the plant journal

SEB SOCIETY FOR EXPERIMENTAL BIOLOGY

The Plant Journal (2019) **100,** 1066–1082 doi: 10.1111/tpj.14500

RESOURCE

Construction and comparison of three reference-quality genome assemblies for soybean

Babu Valliyodan[1,2,†], Steven B. Cannon[3,†], Philipp E. Bayer[4,†], Shengqiang Shu[5], Anne V. Brown[3], Longhui Ren[6], Jerry Jenkins[7], Claire Y.-L. Chung[8], Ting-Fung Chan[8], Christopher G. Daum[5], Christopher Plott[7], Alex Hastie[9], Kobi Baruch[10], Kerrie W. Barry[5], Wei Huang[11], Gunvant Patil[1], Rajeev K. Varshney[12], Haifei Hu[4], Jacqueline Batley[4], Yuxuan Yuan[4], Qijian Song[13], Robert M. Stupar[14], David M. Goodstein[5], Gary Stacey[1], Hon-Ming Lam[8], Scott A. Jackson[15], Jeremy Schmutz[7], Jane Grimwood[7], David Edwards[4] and Henry T. Nguyen[1,]*

[1]Division of Plant Sciences and National Center for Soybean Biotechnology, University of Missouri, Columbia, 65211, MO, USA,

[2]Department of Agriculture and Environmental Sciences, Lincoln University, Jefferson City, 65101, MO, USA,

[3]Corn Insects and Crop Genetics Research Unit, US Department of Agriculture–Agricultural Research Service, Ames, 50011, IA, USA,

[4]School of Biological Sciences, The University of Western Australia, Crawley, 6009, WA, Australia,

[5]Department of Energy Joint Genome Institute, Walnut Creek, 94598, CA, USA,

[6]Interdepartmental Genetics Program, Iowa State University, Ames, 50011, IA, USA,

[7]Hudson-Alpha Institute for Biotechnology, Huntsville, 35806, AL, USA,

[8]Centre for Soybean Research of the State Key Laboratory of Agrobiotechnology and School of Life Sciences, The Chinese University of Hong Kong, Shatin, Hong Kong Special Administrative Region, China,

[9]Bionano Genomics, San Diego, 92121, CA, USA,

[10]NRGene Ltd., Ness Ziona 7403648, Israel,

[11]Department of Agronomy, Iowa State University, Ames, 50011, IA, USA,

[12]Center of Excellence in Genomics and Systems Biology, International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru 502 324, India,

[13]Soybean Genomics and Improvement Lab, US Department of Agriculture – Agricultural Research Service, Beltsville, 20705, MD, USA,

[14]Department of Agronomy and Plant Genetics, University of Minnesota, St. Paul, 55108, MN, USA, and

[15]Center for Applied Genetic Technologies, University of Georgia, Athens, 30602, GA, USA

Received 27 December 2018; revised 10 July 2019; accepted 17 July 2019; published online 21 August 2019.

*For correspondence (e-mail nguyenhenry@missouri.edu).

†These authors contributed equally to this work.

SUMMARY

We report reference-quality genome assemblies and annotations for two accessions of soybean (*Glycine max*) and for one accession of *Glycine soja*, the closest wild relative of *G. max*. The *G. max* assemblies provided are for widely used US cultivars: the northern line Williams 82 (Wm82) and the southern line Lee. The Wm82 assembly improves the prior published assembly, and the Lee and *G. soja* assemblies are new for these accessions. Comparisons among the three accessions show generally high structural conservation, but nucleotide difference of 1.7 single-nucleotide polymorphisms (snps) per kb between Wm82 and Lee, and 4.7 snps per kb between these lines and *G. soja*. snp distributions and comparisons with genotypes of the Lee and Wm82 parents highlight patterns of introgression and haplotype structure. Comparisons against the US germplasm collection show placement of the sequenced accessions relative to global soybean diversity. Analysis of a pan-gene collection shows generally high conservation, with variation occurring primarily in genomically clustered gene families. We found approximately 40–42 inversions per chromosome between either Lee or Wm82v4 and *G. soja*, and approximately 32 inversions per chromosome between Wm82 and Lee. We also investigated five domestication loci. For each locus, we found two different alleles with functional differences between *G. soja* and the two domesticated accessions. The genome assemblies for multiple cultivated accessions and for the closest wild ancestor of soybean provides a valuable set of resources

for identifying causal variants that underlie traits for the domestication and improvement of soybean, serving as a basis for future research and crop improvement efforts for this important crop species.

Keywords: *Glycine max*, *Glycine soja*, soybean, genome assembly, domestication, comparative genomics.

INTRODUCTION

Soybean, *Glycine max* (L.) Merr, is an important crop that is widely used as human food, animal feed and for biofuel production, because of its high protein and oil content of 40% and 21%, respectively (Valliyodan *et al*., 2017). The cultivated species of soybean was domesticated in China approximately 5000 years ago from a wild progenitor related to the closest extant relative, *Glycine soja*. Domestication and selection have since led to a reduction in genetic diversity (Hyten *et al*., 2006).

The reference genome assembly of cultivated soybean is of a northern US cultivar, Williams 82 (Wm82), produced by the soybean research community in collaboration with the Department of Energy Joint Genome Institute (DOE-JGI) (Schmutz *et al*., 2010). This assembly applied a whole-genome shotgun approach complemented by Sanger-sequenced bacterial artificial chromosomes (BACs), and comprised 950 Mbp in 20 pseudomolecules, plus 23.2 Mbp in 1148 unanchored scaffolds. A second assembly, Wm82.a2, was released by DOE-JGI in 2014 and comprised 949 Mbp in 20 pseudomolecules, plus 29.3 Mbp in 1170 unanchored scaffolds (Song *et al*., 2016). Additional *ab initio* (Salamov and Solovyev, 2000) soybean genome assemblies have been released recently: Zhonghuang 13 (Shen *et al*., 2018); Enrei (Shimomura *et al*., 2015); a perennial relative of soybean, *Glycine latifolia* (Liu *et al*., 2018); draft assemblies for seven wild soybean accessions (Li *et al*., 2014); and a high-quality *G. soja* assembly for accession W05 (Xie *et al*., 2019).

These genome assemblies are being used to further understand soybean biology and to accelerate breeding. A small sampling of the many studies that have used the soybean reference assembly includes: population structure and ancestry (Bandillo *et al*., 2015; Zhou *et al*., 2015; Valliyodan *et al*., 2016); identification of the locus determining pod shattering (Dong *et al*., 2014; Funatsuki *et al*., 2014); seed protein content (Hwang *et al*., 2014; Vaughn *et al*., 2014; Bandillo *et al*., 2015); plant architecture (Prince *et al*., 2019); precision gene editing (Curtin *et al*., 2018); and genomic selection (Desta and Ortiz, 2014). Recent advances in next-generation sequencing technologies, including long-read sequencing, long-range scaffolding and advanced bioinformatics for short read assembly (Burton *et al*., 2013), have led to the production of improved assemblies for even large and highly complex genomes (Jiao *et al*., 2017; Ling *et al*., 2018; Raymond *et al*., 2018).

Here, we report three reference-quality *de novo* assemblies, using a combination of short- and long-read technologies, for three accessions: an improved assembly of the northern US accession Wm82; an assembly for the southern US accession Lee (PI 548656); and an assembly for *G. soja* accession PI 483463. All are assembled into pseudomolecules, with scaffolds anchored using a combination of optical maps, high-density genetic maps, and reciprocal structural comparisons between the three genomes. The three assemblies have complementary characteristics, with the Lee assembly being the largest overall (at 1.016 Gbp, approximately 3% bigger than the other two), the Wm82 assembly having the highest contiguity at the scaffold and contig levels (with scaffold N50 of 20 Mbp and contig N50 of 419 kbp), and the *G. soja* assembly providing a useful reference for undomesticated gene forms. Together, these assemblies provide a resource to advance soybean biology and breeding. To illustrate the utility of multiple high-quality genome assemblies, particularly including a wild *G. soja* accession, we examined alleles at loci with established domestication-related functions in soybean, including loci involved in pod dehiscence, determinacy, seed coat color and hard seededness.

RESULTS AND DISCUSSION

Similarity comparisons relative to the US soybean germplasm collection

To determine the similarity between the three genomes described in this study relative to other soybean accessions, we compared these genotypes with the rest of the United States Department of Agriculture (USDA) germplasm collection using two methods: first, using an overall similarity metric between accessions in this study and those in the SoySNP50k genotype matrix (Song *et al*., 2013); and second, using a phylogenetic tree calculated from the SoySNP50k matrix.

Wm82 shows high similarity (≥99%) with 40 other lines. There are 137 accessions with at least 90% similarity, and the median similarity value for Wm82 is 0.628 relative to all other accessions in the US collection (Table S1). Lee shows high similarity (≥99%) to just one other line in the US collection:PI 567789, which is reported to be a mutant of Lee. There are 11 accessions with at least 90% similarity, and the median similarity value for Lee is 0.678 relative to all other accessions in the US collection (notably higher than the similarity value for Wm82) (Table S1). *Glycine soja* PI 483463 is distinct from all other accessions in the US collection: the closest match (PI 597451A) has 90%

similarity, and the median similarity value for *G. soja* PI 483463 is 0.5435 (Table S1). In comparison with each other, Lee and Wm82 are 67.4% similar, Wm82 and *G. soja* PI 483463 are 48.5% similar, and Lee and *G. soja* PI 483463 are 52% similar.

Phylogenetic analysis relative to the US soybean germplasm collection

A phylogeny of the (sampled) US germplasm collection and the accessions described in this paper shows several striking features (Appendixes S1–S3; Figures 1 and S1). The tree is rooted between *G. max* and *G. soja* accessions (top and bottom clades, respectively). Colors indicate countries of origin: yellow, Indonesia and Vietnam; blue, China; orange, South Korea and North Korea; green, Japan; cyan, USA; pink, Brazil; and gray for all others (primarily Russia and India). Most of the clades are predominantly comprised of accessions from particular geographic locations (indicated by countries). Chinese accessions mostly fall into two clades: the large upperclade, with accessions from Southeast Asian lines from Indonesia and Vietnam; and the lower clade, with two nested clades of US accessions. Accessions from North Korea and South Korea mostly fall into one clade, as do those from Japan. Interestingly, the *G. soja* accessions also fall into three clades, by geographic origin: China, North and South Korea, and Japan.

The accessions from this study, as well as some related accessions, are highlighted: Lee and its parents CNS and S-100 are indicated with green text and icons; Wm82 and its progenitors are indicated with red text and icons; and PI 483463 is indicated with violet text and icon. Wm82 and progenitors all fall within the clade of US accessions in the lower clade of Chinese origin, with the exception of Kingwa, which was used in a cross with Williams as the basis for Phytophthora resistance. The Lee progenitors fall into two clades: the large Chinese-dominated clade and a clade of US and Brazilian lines, apparently deriving from an older Japanese lineage. This analysis suggests why the average similarity scores against the US germplasm collection are higher for Lee than for Wm82 (Appendixes S1–S3; Figures 1 and S1). Although Wm82 has high similarity with a large group of US accessions, it is rather different from accessions in the parent clade of Chinese accessions, and is further separated from all other accessions in the large Chinese, Korean and Japanese lines at the top of Figures 1 and S1, where Lee is found.

