Sequencing, Assembly, and Annotation of the Soybean Genome

5

Babu Valliyodan, Suk-Ha Lee and Henry T. Nguyen

Abstract

Genome sequencing yields an exceptional resource of genetic information. Knowledge of whole genome sequence information helps characterize individual genomes, transcriptional states and genetic variation in populations and provide genetic architecture associated with each trait. After the release of the first human genome assembly, other model organism assemblies became available; including the model plant *Arabidopsis thaliana*. The soybean community published the first reference genome of the variety Williams 82 in 2010. Soybean has important syntenic relationships with the other legume species and is a model plant for the legumes. In this chapter, we discuss about the soybean genome assemblies and annotations, and its fine-tuning in view of the next-generation sequencing technologies and bioinformatics tools. In addition, comparison of the structural variations between the cultivated reference genome with the available wild soybean genome information will be discussed. This is followed by the discussion on the opportunities of next-generation sequencing technologies and challenges that we anticipate on the development of more pangenomes and reference genomes for soybean. This will significantly affect the discovery of rare alleles associated with key agronomic and quality traits and shape up the next-generation breeding technologies and crop improvement.

B. Valliyodan (✉) · H.T. Nguyen
Division of Plant Sciences, University of Missouri, Columbia, MO 65211, USA
e-mail: valliyodanb@missouri.edu

S.-H. Lee
Department of Plant Science, Research Institute for Agriculture and Life Sciences and Plant Genomics and Breeding Institute, Seoul National University, Seoul 151-921, Korea

5.1 Introduction

Genetic code is the basis of organismal life. Deciphering the primary DNA sequence, i.e. the genome, and the associated gene, has become a fundamental resource in biology. The speed of genome sequencing is higher in the case of mammalian and microbial genomes when compared to

H.T. Nguyen and M.K. Bhattacharyya (eds.), *The Soybean Genome*, Compendium of Plant Genomes, DOI 10.1007/978-3-319-64198-0_5

plants. Advances in the next-generation sequencing technologies made the whole genome sequencing technically feasible and cost-effective. Application of genomics and the genome data show significant impact on biological investigations irrespective of areas and disciplines. Whole genome sequencing combined with de novo assemblies generates newer reference genomes for genetic investigation and trait discovery in each organism. State-of-the-art sequencing techniques and the associated computational processes provide better resolution to transcriptome, epigenome, and organellar genomes (Varshney et al. 2009). In addition, targeted sequencing including the exomes and regulomes bring major shift to the high-throughput trait diagnostics (Hashmi et al. 2015). A total of 24,002-genome information is available in the NCBI, which includes eukaryotes, prokaryotes, viruses, plasmids, and organelles (https://www.ncbi.nlm.nih.gov/genome/browse/).

The release of the Arabidopsis genome in 2000 (Arabidopsis Genome Consortium 2000) and the technological advances in sequencing methods accelerated the number of sequenced plant genomes closely to 185. Out of these, more than 50% of them are crop genomes and the rest belongs to model plant genomes, orphan crops, and wild relative species. Availability of this genome information shifted the genetic investigation to the next level of research, precision genomics. Also, collection of these genomes representing various species, genera, and classes in the plant kingdom help reveal the evolutionary relationships between plants and also provide better clarity of synteny and gene family evolutions. Above all, the crop-specific genome sequence and transcript information enhances the power of gene and functional marker discovery that will greatly impact the next-generation breeding programs including the genomics-assisted breeding and genomic section.

5.2 Genome Sequencing Platform

The first human genome (Venter et al. 2001; HGSC 2004) and other genome sequencing projects including model plants as well as major crop species such as rice, soybean, sorghum, maize, and grape mainly depended upon Sanger sequencing methods (Sanger et al. 1977), which are expensive and time consuming for complex genomes. Introduction of the next-generation sequencing technologies improved the output/cost ratio of genome sequencing dramatically. In other words, the pyrosequencing approaches were replaced to certain extent by the "sequencing by synthesis" approaches (Margulies et al. 2005; Shendure et al. 2005). At the earlier phase, this technology provided multiplexed DNA fragment sequencing, generated short reads with low-quality data, and after improving the basic chemistry, now it is possible to acquire high-quality, short-read DNA fragment sequences (Fuller et al. 2009). However, the complex genome assemblies using these comparatively short reads show major drawbacks, including assembly and determination of complex genomic regions, identification of gene isoform and other structural rearrangements.

