# SUPPLEMENTAL METHODS

The *Sorghum bicolor* reference genome: improved assembly and annotations, a transcriptome atlas, and signatures of genome organization

Ryan F. McCormick1,2, Sandra K. Truong1,2,, Avinash Sreedasyam3, Jerry Jenkins3, Shengqiang Shu4, David Sims3, Megan Kennedy4, Mojgan Amirebrahimi4, Brock Weers2, Brian McKinley2, Ashley Mattison1,2, Daryl Morishige2, Jane Grimwood3,4, Jeremy Schmutz3,4, and John Mullet2

1. Interdisciplinary Program in Genetics, Texas A&M University, College Station, TX 77843, USA
2. Department of Biochemistry and Biophysics, Texas A&M University, College Station, TX 77843, USA
3. HudsonAlpha Institute for Biotechnology, Huntsville, AL 35806, USA
4. Joint Genome Institute, Department of Energy, Walnut Creek, CA 94598, USA

# Reference genome assembly and improvement

## Genome finishing.

The starting point for the V3.0 assembly was the previously published *S. bicolor* V1.0 genome (Paterson et al., 2009). 320 regions covering the sorghum genome were selected to include any region that contained a gene density greater than 2 genes per 100kb, effectively covering the majority or sorghum gene space. To perform finishing, initial read layouts from the S. bicolor whole genome shotgun assembly were converted into our Phred/Phrap/Consed pipeline (Ewing and Green, 1998; Ewing et al., 1998; Gordon et al., 1998). Following manual inspection of the assembled sequences, finishing was performed by resequencing plasmid subclones and by walking on plasmid subclones or fosmids using custom primers. All finishing reactions were performed with 4:1 BigDye to dGTP BigDye terminator chemistry (Applied Biosystems). Small repeats in the sequence were resolved by transposon-hopping 8kb plasmid clones while 454 small insert libraries were used to improve simple sequence repeats with secondary structure. To fill large gaps, resolve large repeats or to resolve chromosome duplications and extend into chromosome telomere regions, complete fosmid and BAC clones were shotgun sequenced and finished, followed by integration back into the genome sequences. For large clone insert sequencing: 1,018 BAC clones were sequenced, with 861 of these finished, 943 fosmid clones were sequenced with 906 of these individually finished. For targeted regions, 141,270 custom primed sequencing walks were performed from subclones or fosmids. For specialty libraries, 1,418 shatter libraries were sequenced to resolve tandem repeats and 285 small insert shatter libraries were completed to resolve high GC hairpin regions.   
  
Following completion each assembly was validated by an independent quality assessment. This included a visual examination of subclone paired ends and repeat structures and validation of any remaining lower quality regions and regions with high quality base pair discrepancies. These improved sequences have an estimated error rate of less than one error in 100,000 base pairs.

## Integration of finished regions.

A total of 351 Mb of finished sorghum sequence was integrated into the V1.0 published genome. Integration of the finished regions began by aligning the regions to the existing V1.0 assembly and assigning regions to a chromosome. Regions were then oriented with respect to their assigned chromosome. Overlapping regions were merged together to produce merged regions. 50Kb of the front and end of the merged regions was aligned to the genome to identify the start and end of the merged region. Sequence and quality scores were then integrated into the assembly, and statistics on the overlap region were obtained. The integrated assembly was screened for E. coli and adapter sequences (ex. NEB2), and a total of 10,339 bp was identified as foreign and excised from the assembly. In total, 349 clones were integrated representing 344.4 MB of sequence. 4,426 gaps were closed, and a total of 4.96 MB of sequence was added to the assembly. Overall contiguity (contig N50) increased by a factor of 5.8x from 204.5 KB to 1.2 MB (Table 1).