Comparisons of the three assemblies with the 20 087 accessions in the US germplasm collection show high similarity between Wm82 and many other accessions (with 40 accessions having >99% similarity and 137 accessions having ≥90% similarity). This is not surprising considering the importance of Wm82 in research and breeding programs. It is perhaps surprising that Lee, which has also been used widely in southern US breeding programs, shares high similarity with relatively few accessions in the US collection (with only 11 accessions having ≥90% similarity), yet Lee has a higher overall similarity relative to the US collection, compared with Wm82 (median similarity values are 0.628 for Wm82 and 0.67 for Lee). A phylogenetic analysis based on genotype data suggests that the greater median similarity between Lee and the rest of the US collection reflects characteristics of genotype representation in the US collection: namely, that the collection has extensive representation from southern China and Southeast Asia (upper blue plus yellow clades in Figures 1 and S1), North Korea and South Korea, and Japan. In contrast, Wm82 and other northern US cultivars come from a clade of northern Chinese origin (blue and gray clades at bottom of Figures 1 and S1; with gray mostly representing accessions from the Vavilov Institute, http://www.vir.nw.ru). As expected, *G. soja* accession PI 483463 nests within the clade of other *G. soja* accessions (bottom clade in Figures 1 and S1), but sits on a relatively long branch among Chinese *G. soja* accessions, consistent with its phenotypic character as a relative outlier among the available *G. soja* lines, showing unusual salt tolerance (Lee *et al*., 2009; Valliyodan *et al*., 2017).

Germplasm has evidently been under regional selection (Figure 1), probably in relative genetic isolation from other groups. This is suggested by distinct per-country clades. At the same time, breeding efforts also clearly involve periodic wide crosses, as well as the occasional movement of germplasm across country borders. The parents of Lee come from two distinct clades (one predominantly of Chinese origin and another probably of Japanese origin), and Wm82 has genetic material from two distinct clades (with a parent Kingwa from one and Williams from another clade).

Genome assembly and assessment

The *G. max* cultivar Lee and *G. soja* PI 483463 were assembled using a similar approach and combination of technologies, based primarily on NRGene DeNovoMagic assemblies (Avni *et al*., 2017; Springer *et al*., 2018) that usedpaired-end reads of 160–260 bp from size-selected libraries, followed by scaffolding using a two-enzyme Bionano optical map, and manual evaluation and integration into pseudomolecules. The Lee pseudomolecules span 990.7 Mbp, with an additional 25.6 Mbp in 245 unanchored scaffolds, for a total assembly size of 1016.3 Mbp, whereas *G. soja* PI 483463 pseudomolecules span 962.3 Mbp, with 22.9 Mbp in 286 unanchored scaffolds, and with a total assembly size of 985.2 Mbp (Table 1). The Wm82 version 4 assembly (Wm82v4) builds on the widelyused assembly version 2, as well as an incremental version 3 that involved the incorporation of BAC sequences to fill contig gaps in 2016. The Wm82v2 assembly was primarily Sanger-based, and gap-filling in v3 and

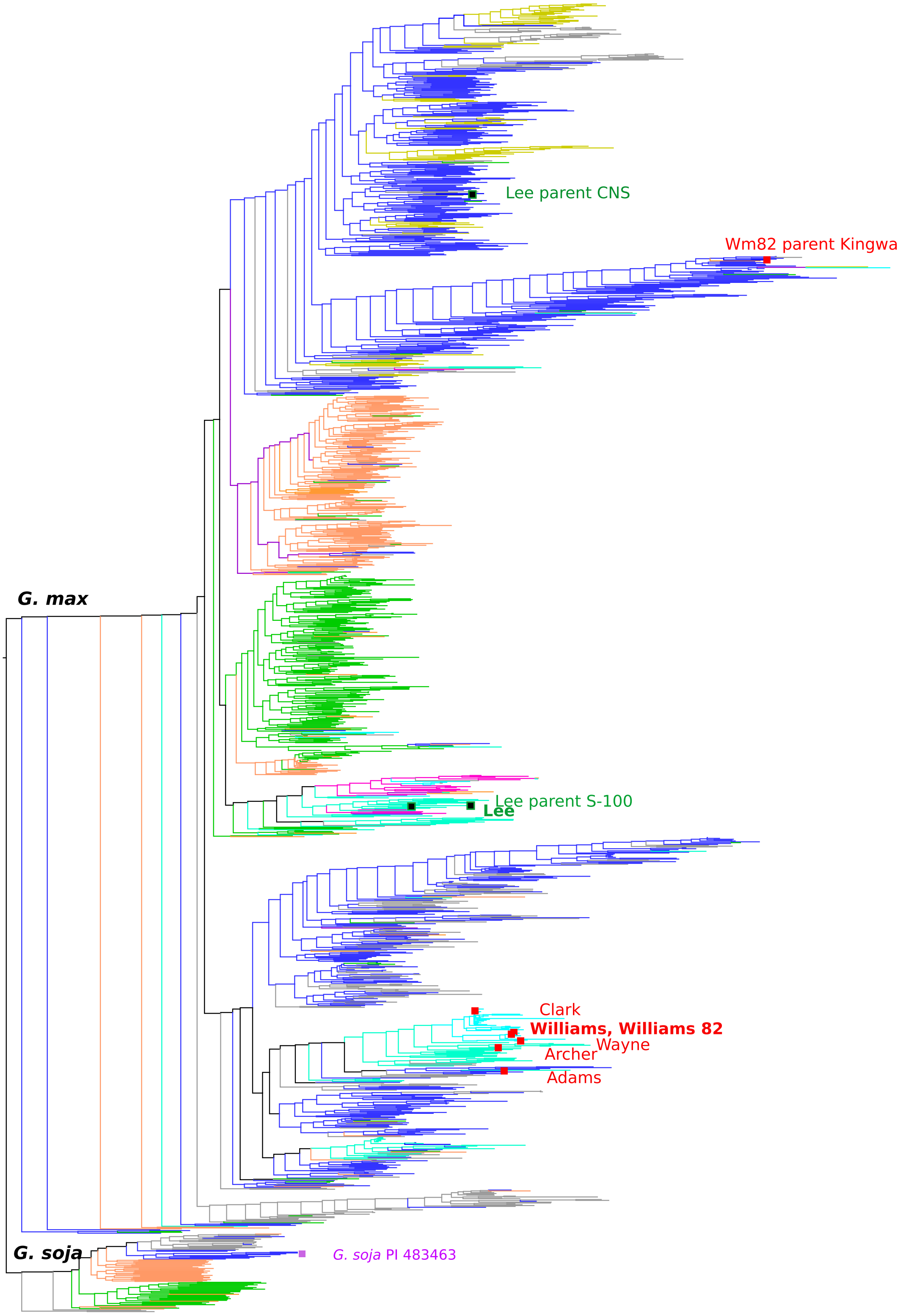

Figure 1. Phylogenetic tree of a random sampling of the US germplasm collection for *Glycine soja* and *Glycine max*. The tree is rooted between *G. max* and *G. soja* accessions, as indicated. Colors indicate countries of origin: yellow, Indonesia and Vietnam; blue, China; orange, South Korea and North Korea; green, Japan; cyan, US; pink, Brazil; gray, all others (predominantly from Russia and India). Country and ID correspondences, as well as tree order, are shown in Appendix S4. Accessions Lee and Wm82 and their immediate progenitors are indicated in green and red text and icons, and *G. soja* PI 483463 is indicated in violet text and icon.

v4 used PacBio-based BAC assemblies targeted to the gap regions. The Wm82v4 assembly closed 3626 gaps and added 5 138 978 bp of sequence relative to Wm82v2, increasing the contig N50 from 233.1 to 419.3 kbp. The Wm82v4 assembly has a total assembly size of 978 Mbp: 961 Mbp in pseudomolecules and 17 Mbp in unanchored scaffolds.

The three assemblies have complementary strengths. The assemblies are closely matched in size: each is within 4% of the total assembly size of the others, and represents 88–91% of the predicted genome size of 1115 Mbp (Arumuganathan *et al*., 1991). In terms of scaffold and contig sizes, the Wm82v4 assembly is the most complete and contiguous, with scaffold N50s being approximately 20 Mbp compared with 14 and 4 Mbp for Lee and *G. soja*, respectively, and with contig N50 being approximately 420 kbp compared with 41 kbp and 28 kbp for Lee and *G. soja*, respectively (Table 1). In general, the scaffold boundaries are different in each assembly, providing an indication of the genomic content in the gap regions in each assembly, and providing a means of assessing scaffold placement and orientation (Figures S2–S4).

Assembly completeness in terms of gene, telomere and centromere capture

Assessing assembly completeness, analyses with CEGMA (v2.5) and BUSCO (Parra *et al*., 2007; Simao *et al*., 2015) demonstrate similar scores, with 4/248 (1.6%) of CEGMA genes missing from each assembly and 79–85/1440 (5.5–5.9%) of BUSCO genes missing from each assembly. This suggests that the gene content was effectively captured in each assembly (Table 2). The Lee and Wm82v4 assemblies were similar in their read realignment rates, with the Wm82v4 assembly showing fewer regions with no reads aligning (two-sided Student's *t*-test, $P < 0.05$), and fewer repetitive regions (Wilcoxon rank sum test, $P < 0.05$), indicating that this has Wm82v4has the highest quality of the three assemblies (Appendix S4; Figure 2; Table S2). Another indication of assembly completeness is the proportion of pseudomolecules that extend into the telomeric repeats at the chromosome ends. The Wm82v4 assembly has telomeric repeats on 26 of the 40 pseudomolecule ends, whereasthe Lee and the *G. soja* assemblies have telomeric repeats on 22 and 18 pseudomolecule ends, respectively.

Soybean has two characteristic centromeric repeat variants: CentGm-1 and CentGm-2, which are 92 and 91 bases long, respectively (Gill *et al*., 2009). Although their sequences are 85% identical, they are sufficiently different to identify distinct arrays on different chromosomes, both on the basis of sequence alignments in pseudomolecule assemblies and by fluorescently labeling these two repeats as fluorescence *in situ* hybridization (FISH) probes (Gill *et al*., 2009; Findley *et al*., 2010). It has been speculated that these repeats may have diverged in distinct *Glycine* species, prior to an allopolyploidy event in *Glycine* approximately 10 Mya (Gill *et al*., 2009). It is unclear how a hypothetical autopolyploidy event (Wang *et al*., 2017) could have impacted centromeric repeat diversity.

Centromeres assembled in all pseudomolecules (Table 3). For a given chromosome, the repeat signatures are similar across the three assemblies; however, the number of repeats and repeat-class ratios differ between chromosomes within an assembly. For example, the number of CentGm-1 repeats is high across all three assemblies for chromosomes 5, 12, 15 and 20, whereas CentGm-2 repeats are absent or near absent for these chromosomes. In contrast, CentGm2 repeats are observed in the centromeres of all three assemblies for chromosomes 6, 7 and 11. Several chromosomes show mixtures of repeat signatures in all three assemblies, and chromosome 1 has a very low abundance of both repeat classes in all three assemblies (Figures S5–S7). All assemblies showed very similar transposon and repeat content (Figures 3, 4 and S5–S7), and the copy number for *copia* is roughly half of that for *gypsy*, as has been previously observed (Du *et al*., 2010; Tian *et al*., 2012; Li *et al*., 2014).

