

Figure S1 (continued)

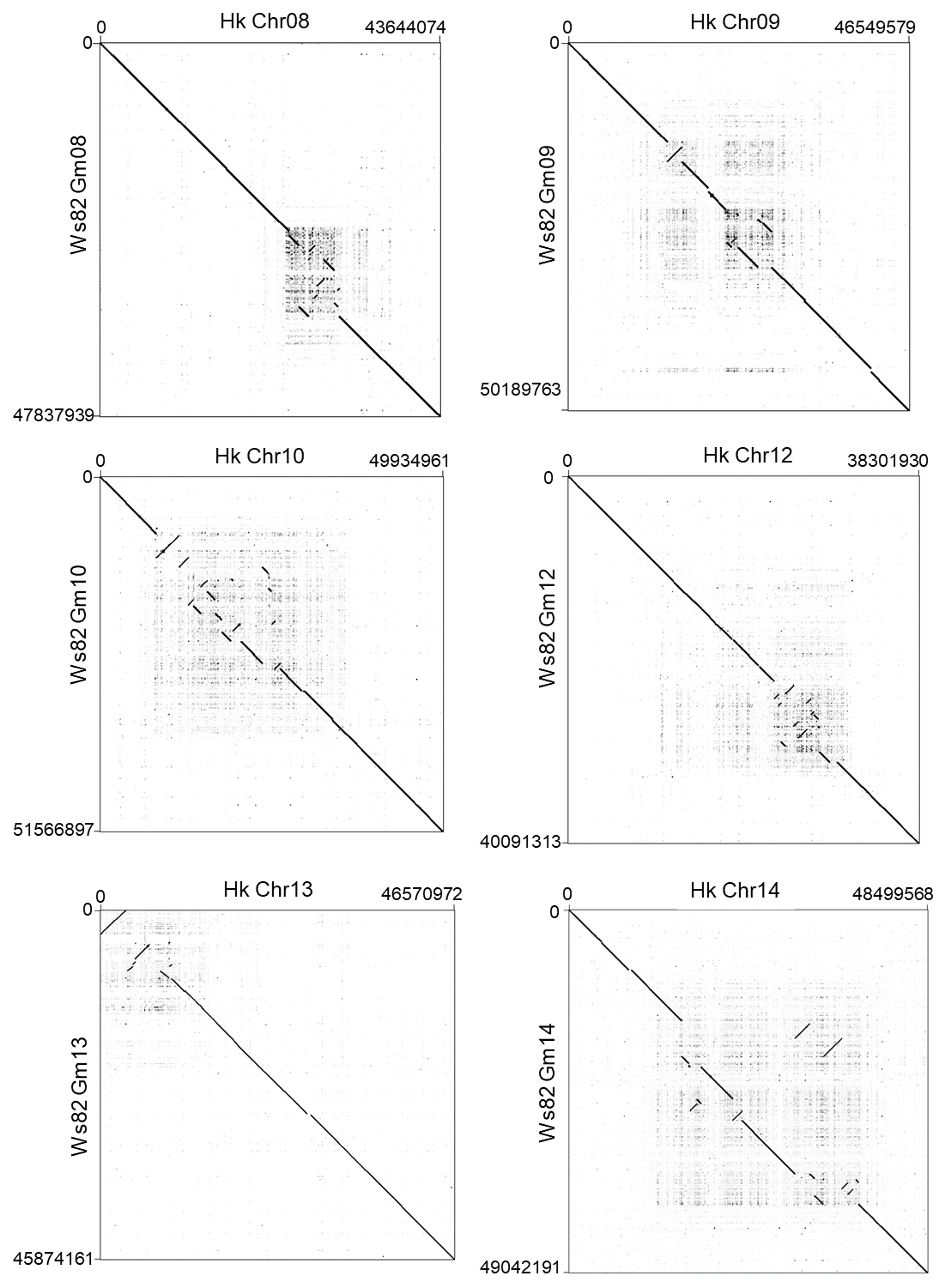


Figure S1 (continued)