Single-molecule, real-time sequencing technology developed by Pacific BioSciences offers longer read lengths and higher consensus accuracies than the short-read technologies, making it well suited for better characterization in genome, transcriptome, and epigenetics research (Eid et al. 2009; Nakano et al. 2017). The major requirement for the de novo reference assembly of genomes is less percentages of gaps and the contigs generated using the longer reads significantly improve the assemblies by closing gaps. In addition, it helps to better characterize the structural variations in the genomes (Utturkar et al. 2014; Rhoads and Au 2015). More recently, development of optical map helped to detect mismatches in assemblies, providing genomic maps with high resolution and to allow assemblies with more accuracy and completeness (Schwartz et al. 1993; Tang et al. 2015; Jiao et al. 2017a). In general, hybrid-sequencing strategies including short reads, long reads, and optical maps are more affordable and scalable for genomic investigations. These high-quality sequences improve the quality of contigs and scaffolds yielding several fold better genome assembly.

5.3 Plant Genomes and Assemblies

The development of reference genome assemblies is essential to identify the DNA sequence variations. Plant genome assembly is a challenging task and even more challenging than the animal genomes. Larger genome size, polyploidy, highly repetitive genomic regions, and genome duplication are the major challenges with most of the plant genomes. For example, the repetitive fraction of the human genome varies between 35 and 45%, whereas in maize, it is 64–73% (Imelfort and Edwards 2009). Even though there are more than 180 plant reference genomes, assembly of only less number of them are at the chromosome level (Schnable et al. 2009; Jiao et al. 2017b). Recent advances in the next-generation sequencing technologies, especially the long-read sequencing, long-range scaffolding, and chromosome capturing (Burton et al. 2013) helped to overcome the challenges in the chromosome-level assemblies of genomes.

Genomes sequenced by Sanger sequencing technique were initially assembled using the TIGR Assembler (Sutton et al. 1995) and Celera Assembler (Myers 1995). Advances in the genome sequencing technology required newer assembly tools as well. In the case of genome assembly using the short-read, de Bruijn graph assemblers (Pevzner et al. 2001) such as Velvet (Zerbino and Birney 2008) and ABySS (Simpson et al. 2009) were used. There are de novo sequence assembly tools available to generate good quality and robust assembly of complex genomes using short reads. These assemblers include SSAKE (Warren et al. 2007), VCAKE (Jeck et al. 2007), Edena (Hernandez et al. 2008), EulerSR (Chaisson and Pevzner 2008) and AllPaths (Butler et al. 2008). Recently, a new assembler called Supernova (Weisenfeld et al. 2017) and new version of ABySS assembler explores the linked reads and optical mapping for improved scaffolding.

A combination of single-molecule sequencing with complementary technologies has also become a common strategy. Sequencing technologies such as PacBio, Nanopore, and optical mapping produce longer reads exceeding 10 kb, and this larger read length and increased error rate of these new technologies required updated assembly methods. New assembly tools such as Canu (Koren et al. 2017), HINGE (Kamath et al. 2017), and Racon (Vaser et al. 2017) designed specifically for long-read PacBio and Nanopore data assembly. In the case of plant genomes, Zinin et al. (2017) assembled the highly repetitive grass *Aegilops tauschii*, by combining PacBio long-read and Illumina short-read sequences. Also, combination of chromosome conformation capture and optical mapping generated a new version of the model plant *Arabidopsis thaliana* genome (Jiao et al. 2017a). This approach helped to improve assembly contiguity reaching chromosome-arm levels.

5.4 Soybean Genome Assembly and Annotation

5.4.1 Soybean Genome Before Whole Genome Sequencing

Grain legumes play significant role in the global food and nutritional security, and soybean is one of the leading oil seed crops in the world. The United States Department of Agriculture (USDA) estimates that the global soybean production in 2016/2017 will be 348.04 million metric tons, around 2.07 million tons more than the previous month's projection (USDA-FAS, May 2017). The development of several genomic tools including better genetic map and genome sequence information have immense contribution towards crop improvement. Soybean the cultivated species *Glycine max* and the close relatives including wild species *Glycine soja* and wild perennial *Glycine tomentella* are members of the tribe *Phaseoleae*, the most economically important of the legume tribes. The genus *Glycine* is paleopolyploid, with $2n = 40$ as its base chromosome number, as compared with other phaseoloid legumes, which are largely $2n = 20$ or 22 (Goldblatt 1981). The estimated average size of soybean genome was estimated at about 1.1×10^9 bp/C based on flow cytometry (Arumuganathan and Earle 1991), and similar values