Structural assessments and comparisons

There are few major structural differences observed between the three genome assemblies (Figures S2–S4); however, small inversions (500–62 045 bp) are frequent,

Table 1 Genome accessions and assembly statistics. Counts are perbase, excluding between-scaffold gaps or between-contig gaps for the indicated statistics

	Glycine max Lee	*Glycine max* Wm82v4	*Glycine soja* PI 483463
GenBank accession no.	GCA_002905335.1	PRJNA48389	GCA_002907465.1
Total assembly size (bp)	1 016 275 704	978 386 919	985 259 765
Pseudomolecules (bp)	990 714 026	961 401 624	962 330 378
Remaining scaffolds (bp)	25 561 678	16 985 295	22 929 387
Remaining scaffolds (count)	272	262	286
Scaffold N50 (bp)	15 020 773	20 441 467	4 430 511
Contig N50 (bp)	37 725	419 290	11 571
Nucleotides (bp)	43 188 592	25 907 548	33 895 040
Ns (percent)	4.25%	2.65%	3.44%

Table 2 Assessments of assembly and annotation coverage in terms of conserved genes captured. Both the CEGMA and BUSCO methods (Core Eukaryotic Genes Mapping Approach and Benchmarking Universal Single-Copy Orthologs, respectively) check for collections of highly conserved protein sequences. For BUSCO, results are reported for the proportion of genes found in the primary assemblies and also for the predicted genes (annotation)

	Subject	*Glycine max* Lee	*Glycine max* Wm82v4	*Glycine soja* PI 483463
CEGMA	Complete	224 (90.32%)	224 (90.32%)	226 (91.13%)
	Complete: single copy	22 (8.87%)	19 (7.66%)	23 (9.27%)
	Complete: duplicated	202 (81.45%)	205 (82.66%)	203 (81.86%)
	Fragmented	20 (8.06%)	20 (8.06%)	18 (7.26%)
	Missing	4 (1.61%)	4 (1.61%)	4 (1.61%)
	Total CEGs searched	248	248	248
BUSCO–genomes	Complete	1342 (93.2%)	1342 (93.2%)	1348 (93.6%)
	Complete: single copy	682 (47.4%)	679 (47.2%)	676 (46.9%)
	Complete: duplicated	660 (45.8%)	663 (46.0%)	672 (46.7%)
	Fragmented	13 (0.9%)	14 (1.0%)	13 (0.9%)
	Missing	85 (5.9%)	84 (5.8%)	79 (5.5%)
	Total BUSCO searched	1440	1440	1440
BUSCO–genes	Complete	1401 (97.3%)	1405 (97.6%)	1393 (96.7%)
	Complete: single copy	569 (39.5%)	588 (40.8%)	568 (39.4%)
	Complete: duplicated	832 (57.8%)	817 (56.7%)	825 (57.3%)
	Fragmented	10 (0.7%)	7 (0.5%)	12 (0.8%)
	Missing	29 (2.0%)	28 (1.9%)	35 (2.4%)
	Total BUSCO searched	1440	1440	1440

with approximately 40–42 inversions per chromosome between either Lee or Wm82v4 and *G. soja*, and approximately 32 inversions per chromosome between Wm82 and Lee. Considering inversions larger than 500 bp, the most frequent variations between assemblies were in the size range of 1–2 kbp, with the largest inversion detected being 62 kbp (detected in both Lee and Wm82, with respect to *G. soja*). A prominent classical gene in soybean, the I locus (controlling seed coat color), is the result of once such inversion, described below under the section on domestication gene analysis.

The three assemblies were assembled into pseudomolecules using two dense high-resolution genetic maps: one with 11 922 markers and another with 21 478 markers (Song *et al.*, 2016). In plots of genetic distance by physical (genomic) distance, most chromosomes show high recombination rates in the gene-rich and chromosome arms, and low recombination rates in the gene-poor, transposon-rich pericentromeric regions (Figures S8–S10). Exceptions are seen on acrocentric chromosome 9, in which the pericentromeric region is found on the leading chromosome arm. There are no substantial deviations in the plots between the three assemblies for corresponding chromosomes.

Comparative gene content

Genome annotation predicted 71 358 transcripts for 47 649 genes in the final Lee assembly, 86 256 transcripts for 52 872 genes in the Wm82v4 assembly and 62 102 transcripts for 46 969 genes in the *G. soja* assembly (Table 4). The differences between annotations are most likely due primarily to the differing gene expression resources used in the annotation pipelines for the three assemblies, with the Wm84v4 annotation using more transcript sequence data, by an order of magnitude, than the other two assemblies, from more diverse tissue libraries, and 2.6 million full-length Iso-Seq transcripts (4.8 billion read-pairs of RNA-seq reads and 2.6 million Iso-Seq reads for Wm82v4, vs. 89 million 2x150 read-pairs for *G. soja* and 180 million 2x150 read-pairs for *G. max* Lee).

Assessing annotation completeness, applying BUSCO (Simao *et al.*, 2015) to predicted genes (rather than to assemblies, presented above) identifies that annotation completeness for the three assemblies ranged from 97.3% to 97.7%, with 0.5% to 0.8% fragmented genes and 1.9% to 2.4% missing (Table 2).

To facilitate comparisons among the assemblies, we identified sets of genes that correspond both by homology (top blastn match of each gene sequence from a query assembly to the comparison genome assembly, at ≥95% identity) and by chromosomal position (overlapping gene models, on the corresponding chromosome, from top GMAP match (Wu and Watanabe, 2005) of each gene between the assemblies). The resulting orthogroups (Table S3) can be considered a genic pan-genome for this set of assemblies and annotations. These orthogroups should also be useful for researchers who wish to translate between the assemblies: for example, to find the gene in Wm82v4 that corresponds with the genes in Wm82v2, Lee or *G. soja*.

The pan-gene analysis identifies 50 686 clusters, each consisting of corresponding gene models from at least two of the four annotation sets. Of these, 41 324 sets have exactly one gene model from each assembly, whereas

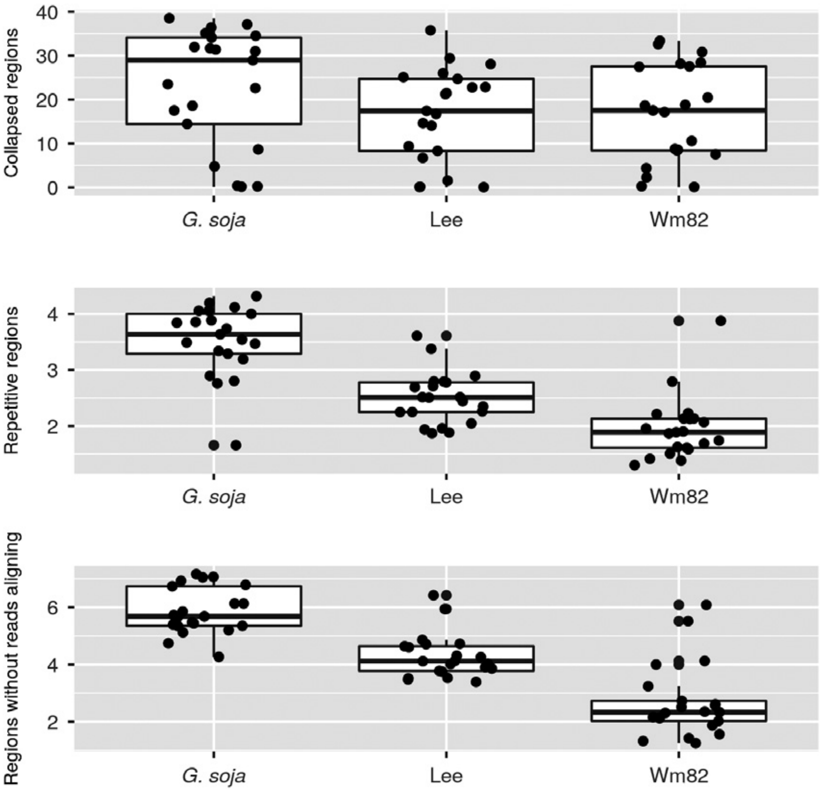

Figure 2. Repetitive and collapsed regions, and regions without reads aligning, for the three assemblies. Each dot is one pseudomolecule. Panels: collapsed regions as percentage of chromosome; repetitive regions as a percentage of chromosome; and regions without reads aligning as a percentage of chromosome. Values in each case were calculated as proportions of reads aligning to each assembly.

Table 3 Counts of characteristic centromeric repeats in the *Glycine* assemblies, by chromosome. Counts for each repeat type were calculated based on the top competitive match (BLASTN) between the two repeat types against genomic windows of 1 kb, and summarized by chromosome

	CentGm-1			CentGm-2			C.Gm-1 Average	C.Gm-2 Average
	G. soja	Lee	Wm82	*G. soja*	Lee	Wm82		
Gm01	0	58	0	0	7	3	19	3
Gm02	2069	3980	896	61	73	25	2315	53
Gm03	237	172	130	241	0	0	180	80
Gm04	642	712	279	186	402	47	544	212
Gm05	3437	2724	1121	0	0	0	2427	0
Gm06	12	49	36	1298	2751	1003	32	1684
Gm07	78	41	12	808	455	261	44	508
Gm08	69	934	17	0	0	0	340	0
Gm09	208	495	346	45	95	44	350	61
Gm10	57	32	24	26	267	78	38	124
Gm11	28	30	21	471	400	264	26	378
Gm12	1899	2237	1023	0	5	0	1720	2
Gm13	488	262	302	1	1	0	351	1
Gm14	256	199	499	13	0	1	318	5
Gm15	2255	4995	1540	0	4	22	2930	9
Gm16	851	1528	603	142	40	32	994	71
Gm17	179	206	63	353	472	7	149	277
Gm18	723	1468	230	0	5	0	807	2
Gm19	460	194	341	31	79	76	332	62
Gm20	2538	6001	2476	0	0	7	3672	2
Sum	16 486	26 317	9959	3676	5056	1870	17 587	3534

3295 have fewer than four members and 2298 have more than four members. Orthogroups that have more than or fewer than four orthologous genes across the four assemblies are enriched for gene families known to occur in genomic clusters. Among families with five or more genes per orthogroup (i.e. above the expected four for

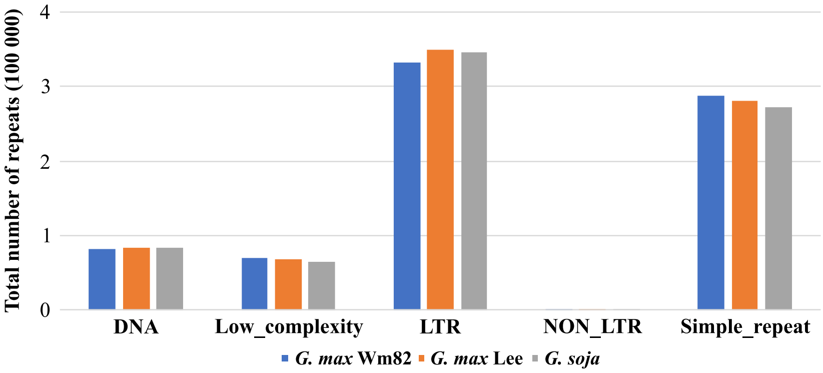

Figure 3. Total number of repeats in each repeat class. The *x*-axis is the name of each repeat class and the *y*-axis is the total number of repeats in each repeat class. Total numbers of repeats in *Glycine max* Wm82 are given by blue bars, total numbers of repeats in *G. max* Lee are given by orange bars and total numbers of repeats in *Glycine soja* are given by blue bars.