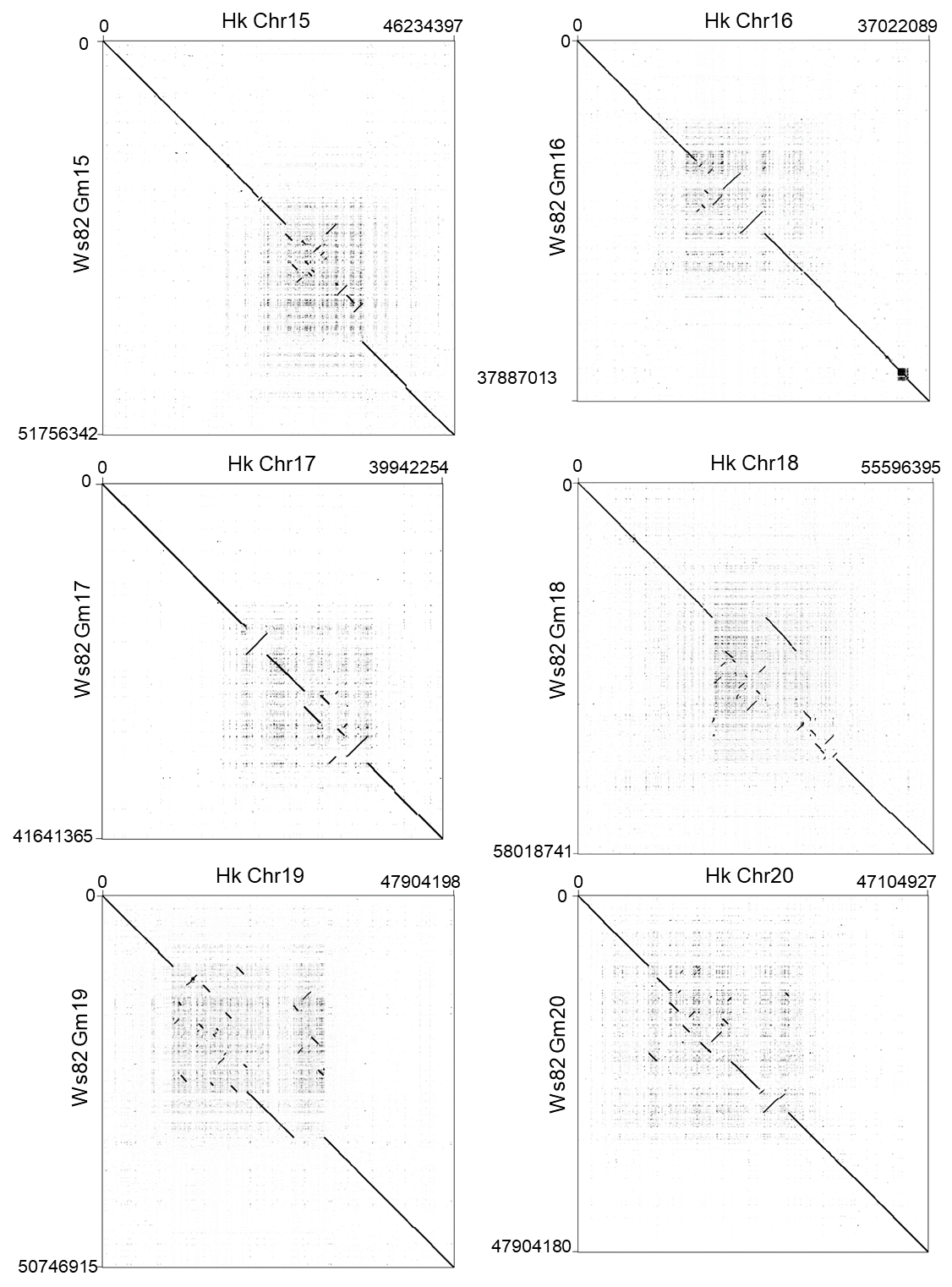


Figure S1. Dot plots showing alignments of corresponding chromosomes between the Hwangkeum and Williams 82 assemblies.