**Table 1: Integration of Finished Regions and Assembly Integration**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Chromosome** | **Regions** | **Gaps Closed** | **Total Ns** | **Integrated Clone Bases** | **Initial Bases** | **Final Bases** | **Bases Gained** |
|
| 1 | 56 | 675 | 690,055 | 55,999,099 | 68,249,213 | 68,853,856 | 604,643 |
| 2 | 47 | 555 | 726,117 | 44,027,412 | 76,503,847 | 77,039,739 | 535,892 |
| 3 | 42 | 471 | 465,865 | 41,975,176 | 73,105,486 | 73,518,897 | 413,411 |
| 4 | 39 | 490 | 530,562 | 40,882,467 | 63,053,210 | 63,608,275 | 555,065 |
| 5 | 28 | 407 | 567,972 | 25,720,260 | 59,092,703 | 59,523,671 | 430,968 |
| 6 | 30 | 334 | 320,363 | 27,740,705 | 61,113,938 | 61,479,048 | 365,110 |
| 7 | 28 | 403 | 626,048 | 28,862,211 | 59,931,894 | 60,627,098 | 695,204 |
| 8 | 22 | 365 | 505,198 | 22,706,396 | 51,957,801 | 52,339,273 | 381,472 |
| 9 | 28 | 367 | 547,621 | 26,161,131 | 55,531,183 | 55,840,465 | 309,282 |
| 10 | 29 | 359 | 411,365 | 30,362,799 | 57,096,972 | 57,761,342 | 664,370 |
|  |  |  |  |  |  |  |  |
| Totals | 349 | 4,426 | 5,391,166 | 344,437,656 |  |  | 4,955,417 |

## Map integration

The newly constructed genetic map (10,789 markers) was used to identify mis-assemblies and integrate the assembled scaffolds into 10 linkage groups. Scaffolds were broken if they contained a putative false join coincident with an area of low BAC/fosmid coverage. A total of 8 breaks were identified in the V1.0 release chromosomes. The new map enabled the identification of 7 previously unmapped scaffolds (Table 2), and they were inserted in the appropriate location. Additionally, a 1.08 MB region of the V1.0 chromosome 6 was moved to chromosome 7. Scaffolds were then ordered and oriented using the marker map. A total of 15 joins were made to form the final assembly containing 10 chromosomes capturing 655.2 Mb (93.8%) of the assembled sequence. Each join is padded with 10,000 Ns. Plots of the marker placements for the 10 chromosomes are given in Figures 1-10.

## Polishing homozygous SNPs/INDELs

Remaining homozygous SNPs and INDELs were corrected in the release using 110x of 2x250 (800 bp insert) Illumina fragments taken from the same strain. Reads were aligned to the integrated assembly using bwa mem (Li and Durbin, 2009). We then applied GATK (McKenna et al., 2010) base quality score recalibration, indel realignment, duplicate removal, and performed SNP and INDEL discovery using standard filtering parameters following the published GATK best practices recommendations (Auwera et al., 2013; DePristo et al., 2011). SNPs and/or INDELs that were within 150bp of one another were not corrected. A total of 1,942 (41% of called) homozygous snps and 1,432 (82% of called) homozygous indels were corrected in the process.

**Table 2**. Summary statistics of the scaffolds integrated into V3.0 that were unanchored in the V1.0 assembly.

|  |  |  |
| --- | --- | --- |
| **Chromosome** | **Previously Unintegrated Scaffolds** | **Integrated Bases** |
|
| 1 | 1 | 7.1 Mb |
| 4 | 1 | 589 Kb |
| 5 | 2 | 9.6 Mb |
| 6 | 1 | 101 Kb |
| 8 | 1 | 7.3 Mb |
| 9 | 1 | 10.3 Kb |
| Totals | 7 | 24.7 Mb |

## Assessment of completeness

Completeness of the euchromatic portion of the V3.0 release was assessed by aligning the previously annotated V 1.0 genes to identify genes that were affected by the region integration. A total of 10 genes out of 29,448 were affected. Table 3 lists the affected genes.

**Table 3:** Summary of the 10 genes annotated in V1.0 that were affected by finished region integration.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Chromosome** | **V1.0 Start** | **V1.0 End** | **PAC ID** | **Gene ID** |
| chromosome\_1 | 48,268,373 | 48,270,455 | 1952401 | Sb01g027750.1 |
| chromosome\_2 | 66,301,267 | 66,303,661 | 1958418 | Sb02g031360.1 |
| chromosome\_3 | 9,581,637 | 9,582,252 | 1961098 | Sb03g008840.1 |
| chromosome\_3 | 59,077,134 | 59,080,784 | 1962810 | Sb03g030790.1 |
| chromosome\_4 | 5,331,079 | 5,331,877 | 1965542 | Sb04g005490.1 |
| chromosome\_4 | 5,335,768 | 5,341,758 | 1965543 | Sb04g005500.1 |
| chromosome\_4 | 66,122,081 | 66,122,557 | 1968514 | Sb04g036336.1 |
| chromosome\_5 | 52,379,367 | 52,380,790 | 1970568 | Sb05g021426.1 |
| chromosome\_5 | 5,487,148 | 5,489,618 | 1969346 | Sb05g004350.1 |
| chromodome\_7 | 56,974,432 | 56,975,236 | 1976260 | Sb07g022445.1 |