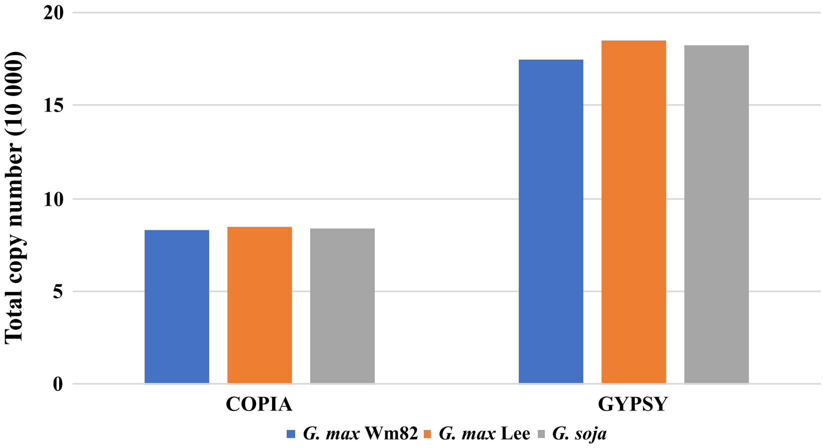

Figure 4. Total copy number of *copia* and *gypsy* retrotransposons. The *y*-axis is the total copy number of the retrotransposons. The left three bars represent the total copy number of *copia*, and right three bars represent the total copy number of *gypsy*. Total copy numbers in *Glycine max* Wm82 are given byblue bars, total copy numbers in *G. max* Lee are given by orange bars and total copy numbers in *Glycine soja* are given by gray bars.

Table 4 Gene prediction and gene clustering results for all assemblies. Counts of genes within clusters are based on OrthoFinder clusters of protein sequences among the three annotation sets. Counts of genes with protein domains are based on INTERPROSCAN matches in the Pfam database

	Glycine max Lee	*Glycine max* Wm82v4	*Glycine soja* PI 483463
Primary transcripts (genes)	47 649	52 872	46 969
All transcripts	71 358	86 256	62 102
Average primary transcript length (aa)	412.3	385.2	412.7
Genes within clusters (%)	46 598 (97.8%)	47 562 (81.2%)	45 403 (96.7%)
Genes with protein domains (%)	39 402 (82.7%)	42 369 (72.3%)	38 954 (82.9%)

conserved single-copy genes), the most frequent families are: tetratricopeptide repeat (TPR)-like genes; nucleotide-binding and leucine-rich repeat genes (NBS-LRR) disease resistance genes; SMALL AUXIN UP RNA (SAUR)-like auxin-responsive proteins; protein kinase superfamily proteins; F-box proteins; cytochrome P450 proteins; and chalcone- and stilbene-synthase family proteins. The families with fewer than four genes per orthogroup overlap substantially with those in the ‘larger than expected’ orthogroups, consistent with a pattern of copy-number increases and decreases occurring in local genomic clusters. Among families with decreased copy numbers in the orthogroups (i.e. orthogroups of three or fewer genes among the four assemblies), the most frequent families are: NBS-LRR disease resistance genes; TTF-type zinc-finger proteins; protein kinase superfamily proteins; PIF1 helicase proteins; α/β-hydrolases superfamily proteins; SAUR-like auxin-responsive proteins; nodulin MtN21 proteins; and pentatricopeptide repeat (PPR) proteins. This comparison of annotations is also supported by gene ontotology (GO)-term enrichment analysis, which shows the strongest biological process responses for defense-response activity (GO:0006952) and several terms related to NBS-LRR genes (Table S4).

We also identified unique matches with predicted genes from any of the assemblies (≥95% identity and ≥85% query coverage in BLAT [Kent, 2002] output) on the corresponding chromosome from which each gene was predicted. This is a more lenient method than the pan-gene method described above, as a search of the genes against the genomes may find pseudogenes or gene models not identified by gene-modeling software. The combined gene queries identified58 762, 58 814 and 58 292 distinct genic regions in Lee, Wm82v4, and *G. soja*, respectively. We finally used annotation LiftOver procedures for all possible pairs of the three annotations to find genes unique to each assembly. This identified 70, 256 and 219 genes with no hits in the other two assemblies for Lee, Wm82v4, and *G. soja*, respectively.

The differences may arise from the greater completeness of the Wm82v4 assembly, although the differences are small (52 more near-identical gene-homologous regions in Wm82v4 than in Lee, and 522 more in Wm82v4 than in *G. soja*) and may reflect true gene presence/absence variation, as observed in many species (Golicz *et al.*, 2016).

Disease resistance gene content

The primary transcripts of the three annotations were mined for resistance gene analogs (also called resistance-gene homologs, RGHs) using RGAugury (Li *et al.*, 2016), and split into three groups: receptor-like kinases (RLKs; TM + LRR/LysM + STTK domain), receptor-like proteins (RLPs; TM + LRR/LysM), and NBS-LRR genes (TIR/Coils + NB-ARC domain + Leucine-Rich-Repeat). The Wm82 annotation contains the most RGH candidates (1886), followed by Lee (1776) and *G. soja* (1750). Wm82 also contained the most NBS-LRR genes (448), followed again by Lee (442) and *G. soja* (419) (Table 5). In most categories of NBS-LRR genes, Wm82 contained the most candidates, except in Coils + NB-ARC + Leucine rich repeats (123 in Lee, 110 in Wm82 and113 in *G. soja*), and 'other' R-genes (20 in Lee, 18 in Wm82 and17 in *G. soja*) (Table 5). The observed differences may also arise from pseudogenization at some loci or from missed gene models where expression support was lacking or where certain features of the gene model were non-standard.

SNP comparisons among the assemblies

In nucleotide comparisons of the Lee, Wm82 and *G. soja* assemblies, the SNP densities between the genotypes were calculated as SNPs per kb. The following SNP densities were observed: 1.7 for Lee versus Wm82v4, 0.13 for Wm82v2 versus Wm02v4, and 4.7 for both Lee and Wm82 versus *G. soja* PI 483463. These rates are consistent with the rates reported by Hyten *et al.* (2006), who reported nucleotide diversity (π) among elite lines of approximately 1.1 per kb, and among *G. soja* accessions of approximately 4.7 per kb. Fine-scale positional comparisons uncovered several striking differences. Figure 5(a) shows comparisons between the Wm82v2 and the Wm82v4 assemblies (red histograms on the left-hand side of each chromosome backbone), and between the Wm82v4 and Lee assemblies (blue histograms on the right-hand side of each chromosome backbone).

Table 5 Resistance gene candidates found in the three annotations

Class	*Glycine max* Lee	*Glycine max* Wm82v4	*Glycine soja* PI 483463
CN	8	11	7
CNL	123	110	113
NBS	29	34	26
NL	92	105	94
OTHER	20	18	17
TN	22	24	22
TNL	99	101	85
TX	49	45	55
Total NLR	442	448	419
RLK	1164	1197	1146
RLP	170	241	185
Total	1776	1886	1750

NBS, only NB-ARC domain; CN, Coils + NB-ARC; TN, TIR + NB-ARC; NL, NB-ARC + Leucine rich repeat; CNL, Coils + NB-ARC + Leucine rich repeat; TNL, TIR + NB-ARC + Leucine rich repeat domain; TX, TIR + unknown domain; OTHER, TIR + Coils domain (missing NB-ARC); RLP, receptor-like protein; RLK, receptor-like protein kinase.

In the comparison between the two Wm82 assembly versions, most genomic regions show very low levels of nucleotide differences, with the exception of regions on Gm03, Gm07, Gm12 and Gm14 (Figure 5a, red histograms). The SNPs observed between Wm82v2 and Wm82v4 (at a genome-wide rate of 0.13 per kb, with 31% of the differences occurring on Gm03 and Gm07) are likely to be caused by differential levels of introgression of the Kingwa parent that was used in breeding Wm82, as reported by Haun *et al.* (2011). Kingwa was selected as a breeding parent of Wm82for its resistance to *Phytophthora sojae* (Dorrance *et al.*, 2004; Gao and Bhattacharyya, 2008), conferred by the locus Rps_1^k. This locus, located on chromosome 3, was introgressed by an initial cross between Williams and Kingwa, followed by multiple crosses to the recurrent parent Williams to recover most of the elite background (while maintaining the KingwaRps_1^k locus) (Bernard and Cremeens, 1988). The largest differences between the two Wm82 assembly versions are observed near this locus. This is explained by the different Wm82 sources used to generate the two assemblies. The initial assemblies of the Wm82 genome were based on DNA samples that came from multiple different individuals of Wm82, rather than a single individual (Haun *et al.*, 2011). Not all Wm82 plants have identical introgression of the Rps_1^klocus (or other loci that were introgressed during the breeding of Wm82). Therefore, these regions of Gm03, Gm07, Gm12 and Gm14 were assembled by reads from