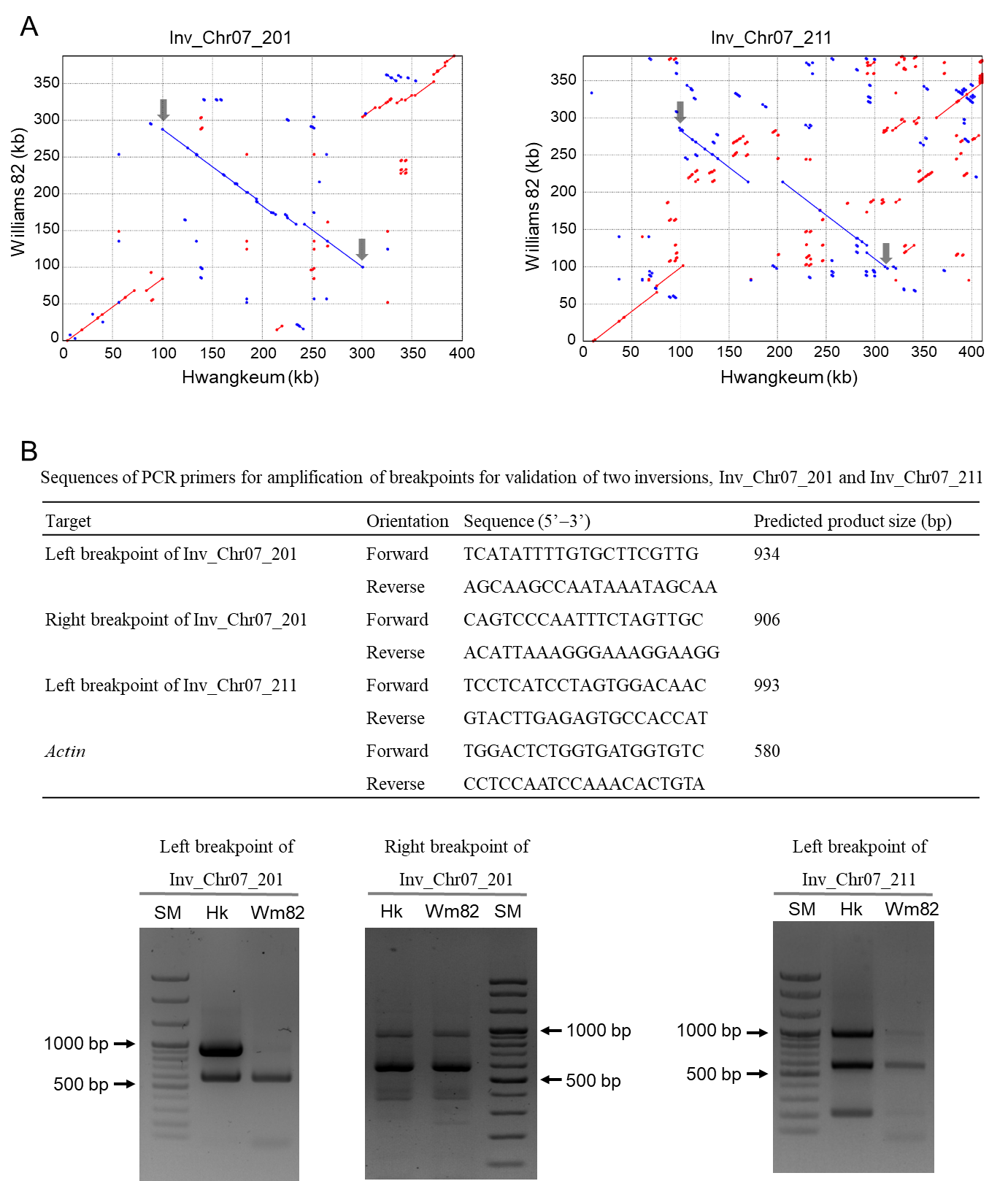


Figure S2. Validation of two large inversions. A. Alignments between Hwangkeum and Williams 82 sequences of two large inversions, Inv\_Chr07\_201 (201 kb) and Inv\_Chr07\_211 (211 kb). Sequences including inversions and their 100-kb flanking sequences were aligned. Sequence coordinates of Inv\_Chr07\_201 are Chr07:31792160-32192735 for the Hwangkeum assembly and Gm07:34127441-34515006 for the Williams 82 assembly. Sequence coordinates of Inv\_Chr07\_211 are Chr07:32165100-32575725 for the Hwangkeum assembly and Gm07:34505863-34888807 for the Williams 82 assembly. Forward maximum unique matches (MUMs) are plotted as red lines/dots while reverse MUMs are plotted as blue lines/dots. Breakpoints of inversions are indicated by gray arrows. B. Multiplex PCR for validation of breakpoints of inversions. PCR primer targets and sample names (SM, a 100-bp DNA ladder, Bioneer, Daejeon, Korea; Hk, Hwangkeum; Wm82, Williams 82) are indicated. Of the three PCR targets attempted, only one target gave a specific product. A common 580-bp band PCR-amplified from the actin gene is shown. Because sequence alignment of right breakpoint of Inv\_Chr07\_211 showed complicated repeat structure, we did not attempt any specific PCR amplification. Although the alignment at the right breakpoint of Inv\_Chr07\_201 appeared to be specific, there is a possibility of non-specific PCR product amplification because this breakpoint contains sequence of a member of a serine/threonine-protein phosphatase multi-gene family.

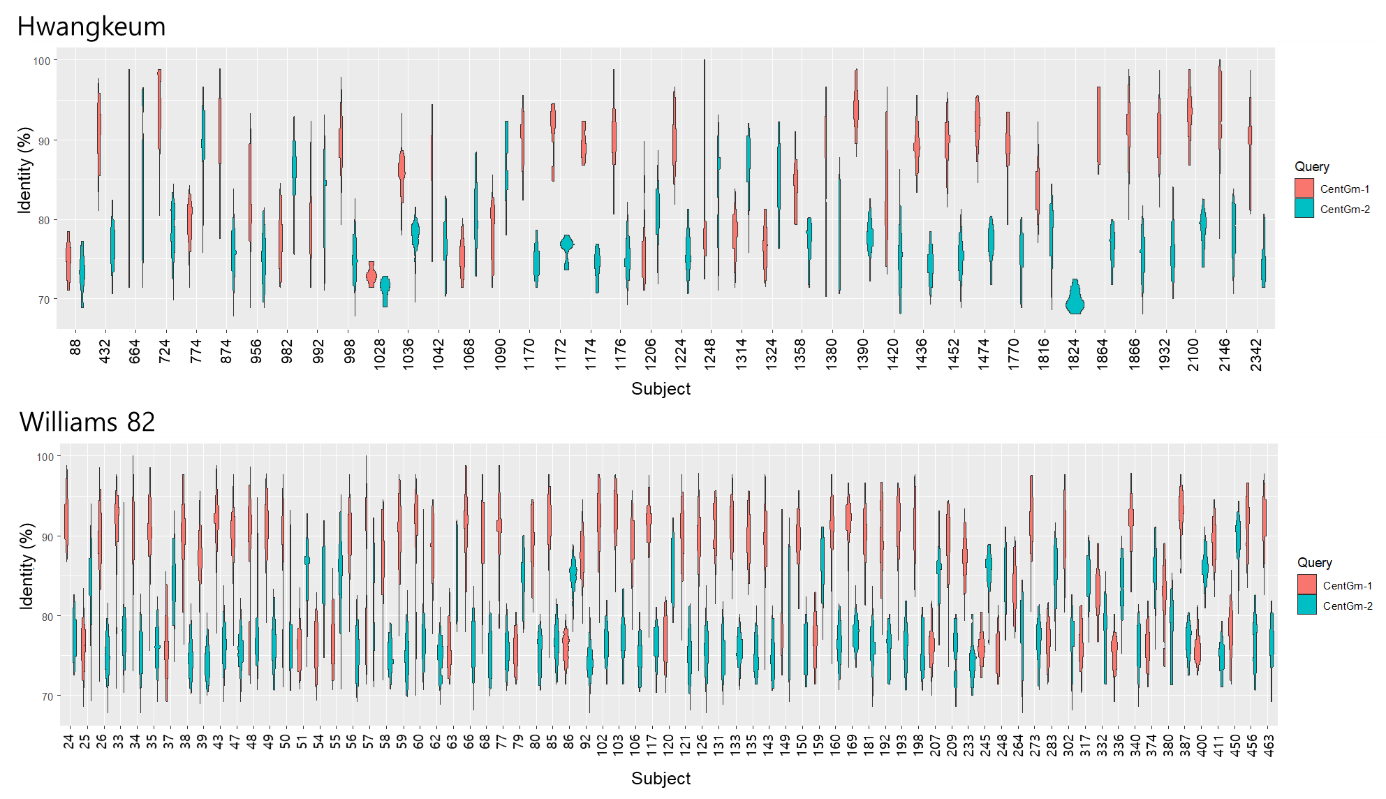


Figure S3. Violin plot distributions of the percent identity of centromeric repeats hit by BLAST searches with CentGm-1 and CentGm-2, respectively, along the unanchored contigs or scaffolds, as sampled in the Hwangkeum and Williams 82 assemblies.