were predicted by DNA re-association kinetics (Cot analysis; Goldberg 1978; Gurley et al. 1979). Both these soybean Cot studies suggested that 40–60% of the genome is repetitive; re-association of different size fragments indicates that the majority of the repetitive sequences are physically linked while a smaller fraction is interspersed with single copy DNA. This latter conclusion is supported by 2700 sequence sampling of nearly 1000 loci across all 20 soybean linkage groups (Marek et al. 2001). The predictions from this research are that the gene rich and repetitive regions occasionally are interspersed, but gene rich "islands" are present as well. Cytological studies show that euchromatin represents ~65% of the genome with heterochromatin mainly localized to the pericentromeric regions and the short arms of four chromosomes (Singh and Hymowitz 1988). Collectively these studies suggest that gene space accounts for about one third of the soybean genome (i.e. $\sim 3.7 \times 10^8$ bp). Two large-scale genome-wide duplication events at 40–50 and 8–10 Mya occurred in the case of soybean and the duplicated regions were segmented and reshuffled after these events (Shultz et al. 2007).

The soybean community has developed substantial amount of marker information and generated physical and genetic maps of soybean, which contributed heavily to the whole genome sequencing and chromosome assembly. As a complimentary approach to the whole genome sequencing, a large number of expressed sequence tags were generated for soybean. Initially around 200,000 expressed sequence tags (ESTs) were generated from more than 50 cDNA libraries representing a wide range of organs, developmental stages, genotypes, and environmental conditions (Shoemaker et al. 2002). In the early 1990s, genome mapping of soybean based on the DNA markers was initiated and several genetic linkage maps of soybean have been published in the last decade. The early maps were developed based primarily on the restriction fragment length polymorphism (RFLP), amplified fragment length polymorphism (AFLP), and simple sequence repeat (SSR) markers, and the more recent maps include single nucleotide polymorphism (SNP) markers. In total, several thousand genetic markers (mostly SSR and SNP markers) were mapped in soybean before release of the first draft genome sequence (Keim et al. 1990; Cregan et al. 1999; Wu et al. 2004, 2010; Song et al. 2004; Kassem et al. 2006; Choi et al. 2007; Xia et al. 2007; Hisano et al. 2008; Yang et al. 2008). Initial physical map for the cultivar Williams 82 was developed from sequencing of the bacterial artificial chromosome (BAC) libraries (Luo et al. 2003; Warren 2006; Shoemaker et al. 2008). Soybean physical maps for "Forrest" and "Williams 82" representing the southern and northern US soybean germplasm base were constructed with different fingerprinting methods. These physical maps are complementary for coverage of gaps on the 20 soybean linkage groups. More than 5000 genetic markers have been anchored onto the Williams 82 physical map, but only a limited number of markers have been anchored to the Forrest physical map. A framework map with almost 1000 genetic markers was constructed using a core set of recombinant inbred lines (RILs) developed from a mapping population of Forrest × Williams 82 (Wu et al. 2011). High-resolution physical maps for both *G. max* and *G. soja* were developed later (Ha et al. 2012), and these maps served as a framework for ordering sequence fragments, comparative genomics, cloning genes, and evolutionary analyses of legume genomes.

5.4.2 Soybean Whole Genome Sequencing

5.4.2.1 Cultivated Soybean Genome Version 1

The first reference genome of the cultivated soybean was released in the year 2010 by the soybean research community in collaboration with the Department of Energy Joint Genome Institute (DOE-JGI) (Schmutz et al. 2010). The northern US cultivar Williams 82 represented the first soybean reference genome. Whole genome shot gun approach was adapted to sequence 1.1 gigabase (Gb) genome. Three sized insert

Table 5.1 Final assembly statistics of *G. max* cv. Williams 82 reference genome version 1 (Glyma1.01)

Total scaffold	1168
Total contig	16,311
Total scaffold sequence	973.3 Mb
Total contig sequence	955.1 Mb
Scaffold N/L50	10/47.8 Mb
Contig N/L50	1492/189.4 Kb

libraries and several BAC libraries were sequenced using Sanger sequencing protocols on ABI 3730XL capillary sequencing machines at the JGI. A total of 15,332,163 sequence reads were assembled into 3363 scaffolds using Arachne assembler (Batzoglou et al. 2002; Jaffe et al. 2003). This assembly covers 969.6 Mb of the soybean genome. To obtain 20 chromosome-level pseudomolecules, integrated the genome assembly with the physical map and high-density genetic map of soybean and the resulting assembly covered 937.3 Mb of the genome size. In addition, 1148 unmapped scaffolds that covered 17.7 Mb of the genome, added to make the total sequence coverage of 8.04×. After the genome assembly and chromosome assignments, this reference genome represented 85% (955 Mb) of the estimated genome size with 1.9% gap (Table 5.1).