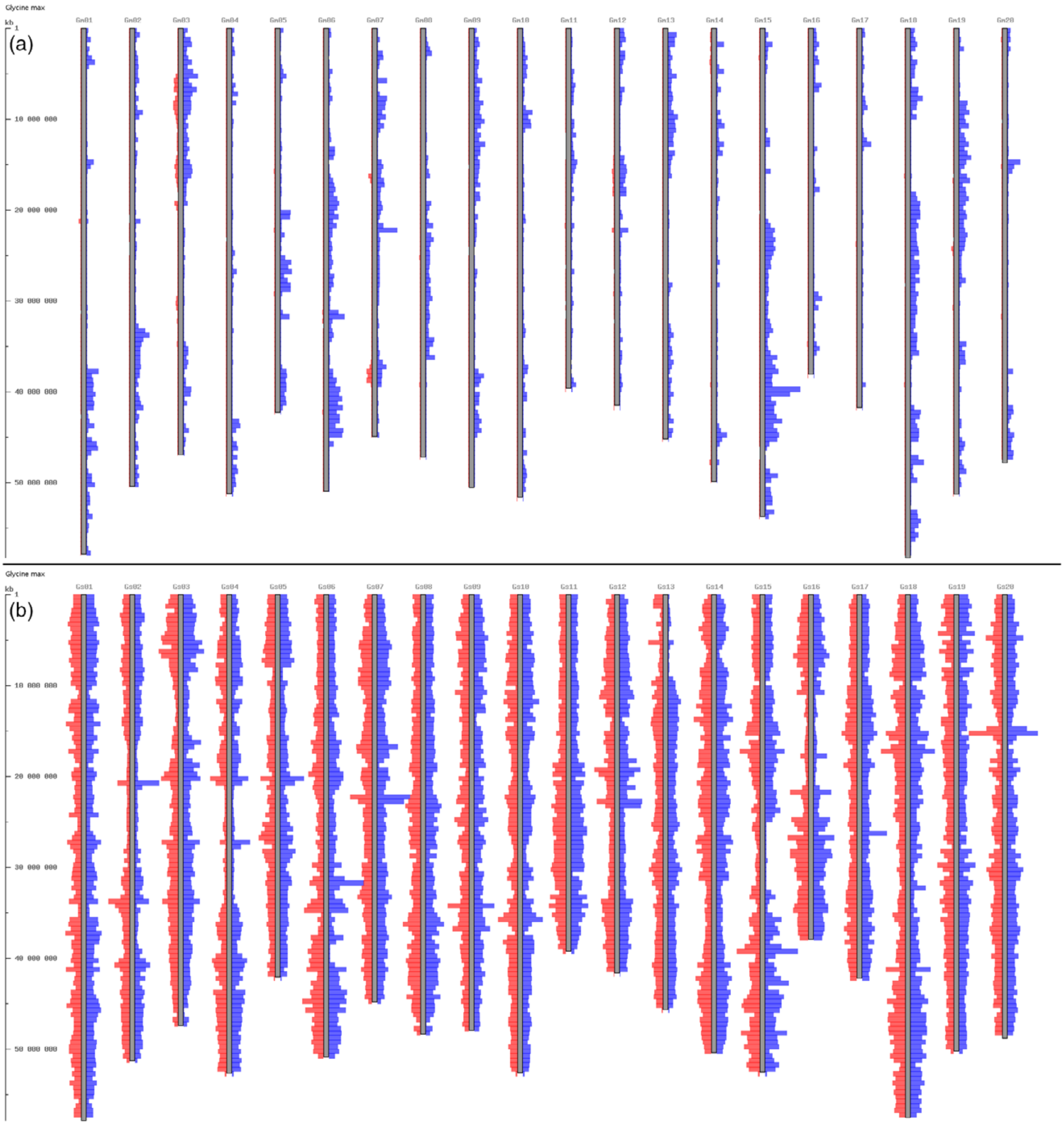

Figure 5. Single-nucleotide polymorphism (SNP) densities between assemblies Wm82v2, Wm82v4, Lee and *Glycine soja* PI 483463. (a) Red histograms (extending leftward from each chromosome) show differences between Wm82v2 and Wm82v4 (with Wm82v4 providing the reference assembly coordinates), whereas blue histograms (extending rightwards) show differences between Wm82v4 and Lee. (b) Red histograms show differences between Wm82v4 and *G. soja* PI 483463, whereas blue histograms show differences between *G. soja* PI 483463 and Lee. Histogram bin sizes are 500 kb for both panels. Only SNP variants are shown (excluding indels and missing data).

several 'Wm82' subtypes, differing in genetic composition in the introgression regions, resulting in a mosaic of Kingwa and Williams reads in those regions. This remained the case in Wm82v2. In Wm82v4, however, in order to better represent this known highlyinbred accession, we used additional sequence reads from the single haplotype of Wm82-ISU-01 in this assembly, thereby changing many of these SNPs to now match either the Kingwa or Williams haplotype throughout any given region (instead of a mosaic of the two).

In the blue histograms on the right-hand side of each chromosome (Figure 5a), showing differences between the

Lee and Wm82v4 assemblies, most chromosomal regions show large differences between the two cultivars, with notable exceptions in particular regions (e.g. the center of Gm01 or distal regions of Gm06, Gm08 andGm09). These appear to be shared haplotypes (regions of identity and therefore shared ancestral history).

Comparisons of Wm82 and Lee against the *G. soja* PI 483463 assembly (Figure 5b) show high levels of difference across almost all regions of all chromosomes, with a few small exceptions. On Gm06 (approximately 38–41 Mbp) and Gm15 (approximately 25–33 Mbp) there are regions of nearidentity with *G. soja*. There are no such large regions of near identity between Wm82 and *G. soja* PI 483463. Although Wm82 does have regions of introgression with respect to *G. soja*, there is considerable diversity among *G. soja* germplasm (Hyten *et al*., 2006; Li *et al*., 2014), and the known introgression regions (particularly Gm03) are not evident with respect to this particular *G. soja* accession.

Parentage analysis of the Lee and Wm82 assemblies

Nucleotide-level characteristics of the assemblies can also be seen in plots of SNP comparisons between the parents of the sequenced cultivars. Figure 6(a,b) shows the

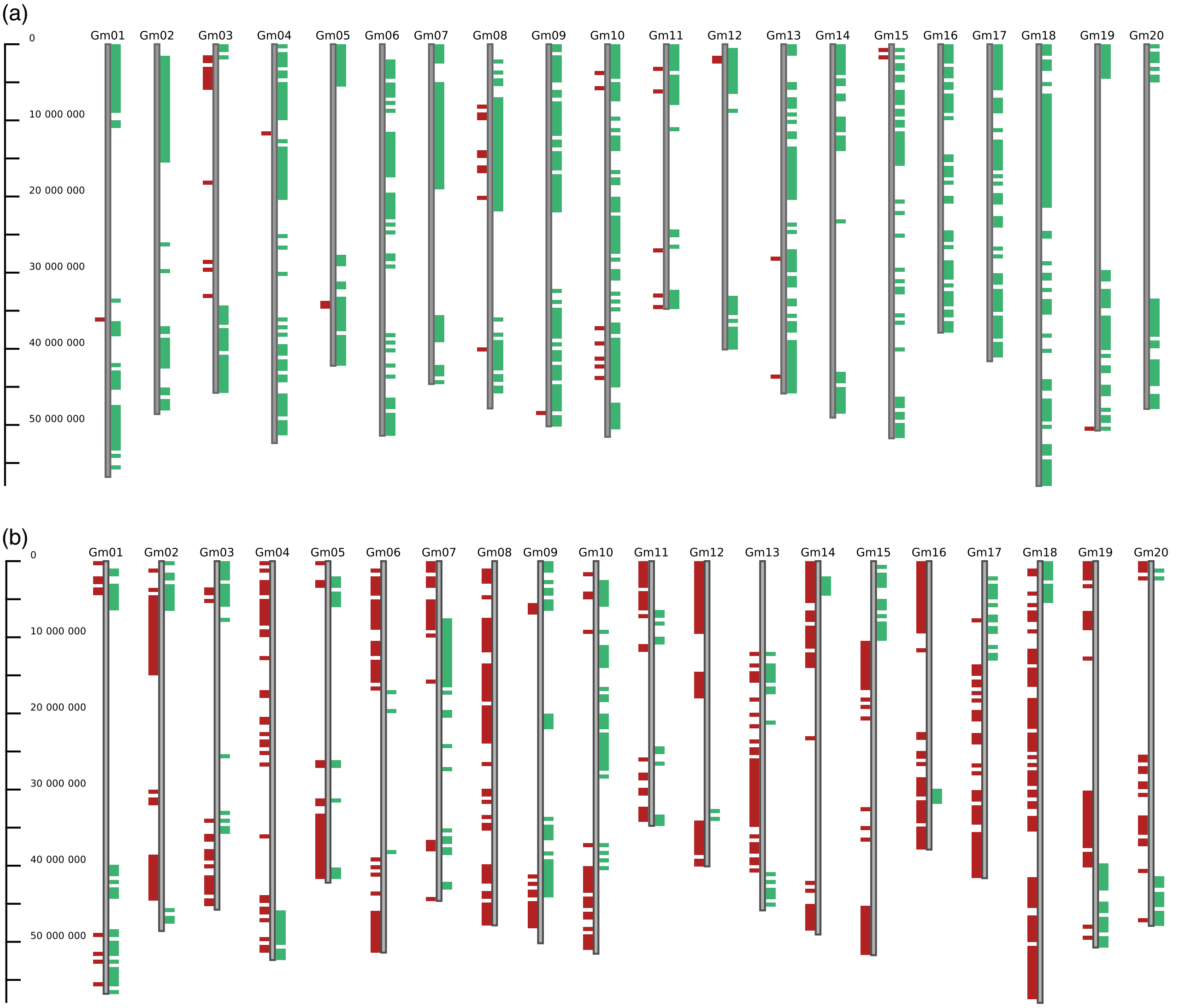

Figure 6. Single-nucleotide polymorphism (SNP) locations for comparisons between assemblies Wm82v2, Lee and their parents. A colored bar indicates when at least five SNPs per 500 kb are present between the indicated parent genotype and the comparison genome. SNPs are from the Soysnp50k array (Song *et al*., 2013). (a) SNP locations plotted relative to the Wm82v2 assembly (which is the coordinate system in which the SNPs are reported). Leftward redbars: differences between the Williams parent and Wm82v2 assembly; rightward green bars: differences between the Kingwa parent and Wm82v2 assembly. (b) SNP differences between Lee (backbone) and its parents are shown: leftward bars show differences from CNS (PI 548445); rightward bars show differences from S-100 (PI 548488).

locations of SNPs from the SoySNP50k genotyping matrix (Song *et al*., 2013), plotted relative to the Wm82v2 assembly (which is the coordinate system in which the SNPs are reported). Bars are indicated where there are at least five SNPs per 500 kb. In Figure 6(a), the green bars indicate relative SNP variation between Wm82v2 and Kingwa (the Rps_1^k donor parent), whereas the red bars indicate relative SNP variation between Wm82v2 and Williams (the recurrent parent). Clearly, Williams and Wm82v2 share nearly identical haplotypes throughout the vast majority of the genome; however, the region of Rps_1^k introgression on chromosome 3 shows a match to Kingwa and abundant SNP differences with Williams. This chromosome 3 region generally matches the major region of variation between the Wm82v2 and Wm82v4 assemblies, as discussed above for Figure 5(a).

In Figure 6(b), the haplotype structure is more typical of an early-generation cross between two distinct parents, with little recurrent backcrossing. The parents of Lee are S-100 (PI 548488) and CNS (PI 548445). Differences between Lee and CNS are shown in red (left), and differences between Lee and S-100 are shown in green (right). Some regions appear to be homozygous for one accession, e.g. for S-100 ('not CNS') on all of Gm08, or for CNS ('not S-100') across much of Gm10, whereas other regions are heterozygous (e.g. top of Gm14).

Domestication gene analysis

Loci responsible for at least two dozen important domestication- or diversification-related traits have been identified in soybean (Sedivy *et al*., 2017), and the pace of such discoveries appears to be increasing with technologies such as GWAS (for locus and allele identification) and CRISPR (for gene knock-out and functional tests). For domestication traits such as determinacy or podshatter (dehiscence), the trait typically involves an identifiable variant with respect to *G. soja*, which is the closest surrogate for the progenitor of the domesticated soybean. To test the utility of genome comparisons between cultivated and 'wild-type' accessions, we evaluated alleles for several important domestication traits in these three genomes.

Pod dehiscence. We examined two loci that condition pod dehiscence: SHAT1-5 (Gao and Zhu, 2013) and Pdh1 (Funatsuki *et al*., 2014). The SHAT1-5 locus was shown to play a prominent role in pod dehiscence in soybean, with most examined cultivated accessions having a loss-of-function allele, with a premature stop codon in the C-terminal end of the protein, 47 residues short of the wild-type gene (Gao and Zhu, 2013). The SHAT1-5 loci in Wm82 and Lee are on chromosome 7 starting at positions 4 314 874 and 4 398 874, respectively, and both contain the premature stop codon, likely conferring the reduction of pod shattering observed in the Wm82 and Lee cultivars.