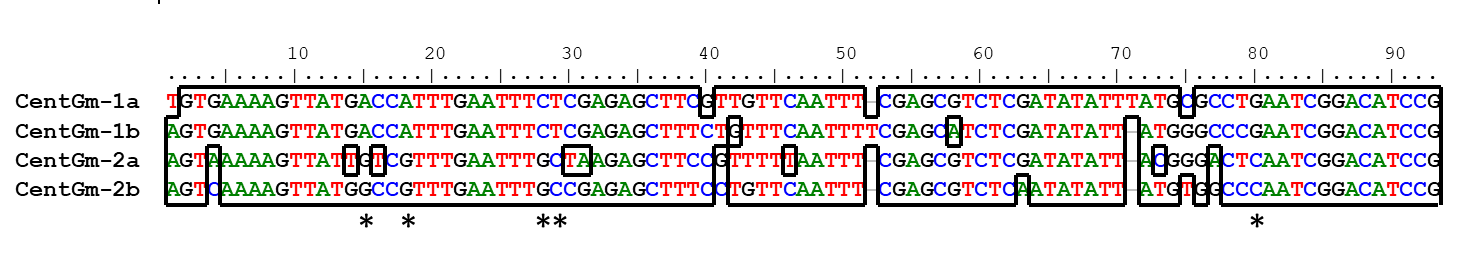


Figure S4. ClustalW alignment of consensus DNA sequences of four CentGm consensus repeats extracted in this study. Asterisks indicate nucleotide positions that distingush between CenGm-1a and -1b and CenGm-2a and -2b. CenGm-1a and CentGm-1b are 92 bp and CenGm-2a and CentGm-2b are 91 bp.

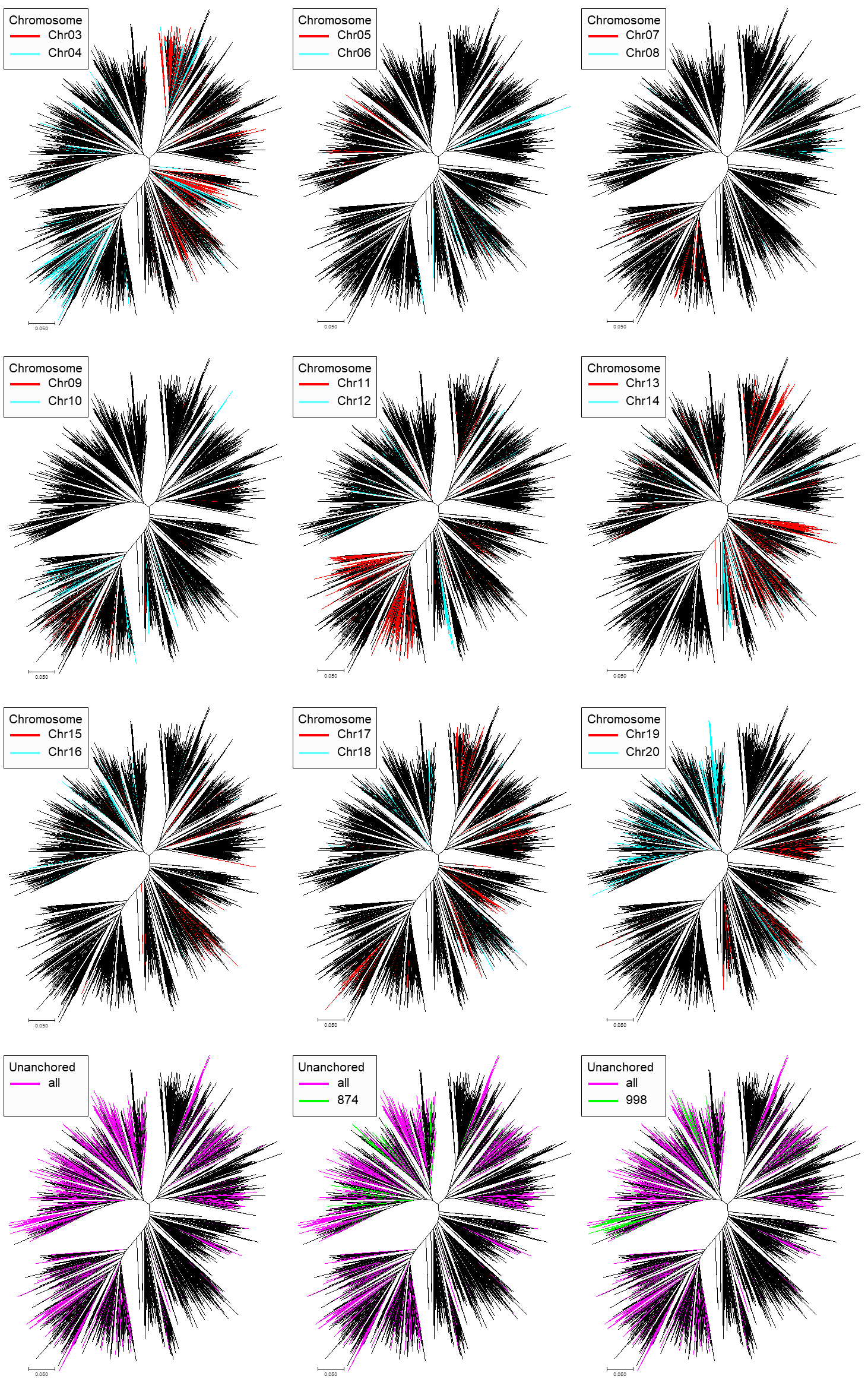


Figure S5. Neighbor-joining phylogenetic tree of centromeric repeat sequences in the Hwangkeum assembly to show clustering of the sequences located on each of chromosome and unanchored contigs. Nine of 4478 sequences are publicly available representative sequences. Sequences on odd number of chromosomes are indicated by red branches and those on even number of chromosomes by light blue branches. Sequences on unanchored contigs are indicated by pink branches and sequences on unanchored contigs Contig\_Un0874 and Contig\_Un0998 are highlighted by green branches to show how repeat sequences of each contig spread over among repeat groups.