Large fractions of plant genomes contain repeated DNA sequences of various types. A combination of genome structure analysis and homology studies identified that 59% of the soybean genome is repeat rich where majority of them are transposable elements (TEs). Analysis of the repetitive elements in the soybean reference genome revealed 57% of the repeat rich low recombination heterochromatic regions are around the centromeres. Percentage of repeat elements are more as compared to Arabidopsis (de la Chaux et al. 2012) and rice (Takata et al. 2007), and lesser compared to maize genome (Schnable et al. 2009). Long terminal repeat (LTR) retrotransposons are the majority of the repetitive elements in the soybean genome (42%) and consist of 510 families containing 14,106 intact elements, including 9733 gypsy-like and 4373 Copia-like transposable elements. Major role of transposable elements is to generate genomic novelty in organisms. This genomic novelty happens through a combination of chromosome rearrangements and associated gene regulation processes. TEs influence genome size, gene content, gene order, and several aspects of nuclear biology (Bennetzen and Wang 2014). Large-scale epigenetic changes due to polyploidization or stress responses affects TEs and all these influence the genomes to generate novel functions (Galindo-González et al. 2017).

The Williams 82 genome contains 46,430 high-confidence protein-coding loci and another ~20,000 predicted loci with low confidence. While comparing these loci with the angiosperm protein families, 283 legume specific gene families harbouring 448 soybean specific genes were identified. A 12.2% of the high-confidence protein-coding loci (5671) represents putative transcription factors in the genome (Schmutz et al. 2010). All these transcription factors from the 28 families were further annotated and a comprehensive soybean transcription factor database, SoyDB, was generated (Wang et al. 2010). The annotations in SoyDB include predicted tertiary structures, protein domains, multiple sequence alignments, DNA binding sites, and consensus sequences for each transcription factor family. Data providing experimental support to the gene content was available after the release of the soybean genome information (Libault et al. 2010; Severin et al. 2010). Transcriptome analysis of soybean tissues collected from 14 different biological conditions revealed the transcription of 55,616 annotated genes and demonstrated that 13,529 annotated soybean genes are putatively pseudogenes (Libault et al. 2010). Mining of the above gene expression atlas identified several tissue specific transcription factors and helped to understand the molecular

Table 5.2 Final assembly statistics of *G. max* cv. Williams 82 reference genome version 2 (Wm82.a2.v1)

Total scaffold	1190
Total contig	17,191
Total scaffold sequence	978.5 Mb
Total contig sequence	955.4 Mb
Scaffold N/L50	10/48.6 Mb
Contig N/L50	1548/189.4 Kb

basis of transcriptional regulation associated with various biological processes.

5.4.2.2 Cultivated Soybean Genome Version 2

Recently, a new version of the soybean genome assembly has been released (Wm82.a2.v1) after correcting several issues in the first version (Glyma1.01). The Glyma1.01 assembly of the whole genome sequence contains 236 unanchored scaffolds with lengths ranging from 10 to 100 kb and 51 unanchored scaffolds with lengths greater than 100 kb (Song et al. 2016). Even though the first assembly was generated using the integrated linkage maps and a genetic map with additional markers, the marker density was not enough to fully cover all regions of the soybean genome. Two high-density genetic linkage maps of soybean based on 21,478 SNP loci mapped in the *G. max* Williams 82 × *G. soja* PI479752 population with 1083 RILs and 11,922 SNP loci mapped in the *G. max* Essex × *G. max* Williams 82 population with 922 RILs were constructed. The high-density genetic linkage maps helped to identify false joins or misplaced scaffolds and unanchored scaffolds in the Glyma1.01 assembly, and the corresponding scaffolds were broken or reassembled to a new Wm82.a2.v1 assembly (Song et al. 2016).

The Wm82.a2.v1 was generated using the latest version of the ARCHNE assembler. A combination of high-density genetic linkage maps and *Phaseolus* synteny was used to identify the false joins in the Glyma1.01 assembly and the scaffold was broken based on these false joins. A total of 63 breaks were identified and broken, and the order and orientation of the broken scaffolds was achieved using the high-density markers and *Phaseolus* synteny. The total sequence including the 1170 unmapped scaffolds was 978.5 Mb and the new build of the 20 chromosomes captured 949.2 Mb (Table 5.2). In the new version of soybean genome assembly, 56,044 protein-coding loci and 88,647 transcripts were predicted. The Wm82.a2.v1 gene set integrates ~1.6 million ESTs, and 1.5 billion paired-end Illumina RNA sequence reads with homology-based gene predictions (https://phytozome.jgi.doe.gov).