|  |
| --- |
| Local HD:Users:jjenkins:Desktop:work:assembly_writeups:Sorghum:S_bicolor_V3_chromosome_1_plot.png |
| **Figure 1:** Marker placements for the genetic map on the *Sorghum bicolor* chromosome 1. |
| Local HD:Users:jjenkins:Desktop:work:assembly_writeups:Sorghum:S_bicolor_V3_chromosome_2_plot.png | |
| **Figure 2:** Marker placements for the genetic map on the *Sorghum bicolor* chromosome 2. | |
| Local HD:Users:jjenkins:Desktop:work:assembly_writeups:Sorghum:S_bicolor_V3_chromosome_3_plot.png | |
| **Figure S3:** Marker placements for the genetic map on the *Sorghum bicolor* chromosome 3. | |
| Local HD:Users:jjenkins:Desktop:work:assembly_writeups:Sorghum:S_bicolor_V3_chromosome_4_plot.png | |
| **Figure 4:** Marker placements for the genetic map on the *Sorghum bicolor* chromosome 4. | |

|  |
| --- |
| Local HD:Users:jjenkins:Desktop:work:assembly_writeups:Sorghum:S_bicolor_V3_chromosome_5_plot.png |
| **Figure 5:** Marker placements for the genetic map on the *Sorghum bicolor* chromosome 5. |
| Local HD:Users:jjenkins:Desktop:work:assembly_writeups:Sorghum:S_bicolor_V3_chromosome_6_plot.png |
| **Figure 6:** Marker placements for the genetic map on the *Sorghum bicolor* chromosome 6. |

|  |
| --- |
| Local HD:Users:jjenkins:Desktop:work:assembly_writeups:Sorghum:S_bicolor_V3_chromosome_7_plot.png |
| **Figure 7:** Marker placements for the genetic map on the *Sorghum bicolor* chromosome 7. |
| Local HD:Users:jjenkins:Desktop:work:assembly_writeups:Sorghum:S_bicolor_V3_chromosome_8_plot.png |
| **Figure 8:** Marker placements for the genetic map on the *Sorghum bicolor* chromosome 8. |

|  |
| --- |
| Local HD:Users:jjenkins:Desktop:work:assembly_writeups:Sorghum:S_bicolor_V3_chromosome_9_plot.png |
| **Figure 9:** Marker placements for the genetic map on the *Sorghum bicolor* chromosome 9. |
| Local HD:Users:jjenkins:Desktop:work:assembly_writeups:Sorghum:S_bicolor_V3_chromosome_10_plot.png |
| **Figure 10:** Marker placements for the genetic map on the *Sorghum bicolor* chromosome 10. |

# Transcriptome atlas construction and analyses

## Plant growth conditions.

For collection of juvenile stage tissue, 15 seeds were planted to a depth of 2.5 cm on 10/31/13 in each of three 3-gallon pots (n=45) containing brown topsoil (All American Stone & Turf) equivalent to native field soil from Brazos County, TX. One teaspoon of Osmocote 13-13-13 was mixed into each pot before planting. The plants were grown under 14 hour day greenhouse conditions with an average daily high of 29.5C and an average daily low of 21.9C. They emerged on 11/3/13 and were harvested on 11/11/13 (8 DAE).