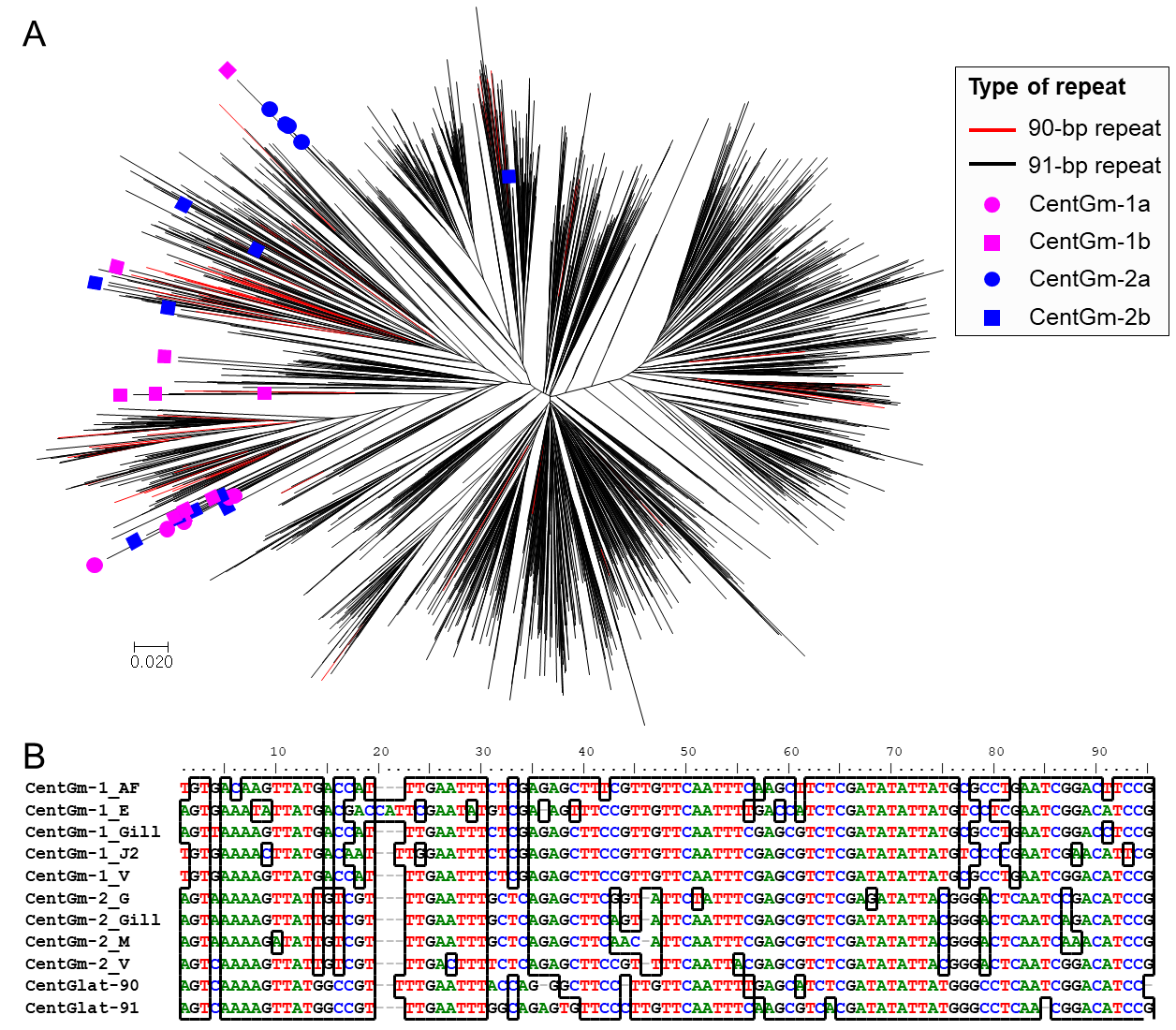


Figure S6. Evolution of centromeric satellite repeats in *Glycine latifolia*. A. Neighbor-joining tree of 3046 satellite repeats consisting of 3107 repeat sequences from *G. latifolia*, five CentGm-1 representative sequences, four CentGm-2 representative sequences, and ten sequences from each of the CentGm-1b and CentGm-2b groups extracted from the Hwangkeum assembly. B. ClustalW alignment of consensus DNA sequences of the publicly available 9 representative consensus CentGm repeats (Findley *et al*. 2010 and Valliyodan *et al*. 2019) and two CentGlat consensus repeats extracted in this study.