5.4.2.3 Wild Soybean Whole Genome Sequencing

Domestication history of cultivated soybean traces back to around 5000 years ago in China and it is considered that wild soybean *Glycine soja* is the closest relative of the cultivated soybean, *Glycine max* (Hymowitz 1970). Both the species have 20 chromosomes ($2n = 40$) and belong to the primary gene pool, where the hybrids are vigorous, exhibit normal meiotic chromosome pairing and normal gene segregation. However, the wild and cultivated soybeans differ in several morphological characteristics (Hymowitz 2004). In the case of *G. soja* publically available DNA sequence resources are limited. It is considered that the wild species is the untapped genetic resource for crop improvement (Valliyodan et al. 2016). The whole genome of wild soybean, *G. soja* var. IT 812932 was sequenced using two platforms, Illumina-GA, and GS-FLX, after the release of the cultivated soybean genome (Kim et al. 2010). Around 48.8 Gb short DNA reads were aligned to the *G. max* reference genome and a consensus sequence was determined for *G. soja*. This consensus sequence spanned 915.4 Mb, with a coverage of 97.65% of the available *G. max* reference genome sequence and an average mapping depth of 43-fold. Also, 32.4 Mb of large deletions and 8.3 Mb of novel sequence contigs in the *G. soja* genome were

detected. The *G. soja* unmapped and unpaired reads were assembled de novo by Velvet (Version 0.7.31) assembler (Zerbino and Birney 2008). The genome structural analysis revealed 5794 deletions and 194 inversions and predicted 8554 insertions in the *G. soja* genome. More than 48% of the deleted regions in the wild soybean genome contain repetitive elements including the major classes of retrotransposons (20.74%). In another study, the 5794 deletions were compared against *G. max* gene positions for regions of overlap. This comparative genomic analysis identified 425 unique genes from the list of higher confidence 46,430 gene model predictions of the Glyma1.01, that are absent in *G. soja* and unique in *G. max* (Joshi et al. 2013). In addition, this study showed that there are significant genomic level differences between *G. max* and *G. soja* that are associated with some functionally important genes for seed, oil, and protein traits. Further investigation of these genes can explain some of the phenotypic differences observed between *G. soja* and *G. max* especially in terms of the major seed composition and agronomic traits.

Recently, de novo assembly of a salt tolerant wild soybean (*G. soja*) accession, W05 genome was reported (Qi et al. 2014). The genome size of this wild soybean was estimated at 1.17 Gb using k-mer statistics (Lander and Waterman 1988), and it is close to the estimated size (1.12 Gb) of the cultivated soybean reference genome, Williams 82. SOAPdenovo software (Li et al. 2010) assembled the 868 Mb of the W05 genome (Table 5.3). About 43.41% of the genome contains repeat elements and majority of them are LTRs (30.89%). The W05 genome contains 52,395 protein-coding genes, and out of these, 49,560 are functionally annotated and the average coding sequence length is 1083.9 bp. Development of wild soybean de novo genomes can facilitate the identification of genetic materials from the untapped genomic resource for crop improvement since these genomes can identify small and large genomic variations.

In order to capture the entire genomic sequence present within the species, including the complete gene set, the pangenome needs to be sequenced (Golicz et al. 2016). Tettelin et al. (2005) introduced the concept of pangenome defining it as a full genomic (genic) makeup of a species. A single genome is insufficient to represent the genomic content in the case of both cultivated and wild soybean where individuals are distinct from one another due to low levels of genetic exchange and recombination. Pangenomes of seven *G. soja* accessions were generated (Li et al. 2014) using the Illumina HiSeq2000 sequencing reads and SOAPdenovo assembler (Li et al. 2010). The average genome coverage was 119.9× and the estimated genome size ranged from 889.33 Mbp for the accession GsojaG (93.6% of the GmaxW82) to 1118.34 Mbp for the accession GsojaD (117.7% of GmaxW82) (Li et al. 2014). In this study, the estimated average number of genes per *G. soja* genome is 55,570, slightly more than the *G. max* Williams 82 genome. These *G. soja* assemblies enabled accurate detection of variation and comparative analyses within genic regions and found significant structural variations when compared to the *G. max* assembly.