For the remaining developmental stages (vegetative, floral induction plus ten days, anthesis, and grain maturity), 36 (10 cm x 150 cm) schedule 40 PVC pipes were used. The pipes were filled with the same soil as above, with the exception that one tablespoon of Osmocote 13-13-13 was mixed in with the soil for each pipe. Each pipe was thoroughly watered and then refilled with soil to bring the soil level to within five cm from the top of the pipe. Seeds were imbibed in water overnight at room temperature under constant aeration on 10/15/13. The seeds were transferred to germination paper on 10/16/13 and grown under greenhouse conditions vertically in a 4L beaker containing five cm of water and a plastic film cover. On 10/19/13 (1 DAE), the seedlings were transferred to pipes and then thoroughly watered. The plants were grown under 14 hour day greenhouse conditions with an average daily high of 32.7C and an average daily low of 23.4C. They were watered as needed - generally every 3-4 days. Harvests were conducted at the vegetative stage (11/11/13, 24 DAE), floral induction plus ten days (12/17/13, 44 DAE), anthesis (1/7/14, 65 DAE), and grain maturity (2/7/14, 96 DAE).

## Tissues collected

Plants were individually harvested for tissues collected and then pooled to generate three biological replicates, each containing nine plants (in the juvenile stage) or three plants (in the remaining stages).

The root tissues were collected for all developmental stages assayed and are comprised of two separate root tissue regions. The soil was washed away from the roots and then (a) root tips (distal 1 cm) and (b) mid sections (2.5 cm) were collected.

The leaf tissues were also collected for all developmental stages and are comprised of separate leaf tissue regions. In the juvenile stage, the following regions were collected: (a) the first emerged leaf, (b) the lower part of the third emerging leaf (still contained in the whorl), and (c) the upper part of the third emerging leaf. In the vegetative stage, after the 1st through 4th emerged leaves were removed, the following regions were collected: (a) the lower region of the shoot containing the remaining leaf sheaths, (b) the region above the sheath including some leaf regions immediately emerging from the whorl, and (c) the upper leaves of the whorl. In the floral initiation stage, after removing all fully emerged leaves except the last fully emerged (ligulated) leaf, the last fully emerged leaf was used to collect (a) the leaf sheath, (b) the lower leaf blade, closer to the sheath, and (c) the upper leaf blade. In the anthesis stage the last fully expanded leaf below the flag leaf was used to collect: (a) the leaf sheath, (b) the lower leaf blade, closer to the sheath, and (c) the upper leaf blade.

The stem tissues were also collected for all developmental stages. In the juvenile and vegetative stage, after separating the stem from its surrounding leaf sheath, the whole stem (2-3 mm, 1 cm in length, respectively) was harvested. In the floral initiation stage, two regions of the stem were collected: (a) mature internode, the middle of stem, and (b) growing internodes, the zone of internode elongation. Growing internodes that were approximately 2 cm in length were selected. All internodes were excised at the nodes. At anthesis and grain maturity stages, two regions of the stem were collected: (a) middle internodes and (b) the first internode below the peduncle. Both regions are fully developed, and all internodes were excised at the nodes.

The panicle tissues were collected at the floral initiation and anthesis stages. In the floral initiation stage, two samples were collected; (a) the peduncle and (b) the entire panicle. In the anthesis stage, two vertical regions of the panicle were collected; (a) the upper panicle (the top half of the panicle), (b) and the lower panicle (the bottom half of the panicle attached to the peduncle (the peduncle was not included in this sample))

## Nitrogen study.

A nitrogen study was conducted in turface under high urea, nitrate, ammonium, and no nitrogen standard conditions using the standard fertilizer recipes (attached). The plants did not grow well and consequentially had little RNA to extract. Therefore, a second study was done in field soil using half of each nitrogen source as well as a no nitrogen control and a NPK (0.5 tablespoon of Osmocote 13-13-13 per pipe) control. 75 (10 cm x 75 cm) schedule 40 PVC pipes were used, each filled with the same soil as above. Each pipe was thoroughly watered and then refilled with soil to bring the soil level to within five cm from the top of the pipe. Seeds were imbibed in water overnight at room temperature under constant aeration on 4/8/14. The seeds were transferred to germination paper on 4/9/14 and grown under greenhouse conditions vertically in a 4L beaker containing five cm of water and a plastic film cover. On 4/12/14 (1 DAE), the seedlings were transferred to the pipes and then thoroughly watered. The plants were grown under 14 hour day greenhouse conditions with an average daily high of 37.5C and an average daily low of 23.9C. For the first ten days of growth, the seedlings were watered as needed - generally every 3-4 days. At day ten, each pipe was given 500 ml of its respective treatment (0.5X of the attached standard solutions or water). Thereafter, every three days the pipes were alternately watered or fertilized at the rate of 500 ml per pipe. At 30 DAE (5/11/14) all pipes were harvested. For each of three reps per treatment (5 plants per rep), soil was washed away from the roots and then the roots and shoots were individually separated and pooled.