Table S1. Summary statistics of initial contig assemblies of Hwangkem sequencing data generated by the PacBio SMRT and Nanopore PromethION sequencing platforms

|  |  |  |  |
| --- | --- | --- | --- |
| Platform | PacBio SMRT | Nanopore PromethION | |
| Assembler | Falcon\_unzip | Shasta | Wtdbg2 |
| Number of contigs | 1,436 | 847 | 7,063 |
| Total length | 963,125,854 | 925,186,764 | 931,302,713 |
| Longest length | 11,577,401 | 24,079,308 | 7,739,327 |
| Min length | 1,391 | 1,008 | 462 |
| Average length | 670,700 | 1,092,310 | 131,856 |
| N50 | 1,713,290 | 6,946,143 | 524,993 |
| L50 | 154 | 37 | 403 |
| N60 | 1,235,573 | 4,325,963 | 365,676 |
| L60 | 219 | 55 | 615 |
| N70 | 895,603 | 2,725,692 | 231,968 |
| L70 | 311 | 81 | 936 |
| N80 | 597,197 | 1,558,438 | 132,392 |
| L80 | 443 | 126 | 1,467 |
| N90 | 341,281 | 838,475 | 56,109 |
| L90 | 652 | 205 | 2,553 |

Table S2. Summary statistics of primary assembly and scaffolding of Hwangkem sequencing data generated by the PacBio SMRT sequencing

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Step | Falcon-Unzip | Error correction | Scaffolding with mate-pair reads | Gap filling with mate-pair reads | Second scaffolding and gap filling with paired-end reads |
| Number of  contigs or scaffolds | 1,436 | 1,331 | 850 | 850 | 730 |
| Total length | 963,125,854 | 966,112,314 | 966,246,010 | 966,655,439 | 966,246,105 |
| N50 | 1,713,290 | 1,753,420 | 2,342,200 | 2,342,263 | 2,540,268 |
| Max length | 11,577,401 | 11,715,299 | 11,715,299 | 11,715,299 | 11,715,299 |
| Min length | 1,391 | 3,916 | 4,588 | 4,588 | 4,588 |
| Average length | 670,700 | 725,854 | 1,136,760 | 1,137,241 | 1,323,624 |
| Gap | - | - | 477 | 467 | 594 |
| N’s | - | - | 133,964 | 119,969 | 134,173 |

Table S3. Summary statistics of chromosome-scale pseudomolecule generation with ALLMAPS for each of the two genetic component maps (HI and WH with equal weights) and final consensus anchoring (‘Anchored’) in the FALCON-Unzip SMRT and Shasta PromethION assemblies

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | HI | WH | Anchored | Oriented | Unplaced |
| FALCON-Unzip SMRT assembly | | | | | |
| Linkage Groups | 20 | 20 | 20 | NA | NA |
| Markers (unique) | 50,867 | 23,708 | 63,267 | 52,204 | 5 |
| Markers per Mb | 54.2 | 28.3 | 67 | 94.3 | 0.2 |
| N50 Scaffolds | 116 | 110 | 116 | 90 | 0 |
| Scaffolds | 627 | 474 | 640 | 201 | 90 |
| Scaffolds with 1 marker | 46 | 68 | 37 | 0 | 0 |
| Scaffolds with 2 markers | 22 | 31 | 22 | 0 | 1 |
| Scaffolds with 3 markers | 26 | 17 | 19 | 0 | 1 |
| Scaffolds with >=4 markers | 533 | 358 | 562 | 201 | 0 |
| Total bases (Mb) | 937.94 | 837.63 | 944.02 | 553.39 | 22.32 |
| (Percent of total assembly) | 97.10% | 86.70% | 97.70% | 57.30% | 2.30% |
| Shasta PromethION assembly | | | | | |
| Linkage Groups | 20 | 20 | 20 | NA | NA |
| Markers (unique) | 48,075 | 23,587 | 60,412 | 54,154 | 9 |
| Markers per Mb | 53 | 28.4 | 66.5 | 85.3 | 0.5 |
| N50 Scaffolds | 37 | 37 | 37 | 37 | 0 |
| Contigs | 387 | 288 | 399 | 90 | 448 |
| Contigs with 1 marker | 55 | 50 | 45 | 0 | 5 |
| Contigs with 2 markers | 29 | 24 | 24 | 0 | 2 |
| Contigs with 3 markers | 24 | 10 | 19 | 0 | 0 |
| Contigs with >=4 markers | 279 | 204 | 311 | 90 | 0 |
| Total bases (Mb) | 906.63 | 829,.60 | 907.9 | 634.65 | 19.74 |
| (Percent of total assembly) | 97.70% | 89.40% | 97.90% | 68.40% | 2.10% |

Markers and scaffolds that failed to be incorporated into the final pseudomolecule assembly are presented in the “unplaced” column. NA represents “not applicable”.

Table S4. BUSCO assessment of the Hwangkeum genome assembly with eukaryota\_odb10 and embryophyta\_odb10 databases

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | eukaryota\_odb10 | |  | embryophyta\_odb10 | |
|  | Count | Ratio (%) |  | Count | Ratio (%) |
| Complete BUSCOs (C) | 250 | 98.1 |  | 1,590 | 98.5 |
| Complete and single-copy BUSCOs (S) | 56 | 22 |  | 778 | 48.2 |
| Complete and duplicated BUSCOs (D) | 194 | 76.1 |  | 812 | 50.3 |
| Fragmented BUSCOs (F) | 3 | 1.2 |  | 9 | 0.6 |
| Missing BUSCOs (M) | 2 | 0.7 |  | 15 | 0.9 |
| Total BUSCO groups searched | 255 | 100 |  | 1,614 | 100 |

Table S5. Number and distributions of 91-bp and 92-bp satellite repeats on chromosomes and unanchored contigs extracted using the Tandem Repeat Finder

|  |  |  |
| --- | --- | --- |
| Chromosome | Number of 91-bp CentGm-2 repeats | Number of 92-bp CentGm-1 repeats |
| Chr01 | 29 | 0 |
| Chr02 | 90 | 226 |
| Chr03 | 910 | 386 |
| Chr04 | 483 | 557 |
| Chr05 | 39 | 883 |
| Chr06 | 147 | 4 |
| Chr07 | 183 | 0 |
| Chr08 | 65 | 157 |
| Chr09 | 141 | 2 |
| Chr10 | 442 | 4 |
| Chr11 | 1756 | 7 |
| Chr12 | 114 | 603 |
| Chr13 | 1028 | 446 |
| Chr14 | 207 | 130 |
| Chr15 | 80 | 62 |
| Chr16 | 20 | 353 |
| Chr17 | 697 | 70 |
| Chr18 | 16 | 273 |
| Chr19 | 284 | 138 |
| Chr20 | 48 | 3363 |
| Unanchored contigs | 2987 | 7429 |
| Total | 9766 | 15093 |