Pdh1 also contributes to the reduction of pod shattering, independent of SHAT1-5 (Funatsuki *et al*., 2014). The Pdh1 shatter-resistant allele is present in Wm82v4, and this allele is distinguished from the wildtype by a stop codon on chromosome 16, 30 amino acids from the start of the genic sequence at 30 161 121 nt. A gene model was not predicted at this location in Wm82, presumably because of the early stop codon. The same allele is present in Lee on chromosome 16, also 30 amino acids from the start of the genic sequence at 31 656 849 nt. In contrast, in *G. soja*, the shatter-susceptible allelic form is present, and was called gene model GlysoPI483463.16G111700.1.

Hardseededness. The GmHs1-1 locus was shown by Sun *et al*. (2015) to confer hard seededness in the wild-type form, and softer seed coat and greater coat permeability in the domesticated variants. They identified a C→T point mutation in Glyma02g43700.1 (equivalent to Glyma.02g269500 in Wm82v2 and Wm82v4) between Wm82 and PI 479752, a *G. soja* accession. The authors showed through genetic, biochemical, and complementation tests that the C→T point mutation, resulting in a transition from threonine to methionine, is causal for the trait. We find that this same single mutation is present in both Lee and Wm82, that this is the only mutation in the coding sequence, and is thus likely the same causal/functional mutation as reported earlier (Sun *et al*., 2015).

Determinacy. The Arabidopsis Terminal flower 1 gene (*Tfl1*) was shown by Tian *et al*. (2010) to have four orthologs in soybean, with the paralog on chromosome 19 (Glyma19g37890 in Wm82v1; equivalent to Glyma.19g194300 in Wm82v2 and Wm82v4) having the largest effect on determinacy. We found that the Wm82 and *G. soja* PI 483463 Dt1 genes at chromosome 19 are identical, as might be expected, as both are indeterminate (although with greater viny-ness in the *G. soja* accession). The Lee allele differs, however, with a C→T mutation, resulting in a transition from proline to leucine. This mutation was identified by Tian *et al*. (2010) as the Gmtfl1-ab mutation, one of four functional missense mutations identified in this gene across the germplasm collection screened in that study. The closest paralog to GmTfl1, on chromosome 3 (Glyma03g35250 in Wm82v1 and Glyma.03G194700 in Wm82v2 and Wm82v4) is identical across the coding sequence for Wm82v4, Lee and *G. soja* PI 483463, so is unlikely to be causal for the determinacy differences observed between these accessions. We conclude that the C→T mutation in Lee is likely to be the same causal mutation as one of the four mutations identified earlier (Tian *et al*., 2010), and this is consistent with the indeterminacy in Wm82 and *G. soja* PI 483463 but with the determinacy in Lee.

Seed coat color. An example of phenotypic consequences from small inversions is seen in the classical I locus, which controls seed coat color in soybean (Wang *et al*., 1994; Clough *et al*., 2004; Tuteja *et al*., 2004, 2009; Cho *et al*., 2017). Although the full story of this complex locus is beyond the scope of this resource paper, the outline is that a cluster of approximately 12 chalcone synthase genes (depending on the accession) on chromosome 8, in an approximately 100-kbp region, has undergone rearrangements that result in the silencing of one of the key chalcone synthase genes. In domesticated accessions, a large inversion and adjacent duplications have rearranged these genes relative to the ancestral state in *G. soja*, bringing one chalcone synthase (CHS1, Glyma.08g110420 in Wm82v4) under transcriptional control of a regulatory element formerly at another location relative to CHS1 in the ancestral genome (likely from the rearranged and fragmented subtilisin-like protease Glyma.08g110380 in Wm82v4). In the rearranged location (Figure S11), the transcription of CHS1 is in reverse orientation (Clough *et al*., 2004; Tuteja *et al*., 2004). The reverse-transcribed CHS1 transcript then pairs with a nearlyidentical but forward-transcribed gene (Glyma.08g110901 in Wm82v4) from downstream in the Wm82 CHS cluster (Figure S11). Post-translational gene silencing (PTGS) then degrades the CHS1 transcripts, yielding seeds with a yellow seed coat (Tuteja *et al*., 2004; Cho *et al*., 2017). Similar organization of the wild-type gene structure at the I locus is reported for the W05 accession of *G. soja* (Xie *et al*., 2019). In tests with the yellow-seeded Wm82, the subtilisin-inverted-CHS1 chimera was able to form double-stranded RNA with the forward-sense CHS mRNAs, thereby likely causing breakdown of the CHS signal via PTGS (Xie *et al.,* 2019).

CONCLUSION

The availability of multiple reference-quality genome assemblies, including an assembly for *G. soja*, will enable basic and applied research in soybean. Multiple assemblies provide confirmation of genomic structure and variations in difficult-to-assemble regions, and comparisons between domesticated accessions and *G. soja* can help to identify the genomic transitions involved in domestication.

The primary reference genome assembly that has been in use for the last decade, Wm82, has been substantially improved, with the closure of more than 3600 gaps, the addition of more than 5 Mbp and with improvements in regions that exhibited high heterozygosity in the previous reference assembly. The use of optical maps and dense genetic maps has resulted in a robust chromosome-scale backbone for soybean, and reciprocal comparisons between the three independent assemblies allowed for the assessment of scaffold contiguity and placement. The presence of centromeric repeats on all chromosomes and terminal telomeric repeats on more than half of the chromosomes in the three assemblies provides an indication of their relative completeness. The incorporation of substantial additional full-length transcript data for the Wm82 gene annotation also strengthens the gene models for soybean.

EXPERIMENTAL PROCEDURES

Similarity comparisons relative to the US soybean germplasm collection

To assess similarities between the three genome assemblies and accessions in the US germplasm collection, a similarity matrix of all 20 087 accessions in the US germplasm against 42 502 SNPs was created using the R program SNPRELATE (Zheng *et al*., 2012). Similarity scores were extracted for all 20 087 US accessions in the SoySNP50K data set for Wm82, Lee and PI 483463 from the similarity matrix using the script extractTop-Match.pl available at Github (https://github.com/avbrown1/SimMatrix-Analysis).

Phylogenetic analysis relative to the US soybean germplasm collection

The SoySNP50k genotype matrix (Song *et al*., 2015) was downloaded from SoyBase Data Store (https://www.soybase.org/data/public/Glycine_max/Wm82.gnm2.div.892R/), in Flapjack format (alleles coded as A, T, C, G or heterozygous sites). This matrix contains genotype data for 20 087 accessions in the US soybean germplasm collection. The main objective of this analysis was to determine the phylogenetic placement of the Lee, Wm82 and *G. soja* PI 483463 accessions relative to other material in the US collection. The size of the US collection presents a challenge, however, as a phylogeny of >20 000 accessions is difficult to visualize. We therefore selected a representative subset of the data by applying several filtering steps. Heterozygous sites were first collapsed to a single allele, selected at random. We set aside genotype data for 11 'focal accessions' to be added back to the analysis at the end. These were: Wm82 and progenitors Williams, Wayne, Clark, Adams, Kingwa and Archer; Lee and progenitors S-100 and CNS; and *G. soja* PI 483463. From the remaining accessions, we selected representative sequences from among near-identical ones, using VSEARCH (Rognes *et al*., 2016), with exemplars being reported from clusters with ≥99% identity. This gave 15 096 accessions. Accessions with large amounts of missing data (>1500 sites out of 42 339) were omitted. To generate an alignment suitable for phylogenetic reconstruction, every 10th SNP was selected, giving an alignment length of 4238 characters. Finally, the 11 focal accessions were added to the sampled set of accessions, giving 1510 representative accessions for phylogenetic analysis. To the genotype identifiers, a tag was added to indicate the country of origin (from the Country of Origin field from GRIN, https://npgsweb.ars-grin.gov). A maximum-likelihood phylogenetic tree was calculated using FASTTREE 2.1.8 (Price *et al*., 2010). Tree visualizations were generated using the ARCHAEOPTERYX tree viewer (Han and Zmasek, 2009).

Glycine max Lee and *G. soja* PI 483463 plant selection for sequencing and assembly

Soybean germplasm seeds were obtained from the USDAGermplasm Resources Information Network (GRIN). A total of 50 seeds were planted in the glasshouse at the University Missouri and after 2 weeks (V1 growth stage) single individuals from each genotype were screened for homozygosity prior to sequencing (Bergelson *et al*., 2016), and a single individual was selected for tissue collection for sequencing. These seeds from selected plants were increased and maintained for further use.

Genome assembly methods

The three assemblies used a combination of technologies: *G. max* Lee used notably NRGene DeNovoMagic scaffolded using Bionano optical maps, *G. soja* PI 483463 used NRGene DeNovoMagic without Bionano scaffolding and *G. max* Wm82v4 was built on top of the existing Wm82v2 with additional information from optical maps, genetic maps and comparison with other assemblies (Appendix S5).

Genome assembly validation

CEGMA 2.5 and BUSCO (Parra *et al*., 2007; Simao *et al*., 2015) analyses as well as read remapping was performed to assess the completeness of the assemblies using the CoReFinder pipeline (Bayer *et al*., 2017), available at http://appliedbioinformatics.com.au/index.php/CoReFinder.

Structural variation between assemblies

Structural comparisons between pseudomolecule assemblies were primarily made using NUCMER from MUMMER 3.23 (Kurtz *et al*., 2004). Visual evaluations were made using dot plots generated by the MUMMERPLOT utility, and alignment summaries generated by the MUMMERSHOW-COORDS utility were analyzed for gaps, inversions and other discontinuities using custom shell scripts.

Nucleotide variation between assemblies

Basepair-level comparisons between pseudomolecule assemblies were made using NUCMER from mummer 3.23 (Kurtz *et al*., 2004), with the SHOW-SNPS utility being used to identify SNPs (with parameters 'show-snps -ClrT'). SNP densities in Figure 5 were generated using CVIT (Cannon and Cannon, 2011). For SNP parentage comparisons (Figure 6), variants from the SoySNP50k array (Song *et al*., 2013) were plotted against the Wm82v2 chromosome coordinates; for Lee, parental lines S-100 (PI 548488) and CNS (PI 548445), and for Wm82, parental lines Williams and Kingwa.

Genome annotation

The *G. max* Wm82v4 and Lee and *G. soja* assemblies were annotated using PERTRAN (Shu *et al*., 2013) and Illumina RNA-seq reads and PASA was used to create transcript assemblies (Haas *et al*., 2003) (Appendix S6). The *G. max* Wm82v4 annotation was additionally improved using Iso-Seq CCSs.

Predicted genes were compared with SWISSPROT and between assemblies using BLAST+ 2.5.0 (Camacho *et al*., 2009) (e-value cutoff: $1e^{-5}$). Gene collinearity was mapped using MCSCANX 2 (Wang *et al*., 2012). ORTHOFINDER 2.2.6 was used for sequence-based clustering (Emms and Kelly, 2015). ORTHOFINDER clusters were functionally annotated using INTERPROSCAN 5.25-64.0 (Jones *et al*., 2014) using SIGNALP 4.1 (Petersen *et al*., 2011) and PFAM 31.0 (Finn *et al*., 2014) and KINFIN 1.0.3 (Laetsch and Blaxter, 2017). For gene loss comparisons, all three annotations were lifted over to the other two assemblies using the flo pipeline (https://github.com/wurmlab/flo), which is based on the University of California, Santa Cruz (UCSC) LiftoverToolkit (Kuhn *et al*., 2013), and genes deleted in the other two assemblies were counted.