Table 5.3 Final assembly statistics of *G. soja*, accession W05 reference genome

Total scaffold	51,178
Total contig	9008
Total scaffold sequence	868 Mb
Total contig sequence	808.67 Mb
Scaffold N/L50	442/401.3 Kb
Contig N/L50	8897/24.17 Kb

5.5 Conclusion and Future Perspectives

To meet the increasing global demand of grain legumes, including soybean, we have to maximize the development of high-yielding cultivars with better plant performances under the biotic and abiotic stress conditions. In order to achieve this, goal precise dissection of genomic regions associated with major traits to the haplotype/allelic level is needed leading to the next-generation breeding strategies. Advances in the next-generation sequencing platforms, cost-effectiveness per bp of genome sequences, and the development of various computational tools including better genome assemblers will expedite the generation of individual crop genomes. Generation of long-sequence reads and the associated genome assemblers will substantially reduce the major challenges in genome assembly including the one posed by genomic repeats and help improve the haplotype resolution and analysis of genomic variation. The accurate detection of genomic variation and comparative analyses within genic region is essential for the discovery of novel genes or alleles associated with the traits of interest and development of new varieties.

A single reference genome is not enough to capture the genomic diversity due to a large amount of structural variation including the copy number variations and presence/absence variation, which significantly alter the individual genomic sequences. Pangenomes representing landrace, elite, and wild soybean from the global collection are needed to address the above issue of genome structure variation and mining for rare alleles. Large-scale sequencing of soybean germplasm and additional reference genome build for cultivated and wild soybean is in pipeline and will be available soon to the soybean community. From these efforts, it is found that long-read sequencing of soybean genome using the PacBio platform helped to improve the whole genome assembly far better. Wild soybean is important resource for novel alleles due to its higher genomic diversity. More reference genome or pangenomes for wild soybean will significantly influence the tapping of rare genomic resources towards crop improvement. Soybean genomes exhibit higher linkage disequilibrium (LD) and more germplasm lines with deep sequencing will help enable precise genotype–phenotype association studies towards pinpointing the candidate gene associate with the specific trait and for the genomic predictions. In addition, development of a robust haplotype map for soybean (HapMap2) is inevitable to find specific genetic variants that affect plant developmental processes and response to disease and climate changes. Generation of large-scale sequencing of germplasm and various genetic populations need a better data visualization tool to compare and mine the genomic information at the allelic level. This will enable breeders and soybean community to easily associate genome sequence information with traits of interest and apply to genome enabled crop improvement programs including genome editing and next-generation breeding strategies.

Acknowledgements The authors are grateful to the United Soybean Board and the United States Department of Agriculture for project support.

References

Arumuganathan K, Earle ED (1991) Nuclear DNA content of some important plant species. Plant Mol Biol Rep 9:208–219

Batzoglou S, Jaffe DB, Stanley K, Butler J, Gnerre S et al (2002) ARACHNE: a whole-genome shotgun assembler. Genome Res 12(1):177–189

Bennetzen JL, Wang H (2014) The contributions of transposable elements to the structure, function, and evolution of plant genomes. Annu Rev Plant Biol 65:505–530

Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK et al (2008) ALLPATHS: de novo assembly of whole-genome shotgun microreads. Genome Res 18:810–820

Chaisson MJ, Pevzner PA (2008) Short read fragment assembly of bacterial genomes. Genome Res 18:324–330

de la Chaux N, Tsuchimatsu T, Shimizu KK, Wagner A (2012) The predominantly selfing plant *Arabidopsis thaliana* experienced a recent reduction in transposable element abundance compared to its outcrossing relative *Arabidopsis lyrata*. Mob DNA 3:2

Eid J, Fehr A, Gray J, Luong K, Luong K et al (2009) Real-time DNA sequencing from single polymerase molecules. Science 323:133–138

Fuller CW, Middendorf LR, Benner SA, Church GM, Harris T et al (2009) The challenges of sequencing by synthesis. Nat Biotechnol 27:1013–1023

Galindo-González L, Mhiri C, Deyholos MK, Grandbastien MA (2017) LTR-retrotransposons in plants: engines of evolution. Gene S0378–1119(17):30322

Goldberg RB (1978) DNA sequence organization in the soybean plant. Biochem Genet 16:45–51

Goldblatt P (1981) Cytology and phylogeny of Leguminosae. In: Polhill RM, Raven PH (eds) Advances in legume systematics, part 2. Royal Botanic Gardens, Kew, pp 427–463

Golicz AA, Batley J, Edwards D (2016) Towards plant pangenomics. Plant Biotechnol J 14:1099–1105

Ha J, Abernathy B, Nelson W, Grant D, Wu X et al (2012) Integration of the draft sequence and physical map as a framework for genomic research in soybean (*Glycine max* (L.) Merr.) and wild soybean (*Glycine soja* Sieb. and Zucc.). Genes Genomes Genet (Bethesda) 2:321–329