## Gene annotation

Information originally documented at <https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Sbicolor_er>

The present v3.1 release, comprising the v3.0 assembly and v3.1 gene set, is a modern annotation using resources used in the original v1.0 release (Sbi1 assembly and Sbi1.4 gene set) and geneAtlas RNA-seq data. The main genome is in 10 chromosomes with small unmapped pieces, some of which contain annotated genes.

127,415 RNAseq transcript assemblies were constructed from about ~3.3B pairs of sorghum stranded paired-end Illumina RNAseq reads using PERTRAN (Shu et. al., unpublished pipeline using GSNAP as read aligners). 111,994 transcript assemblies were constructed using [PASA](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC206470/?tool=pubmed) from *Sorghum bicolor* RNAseq transcript assemblies above and 209,835 ESTs. Loci were determined by transcript assembly alignments and/or EXONERATE alignments of proteins from arabi (*Arabidopsis thaliana*), rice, maize or grape genomes to repeat-soft-masked *S. bicolor* genome using [RepeatMasker](http://www.repeatmasker.org/). Gene models were predicated by homology-based predictors, mainly[FGENESH+](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1810547/?tool=pubmed), FGENESH\_EST (similar to FGENESH+, EST as splice site and intron input instead of protein/translated ORF), and [GenomeScan](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC311055/?tool=pubmed). The best scored predictions for each locus are selected using multiple positive factors including EST and protein support, and one negative factor: overlap with repeats. The selected gene predictions were improved by PASA. Improvement includes adding UTRs, splicing correction, and adding alternative transcripts. PASA-improved gene model proteins were subject to protein homology analysis to above mentioned proteomes to obtain Cscore and protein coverage. Cscore is a protein BLASTP score ratio to MBH (mutual best hit) BLASTP score and protein coverage is highest percentage of protein aligned to the best of homologs. PASA-improved transcripts were selected based on Cscore, protein coverage, EST coverage, and its CDS overlapping with repeats. The transcripts were selected if its Cscore is larger than or equal to 0.5 and protein coverage larger than or equal to 0.5, or it has EST coverage, but its CDS overlapping with repeats is less than 20%. For gene models whose CDS overlaps with repeats for more that 20%, its Cscore must be at least 0.9 and homology coverage at least 70% to be selected. The selected gene models were subject to Pfam analysis and gene models whose protein is more than 30% in Pfam TE domains were removed.

v2.1 loci were tentatively mapped to v3.1 loci by BLAT both v2.1 loci sequence including intron bounded by its CDS range and v2.1 loci sequence including intron bounded by its range extending up to 1K bp to v2.0 assembly. For each loci pairing, their proteins were aligned to each other. When MBH protein is >= 70% identical, v2.1 locus name becomes v3.1 locus name (88% of v2.1 loci mapped this way). When MBH protein is >= 90% identical, v2.1 synonym and defLine if any become v3.1 synonym and defLine respectively.

# REFERENCES

Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy‐Moonshine, A., Jordan, T., Shakir, K., Roazen, D., and Thibault, J. (2013). From FastQ data to high‐confidence variant calls: the genome analysis toolkit best practices pipeline. *Current protocols in bioinformatics*, 11.10. 1-11.10. 33.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., Del Angel, G., Rivas, M. A., and Hanna, M. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* **43**, 491-498.

Ewing, B., and Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* **8**, 186-194.

Ewing, B., Hillier, L., Wendl, M. C., and Green, P. (1998). Base-calling of automated sequencer traces usingPhred. I. Accuracy assessment. *Genome research* **8**, 175-185.

Gordon, D., Abajian, C., and Green, P. (1998). Consed: a graphical tool for sequence finishing. *Genome research* **8**, 195-202.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754-1760.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., and Daly, M. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297-1303.

Paterson, A. H., Bowers, J. E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., Haberer, G., Hellsten, U., Mitros, T., and Poliakov, A. (2009). The Sorghum bicolor genome and the diversification of grasses. *Nature* **457**, 551-556.