Pan-gene comparisons

Pan-gene correspondences between gene models for Wm82v2, Wm82v4, Lee and *G. soja* PI 483463 were calculated both by homology of coding sequences relative to each assembly and by genomic position. Homologies were calculated as the top BLASTN match of each gene sequence from a query assembly to the comparison genome assembly, at ≥95% identity. Chromosomal positions were identified as overlapping gene models, on the corresponding chromosome, from the top GMAP match of each gene between the assemblies. The resulting orthogroups are available in Table S3.

Analysis of resistance genes

Resistance gene candidates (resistance gene analogs or RGAs) were predicted using RGAUGURY (Li *et al*., 2016). Only primary transcripts were used for R-gene prediction, and the class TM-CC was removed from the results.

Analysis of telomeric and centromeric repeats

As a measure of pseudomolecule completeness near the chromosome ends, we checked for characteristic telomeric repeat motifs AAACCCT and AGGGTTT at the leading and trailing ends of a chromosome, respectively, checking for arrays of at least 10 such repeats within 1000 bases of the pseudomolecule ends. We found such telomeric repeat arrays on 26 of the 40 pseudomolecule ends in Wm82, on 22 pseudomolecule ends in Lee and on 18 pseudomolecule ends in *G. soja*.

We searched for the two centromere-specific centromeres CentGm-1 and CentGm-2 (Tek *et al*. 2010; Gill *et al*., 2009) in the three assemblies to identify the assembled centromeric regions. We used tandem repeat finder (trf) to extract all CentGm-1 (92 bp) and CentGm-2 (91 bp)-like long tandem repeats in all three assemblies. The resulting datasets were then merged with CentGm-1 and CentGm-2 and clustered using CD-HIT (identity cut-off: 90%), resulting in consensus sequences TGTGAAAAGTTATGACCATTTGAATTTCTCGAGAGCTTCCGTTGTTCAATTTCGAGCGTCTCGATAT ATTATGCGCCTGAATCGGACATCCG and AGTCAAAAGTTATTGTCGTTTGACTTTTCTCAGAGCTTCCGTTTTCAATTACGAGCGTCTCGA TATATTACGGGACTCAATCGGACATCCG, respectively. CentGm-1 and CentGm-2 were aligned with the reference using BLASTN to identify the location of the centromeres on the pseudomolecules.

ACCESSION NUMBERS

GenBank accession numbers for the genome assemblies are given in Table 1. Assemblies and annotations are also available for download and browsing at both Phytozome (https://phytozome.jgi.doe.gov) and SoyBase (https://soybase.org/data/public/Glycine_max/).

ACKNOWLEDGEMENTS

The authors thank Sarah Kingan (Pacific Biosciences) for providing exploratory sequence and analysis using PacBio sequence data, Nathan Weeks (USDA-ARS, Ames, IA) for providing computational environments for genome analysis, Andrew Farmer (NCGR, Santa Fe, NM) for calculating functional annotations for predicted genes and the HudsonAlpha Genome Sequencing Group for contributions to the *Glycine max* (Wm82) V4 genome assembly and resources. Substantial funding for the work was provided by the United Soybean Board and three industrial partners, Bayer Crop Science (currently BASF), Monsanto (currently Bayer), and Dow AgroSciences (currently Corteva), for funding contributions to the USB soybean genome sequencing project to H.T.N. (1320-532-5615). Funding to T.-F.C. and H.-M.L. in support of this work was supplied by the Hong Kong Research Grants

Council Area of Excellence Scheme (AoE/M-403/16). The work conducted by the US Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported by the Office of Science of the US Department of Energy under contract no. DE-AC02-05CH11231. Additional funding for analysis was provided by National Science Foundation (NSF) award 1444806 to S.B.C. for genome assembly and analysis. Genome finishing work for Wm82 provided by NSF award 0822258 to S.A.J. A portion of the analysis work in this project was provided through in-kind contributions from the USDA Agricultural Research Service, project 5030-21000-069-00-D. The USDA is an equal opportunity provider and employer. Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the US Department of Agriculture. This work was supported by resources provided by the Pawsey Supercomputing Centre with funding from the Australian Government and the Government of Western Australia.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

AUTHOR CONTRIBUTIONS

HTN is the principal investigator for the project. HTN and BV conceived the idea, designed the experiments and managed the project. SBC performed the genome assembly integration. SBC, DE, PEB, AVB, HH, T-FC, JB and YY conducted the genome analysis, and BV and QS also contributed. SS and JS performed the genome annotations. LR performed the repeat analysis. JJ, JG, JS, CP, CGD and SAJ were involved with Wm82 genome improvement, and GS and RS contributed transcription data. JS, KWB and DMG managed the Wm82 genome project. QS constructed the genetic maps. KB developed new assemblies of the variety Lee and PI 483463. AH developed primary genome assembly of the variety Lee using the Illumina reads and Bionano optical maps. H-ML, T-FC and CY-LC conducted the optical map analysis and contributed to the genome analysis. BV and GP contributed to the plant growth, sample preparation and data generation. WH prepared the genome browser. BV, SBC, PEB, DE, SS, JS, AVB, RV, T-FC, H-ML and HTN participated in the first draft of the article, and SBC, BV, PEB, AVB, DE and HTN wrote the final draft.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Phylogenetic tree displayed in Figure 1, in high resolution, with accessions and countries of origin indicated.

Figure S2. Dot plot of Wm82v4 (*x*-axis) by Lee (*y*-axis). Red: forward alignment; blue: inverted alignment.

Figure S3. Dot plot of Wm82v4 (*x*-axis) by *G. soja* (*y*-axis). Red: forward alignment; blue: inverted alignment.

Figure S4. Dot plot of Lee (*x*-axis) by *G. soja* (*y*-axis). Red: forward alignment; blue: inverted alignment.

Figure S5. Densities and locations of repeats (light blue), genes (brown) and centromeric sequences (red) for *G. max* Lee.

Figure S6. Densities and locations of repeats (light blue), genes (brown) and centromeric sequences (red) for *G. soja*.

Figure S7. Densities and locations of repeats (light blue), genes (brown) and centromeric sequences (red) for *G. max* Wm82v4.

Figure S8. Plot of Lee pseudomolecules (*x*-axis) by genetic map (*y*-axis; Wm82 × *G. soja* map). Vertical dotted lines show scaffold boundaries within pseudomolecules.

Figure S9. Plot of *G. soja* pseudomolecules (*x*-axis) by genetic map (*y*-axis; Wm82 × *G. soja* map). Vertical dotted lines show scaffold boundaries within pseudomolecules.

Figure S10. Plot of Wm82 pseudomolecules (*x*-axis) by genetic map (*y*-axis; Wm82 × *G. soja* map). Vertical dotted lines show scaffold boundaries within pseudomolecules.

Figure S11. Plots of the I locus on chromosome 8, controlling seed coat color. (a) Schematic of I locus: wild-type (pigmented) structure at top; cultivated (unpigmented) structure at bottom. Red: subtilisin genes, and partial gene with promoter. Yellow: chalcone synthase (CHS) genes. The subtilisin promoter drives a duplicated and inverted CHS, which causes degradation of the corresponding CHS transcript through post-translational gene silencing (PTGS). (b) Dot plot of Wm82v4 and *G. soja* (*y*- and *x*-axes, respectively). (c) Gene structures from the I locus, from *G. soja*. (d) Gene structures from the I locus, from *G. max* Wm82v4. (e) Dot plot of *G. soja* and Wm82v4 (*y*- and *x*-axes, respectively). Red highlighted regions: subtilisin gene.Yellow highlighted regions: chalcone synthase genes. Dotted lines show approximate inversion and duplication boundaries.

Table S1. Similarity scores between each accession in the US soybean collection and Lee, Wm82 and *G. soja* PI 483463. Similarities and countries of origin are indicated in column headings and plotted by descending similarity scores, as sorted for each of the three comparison genotypes.

Table S2. Repetitiveness and inferred assembly collapses (underlying Figure 2).

Table S3. Pan-genome correspondences of gene models across four assemblies.

Table S4. Gene ontology enrichment from the genome annotation comparisons.

Appendix S1. Sequence alignment used to calculate the genotype phylogeny in Figure 1. Genotypes and SNP positions are sampled as described in Experimental procedures.

Appendix S2. Phylogenetic tree data displayed in Figure 1. The format for the maximum-likelihood tree is ‘phylip’.

Appendix S3. Sequence data, countries of origin and linear tree order for Figure S1 and Figure 1.

Appendix S4. R-language code for analysis of repetitive and collapsed regions (used in Figure 2).

Appendix S5. Genome assembly.

Appendix S6. Genome annotation.

REFERENCES

Arumuganathan, K., Slattery, J.P., Tanksley, S.D. and Earle, E.D. (1991) Preparation and flow cytometric analysis of metaphase chromosomes of tomato. *Theor. Appl. Genet.* **82**, 101–111.

Avni, R., Nave, M., Barad, O. *et al.* (2017) Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. *Science*, **357**, 93–97.

Bandillo, N., Jarquin, D., Song, Q., Nelson, R., Cregan, P., Specht, J. and Lorenz, A. (2015) A population structure and genome-wide association analysis on the USDA soybean germplasm collection. *Plant Genome*, **8**, 1–13.

Bayer, P.E., Hurgobin, B., Golicz, A.A. *et al.* (2017) Assembly and comparison of two closely related Brassica napus genomes. *Plant Biotechnol. J.* **15**, 1602–1610.

Bergelson, J., Buckler, E.S., Ecker, J.R., Nordborg, M. and Weigel, D. (2016) A proposal regarding best practices for validating the identity of genetic stocks and the effects of genetic variants. *Plant Cell*, **28**, 606–609.

Bernard, R.L. and Cremeens, C.R. (1988) Registration of 'Williams 82' soybean. *Crop Sci.* **28**, 1027–1028.

Burton, J.N., Adey, A., Patwardhan, R.P., Qiu, R., Kitzman, J.O. and Shendure, J. (2013) Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31**, 1119.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

Cannon, E.K. and Cannon, S.B. (2011) Chromosome visualization tool: a whole genome viewer. *Int. J. Plant Genomics*, **2011**, 373875.

Cho, Y.B., Jones, S.I. and Vodkin, L.O. (2017) Mutations in Argonaute5 illuminate epistatic interactions of the K1 and I Loci leading to saddle seed color patterns in Glycine max. *Plant Cell*, **29**, 708–725.

Clough, S.J., Tuteja, J.H., Li, M., Marek, L.F., Shoemaker, R.C. and Vodkin, L.O. (2004) Features of a 103-kb gene-rich region in soybean include an inverted perfect repeat cluster of CHS genes comprising the I locus. *Genome*, **47**, 819–831.

Curtin, S.J., Xiong, Y., Michno, J.M., Campbell, B.W., Stec, A.O., Cermak, T., Starker, C., Voytas, D.F., Eamens, A.L. and Stupar, R.M. (2018) CRISPR/Cas9 and TALENs generate heritable mutations for genes involved in small RNA processing of Glycine max and Medicago truncatula. *Plant Biotechnol. J.* **16**, 1125–1137.

Desta, Z.A. and Ortiz, R. (2014) Genomic selection: genome-wide prediction in plant improvement. *Trends Plant Sci.* **19**, 592–601.