Hashmi U, Shafqat S, Khan F, Majid M, Hussain H et al (2015) Plant exomics: concepts, applications and methodologies in crop improvement. Plant Signal Behav 10(1):e976152

Hernandez D, Francois P, Farinelli L, Osterås M, Schrenzel J (2008) De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. Genome Res 18:802–809

Hymowitz T (1970) On the domestication of soybean. Econ Bot 24:408–421

Hymowitz T (2004) Speciation and cytogentics. In Boerma HR, Specht JE (eds) Soybeans: improvement, production and uses. American Society of Agronomy, Madison, pp 97–136

Imelfort M, Edwards D (2009) De novo sequencing of plant genomes using second-generation technologies. Brief Bioinform 10:609–618

International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. Nature 436:793–800

Jackman SD, Vandervalk BP, Mohamadi H, Chu J, Yeo S et al (2017) ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. Genome Res. doi:10.1101/gr.214346.116

Jaffe DB, Butler J, Gnerre S, Mauceli E, Lindblad-Toh K et al (2003) Whole-genome sequence assembly for mammalian genomes: Arachne 2. Genome Res 13:91–96

Jeck WR, Reinhardt JA, Baltrus DA, Hickenbotham MT, Magrini V et al (2007) Extending assembly of short DNA sequences to handle error. Bioinformatics 23:2942–2944

Jiao WB, Accinelli G, Hartwig B, Kiefer C, Baker D et al (2017a) Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data. Genome Res 27:778–786

Jiao WB, Accinelli G, Hartwig B, Kiefer C, Baker D et al (2017b) Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data. Genome Res. doi:10.1101/gr.213652.116

Joshi T, Valliyodan B, Wu JH, Lee SH, Xu D, Nguyen HT (2013) Genomic differences between cultivated soybean, *G. max* and its wild relative *G. soja*. BMC Genomics 14(Suppl 1):S5

Keim P, Diers BW, Olson TC, Shoemaker RC (1990) RFLP mapping in soybean: association between marker loci and variation in quantitative traits. Genetics 126:735–742

Keim P, Schupp JM, Travis SE, Clayton K, Zhu T et al (1997) A high density soybean genetic map based on AFLP markers. Crop Sci 37:537–543

Kim MY, Lee S, Van K, Kim TH, Jeong SC et al (2010) Whole-genome sequencing and intensive analysis of the undomesticated soybean (*Glycine soja* Sieb. and Zucc.) genome. Proc Natl Acad Sci USA 107:22032–22037

Lander ES, Waterman MS (1988) Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics 2:231–239

Li R, Zhu H, Ruan J, Qian W, Fang X et al (2010) De novo assembly of human genomes with massively parallel short read sequencing. Genome Res 20:265–272

Li YH, Zhou G, Ma J, Jiang W, Jin LG et al (2014) De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. Nat Biotechnol 10:1045–1052

Libault M, Farmer A, Joshi T, Takahashi K, Langley RJ et al (2010) An integrated transcriptome atlas of the crop model *Glycine max*, and its use in comparative analyses in plants. Plant J 63:86–99

Luo MC, Thomas C, You FM, Hsiao J, Ouyang S et al (2003) High-throughput fingerprinting of bacterial artificial chromosomes using the SNaPshot labeling kit and sizing of restriction fragments by capillary electrophoresis. Genomics 82:378–389

Marek LF, Mudge J, Darnielle L, Grant D, Hanson N et al (2001) Soybean genomic survey: BAC-end sequences near RFLP and SSR markers. Genome 44:572–581

Margulies M, Margulies M, Egholm M, Altman WE et al (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380

Myers EW (1995) Toward simplifying and accurately formulating fragment assembly. J Comput Biol 2:275–290

Nakano K, Shiroma A, Shimoji M, Tamotsu H, Ashimine N et al (2017) Advantages of genome sequencing by long-read sequencer using SMRT technology in medical area. Hum Cell. doi:10.1007/s13577-017-0168-8

Pevzner PA, Tang H, Waterman MS (2001) An Eulerian path approach to DNA fragment assembly. Proc Natl Acad Sci USA 98:9748–9753

Qi X, Li MW, Xie M, Liu X, Ni M et al (2014) Identification of a novel salt tolerance gene in wild soybean by whole-genome sequencing. Nat Commun 5:4340

Rhoads A, Au KF (2015) PacBio sequencing and its applications. Genom Proteom Bioinform 13:278–289

Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR et al (1977) Nucleotide sequence of bacteriophage φX174 DNA. Nature 24:687–695

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T et al (2010) Genome sequence of the palaeopolyploid soybean. Nature 463:178–183

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F et al (2009) The B73 maize genome: complexity, diversity, and dynamics. Science 326:1112–1115

Schwartz DC, Li X, Hernandez LI, Ramnarain SP, Huff EJ, Wang YK (1993) Ordered restriction maps of *Saccharomyces cerevisiae* chromosomes constructed by optical mapping. Science 262:110–114

Severin AJ, Woody JL, Bolon YT, Joseph B, Diers BW et al (2010) RNA-Seq atlas of *Glycine max*: a guide to the soybean transcriptome. BMC Plant Biol 10:160

Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP et al (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309:1728–1732

Shoemaker R, Keim P, Vodkin L, Retzel E, Clifton SW et al (2002) A compilation of soybean ESTs: generation and analysis. Genome 45:329–338

Shoemaker RC, Grant D, Olson T, Warren WC, Wing R et al (2008) Microsatellite discovery from BAC end sequences and genetic mapping to anchor the soybean physical and genetic maps. Genome 51:294–302

Shultz JL, Ray JD, Lightfoot DA (2007) A sequence based synteny map between soybean and *Arabidopsis thaliana*. BMC Genom 8:8

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I (2009) ABySS: a parallel assembler for short read sequence data. Genome Res 19:1117–1123

Singh RJ, Hymowitz T (1988) The genomic relationship between *Glycine max* (L.) Merr. and *G. soja* Sieb. and Zucc. as revealed by pachytene chromosome analysis. Theor Appl Genet 76:705–711

Song Q, Jenkins J, Jia G, Hyten DL, Pantalone V et al (2016) Construction of high resolution genetic linkage maps to improve the soybean genome sequence assembly Glyma1.01. BMC Genom 17:33

Sutton GG, White O, Adams MD, Kerlavage AR (1995) TIGR assembler: a new tool for assembling large shotgun sequencing projects. Genome Sci Technol 1:9–19

Takata M, Kiyohara A, Takasu A, Kishima Y, Ohtsubo H, Sano Y (2007) Rice transposable elements are characterized by various methylation environments in the genome. BMC Genom 8:469

Tang H, Lyons E, Town CD (2015) Optical mapping in plant comparative genomics. Gigascience 4:3

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815

The International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. Nature 431:931–945

United States Department of Agriculture-Foreign Agricultural Service (2017) World agricultural production, Circular Series WAP 05-17

Utturkar SM, Klingeman DM, Land ML, Schadt CW, Doktycz MJ, Pelletier DA, Brown SD (2014) Evaluation and validation of de novo and hybrid assembly techniques to derive high-quality genome sequences. Bioinformatics 30:2709–2716

Valliyodan B, Qiu D, Patil G, Zeng P, Huang J et al (2016) Landscape of genomic diversity and trait discovery in soybean. Sci Rep 6:23598

Varshney RK, Nayak SN, May GD, Jackson SA (2009) Next-generation sequencing technologies and their implications for crop genetics and breeding. Trends Biotechnol 27(9):522–530

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ et al (2001) The sequence of the human genome. Science 291:1304–1351

Wang Z, Libault M, Joshi T, Valliyodan B, Nguyen HT et al (2010) SoyDB: a knowledge database of soybean transcription factors. BMC Plant Biol 10:14

Warren WC, The Soybean Mapping Consortium (2006) A physical map of the "Williams 82" soybean (*Glycine max*) genome. Abstract W151. In: Plant and animal genomes XIV conference, San Diego, CA, 14–18 Jan 2006

Warren RL, Sutton GG, Jones SJM, Holt RA (2007) Assembling millions of short DNA sequences using SSAKE. Bioinformatics 23:500–501

Weissensteiner MH, Pang AWC, Bunikis I, Höijer I, Vinnere-Petterson O, Suh A, Wolf JBW (2017) Combination of short-read, long-read, and optical mapping assemblies reveals large-scale tandem repeat arrays with population genetic implications. Genome Res. doi:10.1101/gr.215095.116

Wu X, Ren C, Joshi T, Vuong T, Xu D, Nguyen HT (2010) SNP discovery by high-throughput sequencing in soybean. BMC Genom 11:469

Wu X, Vuong TD, Leroy JA, Shannon GJ, Sleper DA, Nguyen HT (2011) Selection of a core set of RILs from Forrest × Williams 82 to develop a framework map in soybean. Theor Appl Genet 122:1179–1187

Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821–829