Table S6. Single nucleotide polymorphisms (SNPs), ratios of transition (ts) to transversion (tv), and insertions and deletions per chromosome called from the Hwangkeum assembly against the Williams 82 reference genome sequence Wm82.a2.v1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Chromosome | SNPs | ts/tv | Insertions | Deletions |
| Chr01 | 77,278 | 1.64 | 12,555 | 14,586 |
| Chr02 | 85,453 | 1.656 | 12,601 | 12,879 |
| Chr03 | 131,879 | 1.76 | 16,501 | 15,932 |
| Chr04 | 41,628 | 1.337 | 8,574 | 9,718 |
| Chr05 | 71,763 | 1.378 | 11,327 | 11,520 |
| Chr06 | 124,611 | 1.722 | 16,053 | 16,782 |
| Chr07 | 91,435 | 1.727 | 13,194 | 13,548 |
| Chr08 | 63,758 | 1.459 | 11,541 | 11,891 |
| Chr09 | 78,989 | 1.61 | 13,055 | 13,694 |
| Chr10 | 92,479 | 1.749 | 12,871 | 13,432 |
| Chr11 | 44,146 | 1.486 | 8,461 | 8,341 |
| Chr12 | 61,231 | 1.62 | 9,625 | 9,434 |
| Chr13 | 116,657 | 1.614 | 15,895 | 16,600 |
| Chr14 | 60,871 | 1.457 | 11,523 | 11,502 |
| Chr15 | 76,964 | 1.576 | 12,770 | 13,893 |
| Chr16 | 148,807 | 1.779 | 15,968 | 16,471 |
| Chr17 | 33,118 | 1.488 | 6,565 | 7,481 |
| Chr18 | 166,090 | 1.77 | 18,072 | 18,907 |
| Chr19 | 134,212 | 1.76 | 14,565 | 15,202 |
| Chr20 | 86,951 | 1.665 | 12,121 | 12,257 |
| Total | 1,788,320 | 1.61265 | 253,837 | 264,070 |

Table S7. Summary statistics of repeat sequences extracted from the Hwangkeum genome assembly

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Categories | | | Number of elements | Length occupied (bp) | Percentage of  sequence |
| Transposable elements | | |  |  |  |
|  | Retroelements | | 469,447 | 392,305,565 | 42.04% |
|  |  | LINEs | 5,464 | 3,396,297 | 0.36% |
|  |  | L1/CIN4 | 3,135 | 1,574,793 | 0.17% |
|  |  | LINEs/Ukn | 2,329 | 1,821,504 | 0.19% |
|  |  | LTR elements | 463,983 | 388,909,268 | 41.68% |
|  |  | Ty1/Copia | 187,387 | 123,338,797 | 13.22% |
|  |  | Gypsy/DIRS1 | 276,596 | 265,570,471 | 28.46% |
|  | DNA transposons | | 32,279 | 5,984,207 | 0.64% |
|  |  | hobo-Activator | 1129 | 371,523 | 0.04% |
|  |  | Tourist/Harbinger | 31,150 | 5,612,684 | 0.60% |
|  | Unclassified | | 164,370 | 61,003,472 | 6.54% |
|  | Total interspersed repeats | |  | 453,309,037 | 48.58% |
| Other repeats | | |  |  |  |
|  | Simple repeats: | | 257,256 | 11,803,751 | 1.26% |
|  | Low complexity: | | 59,297 | 3,074,160 | 0.33% |
| Total repeat sequence | | |  | 468,186,948 | 50.17% |

Table S8. BUSCO assessment of the Hwangkeum gene annotation with eukaryota\_odb10 and embryophyta\_odb10 databases

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | eukaryota\_odb10 | |  | embryophyta\_odb10 | |
|  | Count | Ratio (%) |  | Count | Ratio (%) |
| Complete BUSCOs (C) | 247 | 96.9 |  | 1,562 | 96.8 |
| Complete and single-copy BUSCOs (S) | 55 | 21.6 |  | 545 | 33.8 |
| Complete and duplicated BUSCOs (D) | 192 | 75.4 |  | 1,017 | 63.0 |
| Fragmented BUSCOs (F) | 1 | 0.4 |  | 12 | 0.7 |
| Missing BUSCOs (M) | 7 | 2.7 |  | 40 | 2.5 |
| Total BUSCO groups searched | 255 | 100 |  | 1,614 | 100 |