Dong, Y., Yang, X., Liu, J., Wang, B.H., Liu, B.L. and Wang, Y.Z. (2014) Pod shattering resistance associated with domestication is mediated by a NAC gene in soybean. *Nat. Commun.* **5**, 3352.

Dorrance, A.E., Jia, H. and Abney, T.S. (2004) Evaluation of soybean differentials for their interaction with Phytophthora sojae. *Plant Health Prog.* **5**, 9.

Du, J., Grant, D., Tian, Z., Nelson, R.T., Zhu, L., Shoemaker, R.C. and Ma, J. (2010) SoyTEdb: a comprehensive database of transposable elements in the soybean genome. *BMC Genom.* **11**, 113.

Emms, D.M. and Kelly, S. (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157.

Findley, S.D., Cannon, S., Varala, K., Du, J., Ma, J., Hudson, M.E., Birchler, J.A. and Stacey, G. (2010) A fluorescence in situ hybridization system for karyotyping soybean. *Genetics*, **185**, 727–744.

Finn, R.D., Bateman, A., Clements, J. *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res.* **42**, D222–D230.

Funatsuki, H., Suzuki, M., Hirose, A. *et al.* (2014) Molecular basis of a shattering resistance boosting global dissemination of soybean. *Proc. Natl Acad. Sci. USA*, **111**, 17797–17802.

Gao, H. and Bhattacharyya, M.K. (2008) The soybean-Phytophthora resistance locus Rps1-k encompasses coiled coil-nucleotide binding-leucine rich repeat-like genes and repetitive sequences. *BMC Plant Biol.* **8**, 29.

Gao, M. and Zhu, H. (2013) Fine mapping of a major quantitative trait locus that regulates pod shattering in soybean. *Mol. Breeding*, **32**, 485–491.

Gill, N., Findley, S., Walling, J.G., Hans, C., Ma, J., Doyle, J., Stacey, G. and Jackson, S.A. (2009) Molecular and chromosomal evidence for allopolyploidy in soybean. *Plant Physiol.* **151**, 1167–1174.

Golicz, A.A., Batley, J. and Edwards, D. (2016) Towards plant pangenomics. *Plant Biotechnol. J.* **14**, 1099–1105.

Haas, B.J., Delcher, A.L., Mount, S.M. *et al.* (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666.

Han, M.V. and Zmasek, C.M. (2009) PhyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinformatics*, **10**, 356.

Haun, W.J., Hyten, D.L., Xu, W.W. *et al.* (2011) The composition and origins of genomic variation among individuals of the soybean reference cultivar Williams 82. *Plant Physiol.* **155**, 645–655.

Hwang, E.-Y., Song, Q., Jia, G., Specht, J.E., Hyten, D.L., Costa, J. and Cregan, P.B. (2014) A genome-wide association study of seed protein and oil content in soybean. *BMC Genom.* **15**, 1.

Hyten, D.L., Song, Q., Zhu, Y., Choi, I.Y., Nelson, R.L., Costa, J.M., Specht, J.E., Shoemaker, R.C. and Cregan, P.B. (2006) Impacts of genetic bottlenecks on soybean genome diversity. *Proc. Natl Acad. Sci. USA*, **103**, 16666–16671.

Jiao, Y., Peluso, P., Shi, J. *et al.* (2017) Improved maize reference genome with single-molecule technologies. *Nature*, **546**, 524–527.

Jones, P., Binns, D., Chang, H.Y. *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**, 1236–1240.

Kent, W.J. (2002) BLAT–the BLAST-like alignment tool. *Genome Res.* **12**, 656–664.

Kuhn, R.M., Haussler, D. and Kent, W.J. (2013) The UCSC genome browser and associated tools. *Brief. Bioinform.* **14**, 144–161.

Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C. and Salzberg, S.L. (2004) Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12.

Laetsch, D.R. and Blaxter, M.L. (2017) KinFin: software for Taxon-Aware analysis of clustered protein sequences. *G3 (Bethesda)*, **7**, 3349–3357.

Lee, J.D., Shannon, J.G., Vuong, T.D. and Nguyen, T.N. (2009) Inheritance of salt tolerance in wild soybean (*Glycine soja* Sieb. and Zucc.) accession PI 483463. *J. Hered.* **100**, 798–801.

Li, Y.H., Zhou, G., Ma, J. *et al.* (2014) De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat. Biotechnol.* **32**, 1045–1052.

Li, P., Quan, X., Jia, G., Xiao, J., Cloutier, S. and You, F.M. (2016) RGAugury: a pipeline for genome-wide prediction of resistance gene analogs (RGAs) in plants. *BMC Genom.* **17**, 852.

Ling, H.Q., Ma, B., Shi, X. *et al.* (2018) Genome sequence of the progenitor of wheat A subgenome Triticum urartu. *Nature*, **557**, 424–428.

Liu, Q., Chang, S., Hartman, G.L. and Domier, L.L. (2018) Assembly and annotation of a draft genome sequence for Glycine latifolia, a perennial wild relative of soybean. *Plant J.* **95**, 71–85.

Parra, G., Bradnam, K. and Korf, I. (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, **23**, 1061–1067.

Petersen, T.N., Brunak, S., vonHeijne, G. and Nielsen, H. (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods*, **8**, 785–786.

Price, M.N., Dehal, P.S. and Arkin, A.P. (2010) FastTree 2 – Approximately maximum-likelihood trees for large alignments. *PLoS ONE*, **5**(3), e9490.

Prince, S.J., Valliyodan, B., Ye, H. *et al.* (2019) Understanding genetic control of root system architecture in soybean: insights into the genetic basis of lateral root number. *Plant Cell Environ.* **42**, 212–229.

Raymond, O., Gouzy, J., Just, J. *et al.* (2018) The Rosa genome provides new insights into the domestication of modern roses. *Nat. Genet.* **50**, 772–777.

Rognes, T., Flouri, T., Nichols, B., Quince, C. and Mahé, F. (2016) VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, **4**, e2584.

Salamov, A.A. and Solovyev, V.V. (2000) Ab initio gene finding in Drosophila genomic DNA. *Genome Res.* **10**, 516–522.

Schmutz, J., Cannon, S.B., Schlueter, J. *et al.* (2010) Genome sequence of the palaeopolyploid soybean. *Nature*, **463**, 178–183.

Sedivy, E.J., Wu, F. and Hanzawa, Y. (2017) Soybean domestication: the origin, genetic architecture and molecular bases. *New Phytol.* **214**, 539–553.

Shen, Y., Liu, J., Geng, H., Zhang, J., Liu, Y., Zhang, H., Xing, S., Du, J., Ma, S. and Tian, Z. (2018) De novo assembly of a Chinese soybean genome. *Sci. China Life Sci.* **61**, 871–884.

Shimomura, M., Kanamori, H., Komatsu, S. *et al.* (2015) The Glycine max cv. Enrei genome for improvement of japanese soybean cultivars. *Int. J. Genomics*, **2015**, 358127.

Shu, S., Goodstein, D. and Rokhsar, D. (2013) PERTRAN: Genome-guided RNA-seq Read Assembler. OSTI.gov: U.S. Department of Energy - Office of Scientific and Technical Information.

Simao, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. and Zdobnov, E.M. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**, 3210–3212.

Song, Q., Hyten, D.L., Jia, G., Quigley, C.V., Fickus, E.W., Nelson, R.L. and Cregan, P.B. (2013) Development and evaluation of SoySNP50K, a high-density genotyping array for soybean. *PLoS ONE*, **8**, e54985.

Song, Q., Hyten, D.L., Jia, G., Quigley, C.V., Fickus, E.W., Nelson, R.L. and Cregan, P.B. (2015) Fingerprinting soybean germplasm and its utility in genomic research. *G3 (Bethesda)*, **5**, 1999–2006.

Song, Q., Jenkins, J., Jia, G., Hyten, D.L., Pantalone, V., Jackson, S.A., Schmutz, J. and Cregan, P.B. (2016) Construction of high resolution genetic linkage maps to improve the soybean genome sequence assembly Glyma1.01. *BMC Genom.* **17**, 33.

Springer, N.M., Anderson, S.N., Andorf, C.M. *et al.* (2018) The maize W22 genome provides a foundation for functional genomics and transposon biology. *Nat. Genet.* **50**, 1282–1288.

Sun, L., Miao, Z., Cai, C. *et al.* (2015) GmHs1-1, encoding a calcineurin-like protein, controls hard-seededness in soybean. *Nat. Genet.* **47**, 939.

Tek, A.L., Kashihara, K., Murata, M. and Nagaki, K. (2010) Functional centromeres in soybean include two distinct tandem repeats and a retrotransposon. *Chromosome Res.* **18**, 337–347.

Tian, Z., Wang, X., Lee, R., Li, Y., Specht, J.E., Nelson, R.L., McClean, P.E., Qiu, L. and Ma, J. (2010) Artificial selection for determinate growth habit in soybean. *Proc. Natl Acad. Sci. USA*, **107**, 8563–8568.

Tian, Z., Zhao, M., She, M. *et al.* (2012) Genome-wide characterization of nonreference transposons reveals evolutionary propensities of transposons in soybean. *Plant Cell*, **24**, 4422–4436.

Tuteja, J.H., Clough, S.J., Chan, W.C. and Vodkin, L.O. (2004) Tissue-specific gene silencing mediated by a naturally occurring chalcone synthase gene cluster in Glycine max. *Plant Cell*, **16**, 819–835.

Tuteja, J.H., Zabala, G., Varala, K., Hudson, M. and Vodkin, L.O. (2009) Endogenous, tissue-specific short interfering RNAs silence the chalcone synthase gene family in glycine max seed coats. *Plant Cell*, **21**, 3063–3077.

Valliyodan, B., Dan, Q., Patil, G. *et al.* (2016) Landscape of genomic diversity and trait discovery in soybean. *Sci. Rep.* **6**, 23598.

Valliyodan, B., Ye, H., Song, L., Murphy, M., Shannon, J.G. and Nguyen, H.T. (2017) Genetic diversity and genomic strategies for improving drought and waterlogging tolerance in soybeans. *J. Exp. Bot.* **68**, 1835–1849.

Vaughn, J.N., Nelson, R.L., Song, Q., Cregan, P.B. and Li, Z. (2014) The genetic architecture of seed composition in soybean is refined by genome-wide association scans across multiple populations. *G3 (Bethesda)*, **4**, 2283–2294.

Wang, Y., Tang, H., Debarry, J.D. *et al.* (2012) MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**(7), e49.

Wang, C.S., Todd, J.J. and Vodkin, L.O. (1994) Chalcone synthase mRNA and activity are reduced in yellow soybean seed coats with dominant I alleles. *Plant Physiol.* **105**, 739–748.

Wang, J., Sun, P., Li, Y. *et al.* (2017) Hierarchically aligning 10 legume genomes establishes a family-level genomics platform. *Plant Physiol.* **174**, 284–300.

Wu, T.D. and Watanabe, C.K. (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, **21**, 1859–1875.

Xie, M., Chung, C.Y., Li, M.W. *et al.* (2019) A reference-grade wild soybean genome. *Nat. Commun.* **10**, 1216.

Zheng, X., Levine, D., Shen, J., Gogarten, S., Laurie, C. and Weir, B. (2012) A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, **28**(24), 3326–3328.

Zhou, Z., Jiang, Y., Wang, Z. *et al.* (2015) Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat. Biotechnol.* **33**, 408–414.