4000 platform, yielding 3,293,303,689 2x150bp reads. Sequencing and quality filtering were carried out by Genoscope - National Center of Sequencing, France.

### *Assembly of reads into Sonmez Reference Sequence*

The 970.6 Gbp of PE reads passing quality filters were mapped against the *Triticum aestivum* Chinese Spring (CS) RefSeq v1.0 genome in a 2-step approach, consisting of ungapped and gapped alignments. As the first step, an ungapped alignment was generated using BioKanga v3.4.5 with default parameters except for --substitutions=2 (i.e. 2 mismatches per 100 bp). As the second step, reads that remained unmapped were used to generate a gapped alignment with Bowtie2 v2.3.0 using default parameters except for: --very-sensitive --ignore-quals --mp 999,999 --np 999 --rdg 10,1 --rfg 10,1 --score-min L,-19,0 --n-ceil L,0,0. In doing so, reads were only allowed to map if there were zero mismatches and a single indel of length ≤ 9 bp. Alignments from both steps were using Sambamba v0.6.5.

Regions containing read alignments with indels were identified and re-aligned using GATK v3.7 using default parameters except LODThresholdForCleaning=0.4 and defaultBaseQualities=30. Sequence variations, including Single Nucleotide Polymorphisms (SNPs) and indels, were called by BCFtools v1.3.1 on pileups generated by SAMtools v1.3.1. Homozygous SNP and indel variants were identified using GATK's SelectVariants to retain only variants with zero reads containing the reference (CS) allele and 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. Using all identified homozygous variants, the “Sonmez reference sequence” was recalled.

Regions of the CS reference that were not covered by Sonmez reads were softmasked (lowercase) in the Sonmez genome sequence. GATK's FastaAlternateReferenceMaker tool was used with a slight modification to recall the softmasked CS sequence into a softmasked Sonmez genome sequence. It is important to note that these softmasked regions may represent deletions in Sonmez cultivar and/or insertions in CS.

### *De novo assembly of unmapped reads*

PE reads that remained unmapped following the 2-step approach detailed above were *de novo* assembled into Sonmez-specific genomic contigs. k-mers of length 71 bp and occurring ≥ 9 times in the unmapped reads were extracted using KMC v3.0.1. These extracted k-mers were assembled into contigs using merutensils v0.7.15 kextend 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, referred as “Sonmez unmapped contigs” are 1.05 Gbp in length in total, with the longest contig being 15,887 bp. Sonmez unmapped contigs have an N50 value of 427 bp.