Table S9. EggNOG functional categories of 79,870 transcripts annotated in the Hwangkeum assemblya

|  |  |  |  |
| --- | --- | --- | --- |
| Eggnog | Description | Count | Ratio (%) |
| J | Translation, ribosomal structure and biogenesis | 1785 | 2.2882 |
| A | RNA processing and modification | 393 | 0.5038 |
| K | Transcription | 4375 | 5.6083 |
| L | Replication, recombination and repair | 2614 | 3.3509 |
| B | Chromatin structure and dynamics | 378 | 0.4846 |
| D | Cell cycle control, cell division, chromosome partitioning | 650 | 0.8332 |
| Y | Nuclear structure | 4 | 0.0051 |
| V | Defense mechanisms | 569 | 0.7294 |
| T | Signal transduction mechanisms | 4122 | 5.2839 |
| M | Cell wall/membrane/envelope biogenesis | 873 | 1.1191 |
| N | Cell motility | 21 | 0.0269 |
| Z | Cytoskeleton | 825 | 1.0576 |
| W | Extracellular structures | 0 | 0.0000 |
| U | Intracellular trafficking, secretion, and vesicular transport | 2177 | 2.7907 |
| O | Posttranslational modification, protein turnover, chaperones | 4273 | 5.4775 |
| C | Energy production and conversion | 969 | 1.2421 |
| G | Carbohydrate transport and metabolism | 2727 | 3.4957 |
| E | Amino acid transport and metabolism | 1662 | 2.1305 |
| F | Nucleotide transport and metabolism | 507 | 0.6499 |
| H | Coenzyme transport and metabolism | 541 | 0.6935 |
| I | Lipid transport and metabolism | 1037 | 1.3293 |
| P | Inorganic ion transport and metabolism | 1044 | 1.3383 |
| Q | Secondary metabolites biosynthesis, transport and catabolism | 877 | 1.1242 |
| R | General function prediction only | 8368 | 10.7268 |
| S | Function unknown | 37219 | 47.7105 |
| Total | - | 78010 | 100 |

aOf the total of 79,870 transcripts encoding proteins, 76,823 transcripts matched with EggNOG database proteins and 3,047 had no hit. Of the 76,823 transcripts, 75,930 hit a single EggNOG category and 893 hit multi EggNOG categories.

Table S10. Chromosomal locations of markers generated from genes that regulate anthocyanin biosynthesis in soybean

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Gene name | Marker name | Chrom-osome number | Genetic coordinate (cM) | Coordinate in the Hwangkeum assembly | Coordinate in the Williams 82 assembly |
| Anthocyanidin reductase | SN013 | 8 | 27.1 | 4817677..4817467 | 4785962..4785752 |
| Inverted chalcone synthase repeats | SM303 | 8 | 45.5 | 8617633..8617482 | 8592707..8592572 |
|  | SM304 | 8 | 45.5 | 8613615..8613501 | 8588701..8588587 |
|  | SM305 | 8 | 45.5 | 8563616..8558593 | 8540116..8539975 |
|  | SM306 | 8 | 45.4 | 8404420..8404567 | 8373162..8373309 |
| *ANS1* (anthocyanidin synthetase) | SL015 | 11 | 11.0 | 1982052..1981948 | 1992791..1992687 |
|  | SN014 | 11 | 11.0 | 1981809..1981549 | 1992548..1992288 |
| *ANS2* (anthocyanidin synthetase) | SN015 | 1 | 95.4 | 52648583..52648680 | 54530733..54530830 |
| *DFR1* (dihydroflavonol 4-reductase) | SN016 | 14 | 34.2 | 6082559..6082738 | 6113325..6113504 |
| *DFR1* (dihydroflavonol 4-reductase) | SM004 | 14 | 34.2 | 6083857..6083758 | 6114626..6114525 |
| *DFR3* (dihydroflavonol 4-reductase) | SM005 | 2 | 90.2 | 17236640..17236730 | 17456018..17456109 |
| *DFR3* (dihydroflavonol 4-reductase) | SM325 | 2 | 90.2 | 17235364..17235623 | 17454763..17455001 |
| *DFR2* (dihydroflavonol 4-reductase) | SL016 | 17 | 114.3 | 38956879..38956683 | 40656269..40656072 |
| MYB transcription factor | SL017 | 14 | 60.6 | 28468784..28468965 | 33570807..33570988 |
| *F3'H* (flavonoid 3' hydroxylase) | SL305 | 6 | 75.6 | 18724624..18725117 | 18736389..18736620 |
| *F3'H* (flavonoid 3' hydroxylase) | SN317 | 6 | 76.2 | 18725846..18726153 | 18737347..18737655 |
| Anthocyanidin 3-O-glucosyltransferase | SM326 | 1 | 84.0 | 50469162..50468914 | 52336730..52336496 |
| Anthocyanidin 3-O-glucosyltransferase | SM327 | 1 | 84.0 | 50461429..50461167 | 52329012..52328770 |
| Anthocyanidin 3-O-glucosyltransferase | SM328 | 2 | 51.7 | 9855395..9855588 | 9883176..9883371 |
| Anthocyanidin 3-O-glucosyltransferase | SM329 | 13 | 29.6 | 29578077..29578217 | 28779222..28779416 |
| Anthocyanidin 3-O-glucosyltransferase | SN017 | 16 | 70.5 | 32906717..32906332 | 33683515..33683130 |
| Anthocyanidin 3-O-glucosyltransferase | SN018 | 16 | 70.9 | 32939624..32939937 | 33726307..33726620 |
| *F3H1* (flavanone 3-hydroxylase) | SN318 | 2 | 24.0 | 4406535..4406182 | 4441693..4441339 |
| *F3H2* (flavanone 3-hydroxylase) | SN319 | 2 | 24.0 | 4415087..4414979 | 4450328..4450212 |
| *F3'5'H* (flavonoid 3'5' hydroxylase) | SN019 | 13 | 14.5 | 17697484..17697766 | 17313589..17313871 |
| *F3'5'H* (flavonoid 3'5' hydroxylase) | SL018 | 13 | 14.5 | 17696620..17697213 | 17312724..17313319 |
| *F3'5'H* (flavonoid 3'5' hydroxylase) | SN020 | 13 | 14.5 | 17700675..17700903 | 17316832..17317060 |
| *F3'5'H* (flavonoid 3'5' hydroxylase) | SL019 | 13 | 14.5 | 17700710..17700408 | 17316867..17316512 |